Learning to Detect an Odd Markov Arm

P. N. Karthik and Rajesh Sundaresan

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. Our analysis of the rested Markov setting is a key first step in understanding the difficult case of restless Markov setting, which is still open.

Index Terms

Multi-armed bandits, rested bandits, Markov rewards, odd arm identification, anomaly detection, forced exploration.

I. Introduction

THE classical single player multi-armed bandit problem initially proposed by Robbins [1] is one of sequential decision making under uncertainty. A player (also referred to as a learner) has to decide between K alternatives (or arms) when allowed to select only one at each time. Each arm returns a reward when selected. The arm selected at any time is a function only of the past choices of arms and the corresponding rewards obtained. In this setting, the goal of the learner is to maximise the long-term average reward by exercising a balanced trade-off between the decision to *exploit* an arm that, based on the current state of knowledge, appears to yield high rewards, versus the decision to *explore* other arms which could potentially yield high rewards in the future. Thus, every sequential arm selection strategy for the multi-armed bandit problem involves an exploration versus exploitation trade-off.

A. Prior Work and Motivation

While the work of Robbins considers the setting in which the successive rewards received from any given arm are independent and identically distributed (iid) according to a fixed but unknown distribution, extensions to this work that consider more general structures on the rewards have appeared in the literature. The seminal paper of Gittins [2] considers the setting in which the rewards from each arm constitute a time homogeneous Markov process with a *known* transition law. Furthermore, at any given time, only the arm chosen by the learner exhibits a state transition, while rest of the arms remain frozen (or *rested*) at their previously observed states. Such a structure on the arms is not restrictive, and closely models a host of real-life scenarios, as outlined in [3, Chapter 1]. The central problem in [2] is one of maximising the long-term average discounted reward obtained by sequentially selecting the arms. For this problem, Gittins proposed and demonstrated the optimality of a simple index-based policy that involves constructing an index for each arm using the knowledge of the transition laws of the arms, and selecting at any given time the arm with the largest index.

Agarwal et al. [4] consider a similar setting as Gittins', where the rewards from each arm are Markovian. However, an important feature of the work in [4] is that, unlike Gittins [2], the authors do not assume the knowledge of the transition laws of the arms. They then provide a strengthening of the results of [2]. Since the setting and the results of Agarwal et al. will be of relevance to us in this paper, in what follows, we describe their work in some detail. In [4], a system whose transition law is parameterised by an unknown parameter belonging to a known, finite, parameter space is considered. At any given time, one out of finitely many possible controls (or actions) is applied to the system, resulting in a reward based on the action applied and the system's current state, followed by a transition of the system's current state. The choice of action to be applied at any given time is based on the past history of actions and states, and the current state of the system. The performance of any sequence of actions and states (known as a policy) is measured by the infinite horizon expected sum of the rewards generated by the policy.

The authors of [4] compare the performance of any given policy in relation to a policy that has knowledge of the true parameter of the system by introducing a loss function that is defined, at any given time n, as the total reward generated up to time n by the policy that has knowledge of the system parameter minus the expected value of the total reward generated by the

The authors are with the Department of Electrical Communication Engineering at the Indian Institute of Science, Bangalore 560012, Karnataka, India.

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.

This work will be presented in part at the 2019 IEEE International Symposium on Information Theory (ISIT).

given policy up to time n. Subsequently, the authors of [4] define a notion of optimality for a policy as one for which the loss function of the policy grows as $o(n^\alpha)$ for every $\alpha > 0$, and look for policies that are optimal within this *regret minimisation* framework.

The authors of [4] then show that the problem described above can be reduced to a problem of a multi-armed bandit with finitely many arms, each generating Markov rewards, as considered by Gittins [2], but with transition laws parameterised by an unknown parameter coming from a finite parameter set. Exploiting the structure arising from the multi-armed bandit setting, the authors of [4] then show that for any policy, the ratio of its loss function to that of $\log(n)$ may be lower bounded in terms of a problem-instance dependent constant that quantifies the effort required by any policy to distinguish the true parameter of the system from its nearest alternative. Furthermore, the authors of [4] propose a policy and demonstrate that its performance meets the lower bound, hence proving that their policy is optimal.

While the lower bound in [4] quantifies the growth rate of the loss in performance of a policy in relation to the policy that has full knowledge of the system's true parameter, it does not reflect the quickness of learning by a policy. That is, the results in [4] do not answer the question of how many controls (or actions) are needed, on the average, in order to learn the true parameter of the system to a desired level of accuracy. In this paper, we answer this question at the same level of generality of [4], when one of the Markov arms is anomalous.

B. Setting and Problem Description

Let us recall our setting. We have a multi-armed bandit problem with finitely many arms in which each arm is identified with a time homogeneous, irreducible and aperiodic discrete time Markov process on a common finite state space, as in [4]. We assume that each arm yields Markov rewards independently of the other arms. Further, we impose the following condition on the arms: the evolution of states on one of the arms is governed by a transition matrix P_1, while those of the other arms is governed by a common transition matrix P_2, where $P_2 \neq P_1$, hence making one of the arms anomalous (hereinafter referred to as the odd arm). A learner, who has no knowledge of the above transition matrices, seeks to identify the index of the odd arm as quickly as possible. In order to do so, the learner devises sequential arm selection schemes in which at each time step, he chooses one of the arms and observes a state transition on the chosen arm. During this time, all the unobserved arms remain frozen at their previously observed states and do not exhibit state transitions, thus making the arms *rested* as in [2].

Given a maximum acceptable error probability, the goal of the learner is to devise sequential schemes for identifying the index of the odd arm with (a) the least expected number of arm selections, and (b) the probability of error at stoppage being within the acceptable level. We note here that the unknown parameters of our problem are the transition laws of the odd arm and the non-odd arm Markov processes, and the index of the odd arm, thus making our parameter set a continuum, unlike a finite parameter set in [4]. However, our goal is only to identify the index of the odd arm.

The structure of anomaly imposed on the arms in the context of odd arm identification is not new, and has been dealt with in the recent works of Vaidhiyan et al. [5] for the case of iid Poisson observations from each arm, and of Prabhu et al. [6] for the case of iid observations belonging to a generic exponential family. The works [5] and [6] can be embedded within the classical works of Chernoff [7] and Albert [8], and provide a general framework for the analysis of lower bounds on expected number of samples required for identifying the index of the odd arm. In addition, they also provide explicit schemes that achieve these lower bounds in the asymptotic regime as error probability vanishes. We refer the reader to also [9]–[16] for other related works on iid observations.

In this paper, we present results similar in spirit to those of [5]–[8], but for the important setting of Markov arms. To the best of our knowledge, there is no prior work on odd arm identification for the case of multi-armed bandits with Markov rewards, and this paper is the first to study this setting.

C. Contributions

Below, we highlight the key contributions of our work. Further, we mention the similarities and differences of our work with the aforementioned ones, and also bring out the challenges that we need to overcome in the analysis for the Markov setting.

1) We derive an asymptotic lower bound on the expected number of arm selections required by any policy that the learner may use to identify the index of the odd arm. Here, the asymptotics is as the error probability vanishes. Similar to the lower bounds appearing in [4]- [6], our lower bound has a problem-instance dependent constant that quantifies the effort required by any policy to identify the true index of the odd arm by guarding itself against identifying the nearest, incorrect alternative.
2) We characterise the growth rate of the expected number of arm selections required by any policy as a function of the maximum acceptable error probability, and show that in the regime of vanishingly small error probabilitys, this growth rate is logarithmic in the inverse of the error probability. The analysis of the lower bounds in [5] and [6] uses the familiar data processing inequality presented in the work of Kaufmann et al. [10] that is based on Wald's identity [17] for iid processes. However, the Markov setting in our problem does not permit the use of Wald's identity. Therefore, we derive results for our Markov setting generalising those appearing in [10], and subsequently use these generalisations to arrive at the lower bound. See Section III for the details.

3) In the analysis of the lower bound, we bring out the key idea that the empirical proportion of times an arm is observed to exit out of a state is equal, in the long run, to the empirical proportion of times it is observed to enter into the same state. We then replace these common proportions by the probability of observing the arm occupying this state under the arm's stationary distribution. This is possible due to the rested nature of the arms, and may not hold in a more general setting of "restless" arms where the unobserved arms continue to undergo state transitions.
4) We propose a sequential arm selection scheme that takes as inputs two parameters, one of which may be chosen appropriately to meet the acceptable error probability, while the other may be tuned to ensure that the performance of our scheme comes arbitrarily close to the lower bound, thus making our scheme near-optimal.
   We now contrast the near-optimality of our scheme with the near-optimality of the scheme proposed by Vaidhiyan et al. in [5], and highlight a key simplification that our scheme entails. The scheme of Vaidhiyan et al. is built around the important fact that each arm is sampled at a non-trivial, strictly positive and optimal rate that is bounded away from zero, as given by the lower bound, thereby allowing for exploration of the arms in an optimal manner. This stemmed from their specific Poisson observations. However, the lower bound presented in Section III may not have this property in the context of Markov observations. Therefore, recognising the requirement of sampling the arms at a non-trivial rate for good performance of our scheme, in this paper, we use the idea of "forced exploration" proposed by Albert in [8]. In particular, we propose a simplified way of sampling the arms by considering a mixture of uniform sampling and the optimal sampling given by the lower bound in Section III. We do this by introducing an appropriately tuneable parameter that controls the probability of switching between uniform sampling and optimal sampling, the latter being given by the lower bound. While this ensures that our policy samples each arm with a strictly positive probability at each step, it also gives us the flexibility to select an appropriate value for this parameter so that the upper bound on the performance of our scheme may be made arbitrarily close to our lower bound. We refer the reader to Section IV for the details.

D. Organisation

The rest of the paper is organised as follows. In Section II, we set up some of the basic notations that will be used throughout the paper. In Section III, we present a lower bound on the performance of any policy. In Section IV, we present a sequential arm selection policy and demonstrate its near optimality. We present the main result of this paper in Section V, combining the results of Sections III and IV. In Section VI, we provide some simulation to support the theoretical development, and provide concluding remarks in Section VIII. We present the proofs of the main results in Section VII.

II. Notations

In this section, we set up the notations that will be used throughout the rest of this paper. Let $K \geq 3$ denote the number of arms, and let $\mathcal{A} = \{1, 2, \ldots, K\}$ denote the set of arms. We associate with each arm an irreducible, aperiodic, time homogeneous discrete-time Markov process on a finite state space $\mathcal{S}$, where the Markov process of each arm is independent of the Markov processes of the other arms. We denote by $|\mathcal{S}|$ the cardinality of $\mathcal{S}$. Without loss of generality, we take $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$. Hereinafter, we use the phrase 'Markov process of arm a' to refer to the Markov process associated with arm $a \in \mathcal{A}$.

At each discrete time instant, one out of the K arms is selected and its state is observed. We let A_n denote the arm selected at time n, and let $\bar{X}_n$ denote the state of arm A_n, where $n \in \{0, 1, 2, \ldots\}$. We treat A_0 as the zeroth arm selection and Z_0 as the zeroth observation. Selection of an arm at time n is based on the history $(\bar{X}^{n-1}, A^{n-1})$ 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 the collection $(X_k^a)_{k \geq 0}$ of random variables. Further, we denote by $N_a(n)$ the number of times arm a is selected by a policy up to (and including) time n, i.e.,

$$N_a(n) = \sum_{t=0}^{n} 1_{\{A_t = a\}}. \tag{1}$$

Then, for each $n \geq 0$, we have the observation

$$\bar{X}_n = X^{A_n}_{N_{A_n}(n)-1}. \tag{2}$$

We consider a scenario in which the Markov process of one of the arms (hereinafter referred to as the odd arm) follows a transition matrix $P_1 = (P_1(j|i))_{i,j \in \mathcal{S}}$, while those of rest of the arms follow a transition matrix $P_2 = (P_2(j|i))_{i,j \in \mathcal{S}}$, where $P_2 \neq P_1$; here, $P(j|i)$ denotes the entry in the ith row and jth column of the matrix P. Further, we let μ_1 and μ_2 denote the unique stationary distributions of P_1 and P_2 respectively. We denote by ν the common distribution for 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 operate in a setting where the transition matrices and their associated stationary distributions are unknown to the learner.

For each $a \in \mathcal{A}$ and state $i \in \mathcal{S}$, we denote by $N_a(n, i)$ the number of times up to (and including) time n the Markov process of arm a is observed to *exit* out of state i, i.e.,

$$N_a(n, i) = \sum_{m=1}^{N_a(n)-1} 1_{\{X_{m-1}^a = i\}}. \tag{3}$$

Similarly, for each $i, j \in \mathcal{S}$, we denote by $N_a(n,i,j)$ the number of times up to (and including) time n the Markov process of arm a is observed to exit out of state i and enter into state j, i.e.,

$$N_a(n,i,j) = \sum_{m=1}^{N_a(n)-1} 1_{\{X^a_{m-1}=i,\, X^a_m=j\}}. \tag{4}$$

Clearly, then, the following hold:

1) For each $a \in \mathcal{A}$ and $i \in \mathcal{S}$,
$$\sum_{j\in\mathcal{S}} N_a(n,i,j) = N_a(n,i). \tag{5a}$$
2) For each $a \in \mathcal{A}$,
$$\sum_{i\in\mathcal{S}} N_a(n,i) = N_a(n) - 1. \tag{5b}$$
3) For each n,
$$\sum_{a\in\mathcal{A}} N_a(n) = n + 1. \tag{5c}$$

We note here that the upper index of the summation in (3) is $N_a(n) - 1$, and not $N_a(n)$, since the last observed transition on arm a would be an exit out of the state given by $X^a_{N_a(n)-2}$ and an entry into the state given by $X^a_{N_a(n)-1}$. This is further reflected by the summation in (5b).

Fix transition matrices P_1 and P_2, where $P_2 \neq P_1$, and let H_h denote the hypothesis that h is the index of the odd arm. The transition matrix of arm h is P_1; all other arms have P_2. We refer to the triplet $C = (h, P_1, P_2)$ as a configuration. Our problem is one of detecting the true hypothesis among all possible configurations given by

$$\mathcal{C} = \{C = (h, P_1, P_2) : h \in \mathcal{A},\, P_2 \neq P_1\}$$

when P_1 and P_2 are unknown. Let $C = (h, P_1, P_2)$ denote the underlying configuration of the arms. For each $a \in \mathcal{A}$, we denote by $(Z^a_h(n))_{n\geq 0}$ the log-likelihood process of arm a under configuration C, with h being the true index of the odd arm. Using the notations introduced above, we may then express $Z^a_h(n)$ as

$$Z^a_h(n) = \begin{cases} 0, & N_a(n) = 0, \\ \log \nu(X^a_0), & N_a(n) = 1, \\ \log \nu(X^a_0) + \sum\limits_{m=1}^{N_a(n)-1} \log P^a_h(X^a_m | X^a_{m-1}), & N_a(n) \geq 2, \end{cases} \tag{6}$$

where $P^a_h(j|i)$ denotes the conditional probability under hypothesis H_h of observing state j on arm a given that state i was observed on arm a at the previous sampling instant, and is given by

$$P^a_h(j|i) = \begin{cases} P_1(j|i), & a = h, \\ P_2(j|i), & a \neq h. \end{cases} \tag{7}$$

Then, since the Markov processes of all the arms are independent of one another, for a given sequence $(A^n, \bar{X}^n)$ of arm selections and observations under a policy π and a configuration $C = (h, P_1, P_2)$, denoting by $(Z_h(n))_{n\geq 0}$ the log-likelihood process under hypothesis H_h of all arm selections and observations up to time n, we have

$$Z_h(n) = \sum_{a=1}^{K} Z^a_h(n), \tag{8}$$

where $Z^a_h(n)$ is as given in (6). On similar lines, for any two configurations $C = (h, P_1, P_2)$ and $C' = (h', P'_1, P'_2)$, where $P'_2 \neq P'_1$ and $h' \neq h$, for each $a \in \mathcal{A}$, we define the log-likelihood process $(Z^a_{hh'}(n))_{n\geq 0}$ of configuration C with respect to configuration C' for arm a as

$$\begin{aligned} Z^a_{hh'}(n) &= Z^a_h(n) - Z^a_{h'}(n) \\ &= \begin{cases} 0, & N_a(n) = 0, 1, \\ \sum\limits_{m=1}^{N_a(n)-1} \log \dfrac{P^a_h(X^a_m | X^a_{m-1})}{P^a_{h'}(X^a_m | X^a_{m-1})}, & N_a(n) \geq 2. \end{cases} \end{aligned} \tag{9}$$

We note that in the above equation, for P^a_h, we should use (7), and for $P^a_{h'}$, we shall use, for all $a \in \mathcal{A}$ and $i, j \in \mathcal{S}$,

$$P^a_{h'}(j|i) = \begin{cases} P'_1(j|i), & a = h', \\ P'_2(j|i), & a \neq h'. \end{cases} \tag{10}$$

Finally, we denote by $(Z_{hh'}(n))_{n\geq 0}$ the log-likelihood process of configuration C with respect to C' as

$$Z_{hh'}(n) = \sum_{a=1}^{K} Z^{a}_{hh'}(n), \tag{11}$$

which includes all arm selections and observations.

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 $(\Omega, \mathcal{F}, P)$. We define the filtration $(\mathcal{F}_n)_{n\geq 0}$ as

$$\mathcal{F}_n = \sigma(A^n, \bar{X}^n), \quad n \geq 0. \tag{12}$$

We use the convention that the zeroth arm selection A_0 is measurable with respect to the sigma algebra $\{\phi, \Omega\}$, whereas for all $n \geq 1$, the nth arm selection A_n is $\mathcal{F}_{n-1}$-measurable. For any stopping time τ with respect to the filtration in (12), we denote by $\mathcal{F}_\tau$ the σ-algebra

$$\mathcal{F}_\tau = \{E \in \mathcal{F} : E \cap \{\tau = n\} \in \mathcal{F}_n \text{ for all } n \geq 0\}. \tag{13}$$

Our focus will be on policies π that identify the index of the odd arm by sequentially sampling the arms, one at every time instant, and learning from the arms selected and observations obtained in the past. Specifically, at any given time, a policy π prescribes one of the following alternatives:

1) Select an arm, based on the history of past observations and arms selected, according to a fixed distribution λ independent of the underlying configuration of the arms, i.e., for each $n \geq 1$,

$$P(A_n = a | A^{n-1}, \bar{X}^{n-1}) = \lambda(a). \tag{14}$$

2) Stop selecting arms, and declare the index $I(\pi)$ as the odd arm.

Given a maximum acceptable error probability $\epsilon > 0$, we denote by $\Pi(\epsilon)$ the family of all policies whose probability of error at stoppage for any underlying configuration of the arms is at most ϵ. That is,

$$\Pi(\epsilon) = \left\{ \pi : P^{\pi}(I(\pi) \neq h | C) \leq \epsilon \ \forall \ C = (h, P_1, P_2), \text{ where } h \in \mathcal{A} \text{ and } P_2 \neq P_1 \right\}. \tag{15}$$

For a policy π, we denote its stopping time by $\tau(\pi)$. Further, we write $E^{\pi}[\cdot|C]$ and $P^{\pi}(\cdot|C)$ to denote expectations and probabilities given that the underlying configuration of the arms is C. In this paper, we characterise the behaviour of $E^{\pi}[\tau(\pi)|C]$ for any policy $\pi \in \Pi(\epsilon)$, as ϵ approaches zero. 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, $P_2 \neq P_1$. 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* (15)*, it follows that*

$$\Pi(\epsilon) = \bigcap_{P_1, P_2 : P_2 \neq P_1} \Pi(\epsilon | P_1, P_2). \tag{16}$$

That is, policies in $\Pi(\epsilon)$ work for any P_1, P_2, with $P_2 \neq P_1$. 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 [7]. In this paper, we show that $\Pi(\epsilon)$ is nonempty even for the setting of rested and Markov arms.

In the next section, we provide a configuration dependent lower bound on $E^{\pi}[\tau(\pi)|C]$ for any policy $\pi \in \Pi(\epsilon)$. In Section IV, we propose a sequential arm selection policy that achieves the lower bound asymptotically as the probability of error vanishes. We present the proofs in Section VII.

III. The 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

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

with the convention $0 \log 0 = 0 \log \frac{0}{0} = 0$. 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 underlying configuration of the arms. Then,*

$$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^{\pi}[\tau(\pi)|C]}{\log \frac{1}{\epsilon}} \geq \frac{1}{D^*(h, P_1, P_2)}, \tag{17}$$

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

$$D^*(h, P_1, P_2) = \max_{0 \leq \lambda_1 \leq 1} \left\{ \lambda_1 \, D(P_1||P|\mu_1) + (1-\lambda_1) \frac{(K-2)}{(K-1)} D(P_2||P|\mu_2) \right\}. \tag{18}$$

In (18)*, P is a transition probability matrix whose entry in the ith row and jth column is given by*

$$P(j|i) = \frac{\lambda_1 \mu_1(i) P_1(j|i) + (1-\lambda_1) \frac{(K-2)}{(K-1)} \mu_2(i) P_2(j|i)}{\lambda_1 \mu_1(i) + (1-\lambda_1) \frac{(K-2)}{(K-1)} \mu_2(i)}. \tag{19}$$

□

The proof of Proposition 1 broadly follows the outline of the proof of the lower bound in [10], with necessary modifications for the setting of Markov rewards. We now outline some of the key steps in the proof. For an arbitrary choice of error probability $\epsilon > 0$, we first show that for any policy $\pi \in \Pi(\epsilon)$, the expected value of the total sum of log-likelihoods up to the stopping time $\tau(\pi)$ can be lower bounded by the binary relative entropy function

$$d(\epsilon, 1-\epsilon) \coloneqq \epsilon \log \frac{\epsilon}{1-\epsilon} + (1-\epsilon) \log \frac{1-\epsilon}{\epsilon}. \tag{20}$$

Next, we express the expected sum of log-likelihoods up to the stopping time $\tau(\pi)$ in terms of the expected value of the stopping time. It is in obtaining such an expression that works such as [10], [5] and [6] that are based on iid observations use Wald's identity, which greatly simplifies their analysis of the lower bound. Our setting of Markov rewards does not permit us to use Wald's identity. Therefore, we first obtain a generalisation of [10, Lemma 18], a change of measure based argument, to the setting of Markov rewards, and subsequently use this generalisation to obtain the desired relation.

We then show that for any arm $a \in \mathcal{A}$, the long run frequency of observing the arm exit out of a state $i \in \mathcal{S}$ is equal to that of observing arm a enter into the state i, and note that this common frequency is the stationary probability of observing the arm in state i. This explains the appearance of the unique stationary distributions μ_1 and μ_2 of the odd arm and the non-odd arms respectively in the expression (18). We wish to emphasise that this step in the proof is possible due to the rested nature of the arms. The lower bound in the more general setting of "restless" arms in which the unobserved arms continue to undergo state transitions is still open.

Finally, combining the above steps and using $d(\epsilon, 1-\epsilon)/\log \frac{1}{\epsilon} \to 1$ as $\epsilon \downarrow 0$, we arrive at the lower bound in (17). The details may be found in Section VII-A.

Remark 2. *The right-hand side of* (18) *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, and we deduce that $D^*(h, P_1, P_2)$ does not depend on h. However, we include the index h of the odd arm for the sake of consistency with the notation $C = (h, P_1, P_2)$ used to denote arm configurations.*

Going further, we let $\lambda^* \in [0,1]$ denote the value of λ that achieves the maximum in (18). 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 on $\mathcal{A}$ given by

$$\lambda_{opt}(h, P_1, P_2)(a) \coloneqq \begin{cases} \lambda^*, & a = h, \\ \frac{1-\lambda^*}{K-1}, & a \neq h. \end{cases} \tag{21}$$

In the next section, we construct a policy that, at each time step, chooses arms with probabilities that match with those in (21) 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 propose a scheme that asymptotically achieves the lower bound of Section III, as the probability of error vanishes. Our policy is a modification of the policy proposed by Prabhu et al. [6] for the case of K iid processes. We denote our policy by $\pi^\star(L, \delta)$, where $L \geq 1$ and $\delta \in (0,1)$ are the parameters of the policy.

Our policy is based on a modification of the classical generalised likelihood ratio (GLR) test in which we replace the maximum that appears in the numerator of the classical GLR statistic by an average computed with respect to a carefully constructed artificial prior over the space $\mathcal{P}(\mathcal{S})$ of all probability distributions on the state space $\mathcal{S}$. We describe this modified GLR statistic in the next section.

A. The Modified GLR Statistic

We revisit (8), and 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

$$Z_h(n)=\sum_{a=1}^{K}\log\nu(X_0^a)+\sum_{i,j\in\mathcal{S}}N_h(n,i,j)\log P_1(j|i)+\sum_{i,j\in\mathcal{S}}\left(\sum_{a\neq h}N_a(n,i,j)\right)\log P_2(j|i), \tag{22}$$

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

$$f(A^n,\bar{X}^n|C)=\prod_{a=1}^{K}\nu(X_0^a)\prod_{i,j\in\mathcal{S}}(P_1(j|i))^{N_h(n,i,j)}\cdot\prod_{i,j\in\mathcal{S}}(P_2(j|i))^{\sum\limits_{a\neq h}N_a(n,i,j)}. \tag{23}$$

We now introduce an artificial prior on the space of all transition probability matrices for the state space $\mathcal{S}$. Let $\mathrm{Dir}(1,\ldots,1)$ denote a Dirichlet distribution with $|\mathcal{S}|$ parameters $\alpha_1,\ldots,\alpha_{|\mathcal{S}|}$, where $\alpha_j=1$ 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, independently 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

$$\begin{aligned}\mathscr{D}(P_1,P_2)&\coloneqq\prod_{i\in\mathcal{S}}\frac{\prod\limits_{j\in\mathcal{S}}(P_1(j|i))^{\alpha_j-1}}{B(1\ldots,1)}\prod_{i\in\mathcal{S}}\frac{\prod\limits_{j\in\mathcal{S}}(P_2(j|i))^{\alpha_j-1}}{B(1\ldots,1)}\\&=\frac{1}{B(1,\ldots,1)^{2|\mathcal{S}|}},\end{aligned} \tag{24}$$

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 (23) computed with respect to the prior in (24). From the property that the Dirichlet distribution is the appropriate conjugate prior for the observation process,

$$f(A^n,\bar{X}^n|H_h)=\prod_{a=1}^{K}\nu(X_0^a)\prod_{i\in\mathcal{S}}\frac{B((N_h(n,i,j)+1)_{j\in\mathcal{S}})}{B(1,\ldots,1)}\prod_{i\in\mathcal{S}}\frac{B\left(\left(\sum\limits_{a\neq h}N_a(n,i,j)+1\right)_{j\in\mathcal{S}}\right)}{B(1,\ldots,1)}, \tag{25}$$

where in the above expression, $B((N_h(n,i,j)+1)_{j\in\mathcal{S}})$ denotes the normalisation factor for a Dirichlet distribution with parameters $(N_h(n,i,j)+1)_{j\in\mathcal{S}}$. It can be shown that $f(A^n,\bar{X}^n|H_h)$ is also the expected value of the likelihood in (23) computed with respect to the prior in (24), i.e.,

$$f(A^n,\bar{X}^n|H_h)=\prod_{a=1}^{K}\nu(X_0^a)\prod_{i\in\mathcal{S}}E\left[\prod_{j\in\mathcal{S}}X_{ij}^{N_h(n,i,j)}\cdot Y_{ij}^{\sum\limits_{a\neq h}N_a(n,i,j)}\right] \tag{26}$$

where in the above set of equations, the random vectors $(X_{ij})_{i,j\in\mathcal{S}}$ and $(Y_{ij})_{i,j\in\mathcal{S}}$ are independent with independent components, and jointly distributed according to (24), and the expectation is also with respect to this joint density.

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 (23) with respect to $P_1(j|i)$ and $P_2(j|i)$ for each $i,j\in\mathcal{S}$, and setting each of these derivatives to zero, we get

$$\hat{P}_{h,1}^n(j|i)=\frac{N_h(n,i,j)}{N_h(n,i)},\quad \hat{P}_{h,2}^n(j|i)=\frac{\sum\limits_{a\neq h}N_a(n,i,j)}{\sum\limits_{a\neq h}N_a(n,i)}. \tag{27}$$

Plugging the estimates in (27) back into (23), we get the maximum likelihood of all observations and actions under hypothesis H_h:

$$\begin{aligned}\hat{f}(A^n,\bar{X}^n|H_h)&\coloneqq\max_{C=(h,\cdot,\cdot)}f(A^n,\bar{X}^n|C)\\&=\prod_{a=1}^{K}\nu(X_0^a)\prod_{i,j\in\mathcal{S}}\left\{\left(\frac{N_h(n,i,j)}{N_h(n,i)}\right)^{N_h(n,i,j)}\left(\frac{\sum\limits_{a\neq h}N_a(n,i,j)}{\sum\limits_{a\neq h}N_a(n,i)}\right)^{\sum\limits_{a\neq h}N_a(n,i,j)}\right\}.\end{aligned} \tag{28}$$

We now define our modified GLR statistic. Let H_h and $H_{h'}$ be any two hypotheses, with $h' \neq h$. Let π be a policy whose sequence of arm selections and observations up to (and including) time n is $(A^n, \bar{X}^n)$. Then, the modified GLR statistic of H_h with respect to $H_{h'}$ up to time n is denoted by $M_{hh'}(n)$, and is defined as

$$\begin{aligned} M_{hh'}(n) &= \log \frac{f(A^n, \bar{X}^n | H_h)}{\hat{f}(A^n, \bar{X}^n | H_{h'})} \\ &= T_1 + T_2(n) + T_3(n) + T_4(n) + T_5(n), \end{aligned} \tag{29}$$

where the terms appearing in (29) are as follows.

1) The term T_1 is given by
$$T_1 = 2|\mathcal{S}| \log \left(\frac{1}{B(1, \ldots, 1)} \right). \tag{30}$$

2) The term $T_2(n)$ is given by
$$T_2(n) = \sum_{i \in \mathcal{S}} \log B((N_h(n,i,j)+1)_{j \in \mathcal{S}}). \tag{31}$$

3) The term $T_3(n)$ is given by
$$T_3(n) = \sum_{i \in \mathcal{S}} \log B\left(\left(\sum_{a \neq h} N_a(n,i,j) + 1 \right)_{j \in \mathcal{S}} \right). \tag{32}$$

4) The term $T_4(n)$ is given by
$$T_4(n) = -\sum_{i,j \in \mathcal{S}} N_{h'}(n,i,j) \log \frac{N_{h'}(n,i,j)}{N_{h'}(n,i)}. \tag{33}$$

5) The term $T_5(n)$ is given by
$$T_5(n) = -\sum_{i,j \in \mathcal{S}} \sum_{a \neq h'} N_a(n,i,j) \log \frac{\sum\limits_{a \neq h'} N_a(n,i,j)}{\sum\limits_{a \neq h'} N_a(n,i)}. \tag{34}$$

Note that ν, the distribution of the initial state of any arm, is irrelevant since it appears in both (25) and (28), and thus cancels out in writing (29). Let us emphasise that our modified GLR statistic is one in which the maximum in the numerator of the usual GLR statistic is replaced by an average in (25) computed with respect to the artificial prior over the space $\mathscr{P}(\mathcal{S})$ introduced in (24).

Remark 3. *We wish to mention here that the expression on the right-hand side of* (23) *for* $f(A^n, \bar{X}^n | C)$ *represents the likelihood of all observations up to (and including) time* n *"conditioned on" the actions* A^n *up to (and including) time* n*. In other words, a more precise expression for* $f(A^n, \bar{X}^n | C)$ *is as follows:*

$$f(A^n, \bar{X}^n | C) = \left[\prod_{t=0}^{n} P_h(A_t | A^{t-1}, \bar{X}^{t-1}) \right] \prod_{a=1}^{K} \nu(X_0^a) \prod_{i,j \in \mathcal{S}} (P_1(j|i))^{N_h(n,i,j)} \cdot \prod_{i,j \in \mathcal{S}} (P_2(j|i))^{\sum_{a \neq h} N_a(n,i,j)}, \tag{35}$$

where $P_h(A_t | A^{t-1}, \bar{X}^{t-1})$ *represents the probability of selecting arm* A_t *at time* t *when the true hypothesis is* H_h *(i.e., when* h *is the index of the odd arm), with the convention that at time* $t = 0$*, this term represents* $P_h(A_0)$*. Note that for any policy (see description in the paragraph containing* (14) *and* (15)*), this must be independent of the true hypothesis* H_h*, and is thus the same for any two hypotheses* H_h *and* $H_{h'}$*, where* $h' \neq h$*.*

As a consequence of this, the first term within square brackets on the right-hand side of (35) *appears in both the numerator and the denominator terms of the modified GLR statistic of* (29)*, and thus cancels out. Hence, we omit writing this term in the expressions of* (23)*,* (25) *and* (28)*.*

B. The Policy $\pi^\star(L, \delta)$

With the above ingredients in place, we now describe our policy based on the modified GLR statistic of (29). For each $h \in \mathcal{A}$, let
$$M_h(n) \coloneqq \min_{h' \neq h} M_{hh'}(n) \tag{36}$$
denote the modified GLR of hypothesis H_h with respect to its nearest alternative.

Policy $\pi^\star(L, \delta)$:
Fix parameters $L \geq 1$ 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 $n \in \{0, 1, 2, \ldots\}$. We choose each of the K arms once in the first K time steps $n = 0, \ldots, K-1$. For each $n \geq K-1$, at time n, we follow the procedure described below:

1) Let $h^*(n) = \arg\max_{h \in \mathcal{A}} M_h(n)$ be the index with the largest modified GLR after n time steps. We resolve ties uniformly at random.
2) If $M_{h^*(n)}(n) < \log((K-1)L$, then we choose the next arm A_{n+1} based on the sequence $(A^n, \bar{X}^n)$ of observations and arms selected until time n as per the following rule:
   a) If $B_{n+1} = 1$, then we choose an arm uniformly at random.
   b) If $B_{n+1} = 0$, then we choose A_{n+1} according to the distribution $\lambda_{opt}(h^*(n), \hat{P}^n_{h^*(n),1}, \hat{P}^n_{h^*(n),2})$.
3) If $M_{h^*(n)}(n) \geq \log((K-1)L)$, then we stop selecting arms and declare $h^*(n)$ as the true index of the odd arm.

In the above policy, $h^*(n)$ provides the best estimate of the odd arm at time 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 this indicates that the policy is confident that $h^*(n)$ is the odd arm. At this stage, the policy stops taking further samples and declares $h^*(n)$ as the index of the odd arm. If not, the policy 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 [8]. Based on the description in items (2(a)) and (2(b)) above, it follows that for each $a \in \mathcal{A}$,

$$\begin{aligned} P(A_{n+1} = a | A^n, \bar{X}^n) &= \frac{\delta}{K} + (1-\delta)\, \lambda_{opt}(h^*(n), \hat{P}^n_{h^*(n),1}, \hat{P}^n_{h^*(n),2})(a) \\ &\geq \frac{\delta}{K} > 0. \end{aligned} \tag{37}$$

As we will see, the strictly positive lower bound in (37) will ensure that the policy selects each arm at a non-trivial frequency so as to allow for sufficient exploration of all arms. Also, we will show that the parameters L and δ may be selected so that our policy achieves a desired target error probability, while also ensuring that the normalised expected stopping time of the policy is arbitrarily close to the lower bound in (17).

Remark 4. *Evaluating the distribution $\lambda_{opt}(h^*(n), \hat{P}^n_{h^*(n),1}, \hat{P}^n_{h^*(n),2})$ in step (2(a)) of the policy involves solving the maximisation problem in* (18) *with the transition matrices P_1 and P_2 replaced by their corresponding ML estimates $\hat{P}^n_{h^*(n),1}$ and $\hat{P}^n_{h^*(n),2}$ respectively at each time $n \geq K-1$ until stoppage. In the event when any of the rows of the estimated matrices has all its entries as zero, we substitute the corresponding zero row by a row with a single '1' in one of the $|S|$ positions picked uniformly at random. Since the ML estimates converge to their respective true values as more observations are accumulated, we note that such a substitution operation (or any modification thereof that replaces the all-zero rows by an arbitrary probability vector) needs to be carried out only for finitely many time slots, and does not affect the asymptotic performance of the policy.*

C. Performance of $\pi^\star(L, \delta)$

In this subsection, we show that the expected number of samples required by policy $\pi^\star(L, \delta)$ to identify the index of the odd arm can be made arbitrarily close to that in (17) in the regime of vanishing error probabilities. We show that this can be achieved by choosing the parameters L and δ carefully. We organise this subsection as follows:

1) First, we show that when the true index of the odd arm is h, the modified GLR $M_h(n)$ of hypothesis H_h with respect to its nearest alternative has a strictly positive drift under our policy. We then use this to show that our policy stops in finite time with probability 1.
2) For any fixed target error probability $\epsilon > 0$, we show that for an appropriate choice of the threshold parameter L, our policy belongs to the family $\Pi(\epsilon)$, i.e., its probability of error at stoppage is within ϵ.
3) We obtain an upper bound on the expected stopping time of our policy, and demonstrate that this upper bound may be made arbitrarily close to the lower bound in (17) by choosing an appropriate value of $\delta \in (0,1)$.

1) Strictly Positive Drift of the Modified GLR Statistic: The main result on the strictly 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 a version of the policy $\pi^\star(L, \delta)$ that never stops. Let $C = (h, P_1, P_2)$ be the underlying configuration of the arms. Then, for all $h' \neq h$, under the non-stopping version of our policy, we have*

$$\liminf_{n \to \infty} \frac{M_{hh'}(n)}{n} > 0. \tag{38}$$

□

The proof is based on the key idea that forced exploration with parameter $\delta \in (0,1)$ (see items (2(a)) and (2(b)) of policy $\pi^\star(L, \delta)$) results in sampling each arm with a strictly positive rate that grows linearly. It is in showing an analogue of Proposition 2 for iid Poisson observations that the authors of [5] use their result of [5, Proposition 3] on guaranteed exploration at a strictly positive rate. Since it is not clear if the analogue of [5, Proposition 3] holds in general, we use the idea in [8] of forced exploration. We present the details in Section VII-B. We refer the reader to [9] on how to make do with forced exploration at a sublinear rate.

As an immediate consequence of the above proposition, we have the following: suppose $C = (h, P_1, P_2)$ is the underlying configuration of the arms. Then, a.s.,

$$\liminf_{n\to\infty} M_h(n) = \liminf_{n\to\infty} \min_{h'\neq h} M_{hh'}(n) > 0. \tag{39}$$

The result in (39) has the following implication. For any $h' \neq h$, we have the following set of inequalities holding a.s.:

$$\begin{aligned}\limsup_{n\to\infty} M_{h'}(n) &= \limsup_{n\to\infty} \min_{a\neq h'} M_{h'a}(n)\\ &\leq \limsup_{n\to\infty} M_{h'h}(n)\\ &= \limsup_{n\to\infty} -M_{hh'}(n)\\ &= -\liminf_{n\to\infty} M_{hh'}(n)\\ &\leq -\liminf_{n\to\infty} M_h(n)\\ &< 0.\end{aligned} \tag{40}$$

From the above set of inequalities, it follows that under policy $\pi^\star(L,\delta)$,

$$h^*(n) = \arg\max_{h\in\mathcal{A}} M_h(n) = h \text{ a.s.} \tag{41}$$

for all sufficiently large values of n.

We note here that when $C = (h, P_1, P_2)$ is the underlying configuration of the arms, (41) seems to suggest that policy $\pi^\star(L,\delta)$ a.s. outputs h as the true index of the odd arm at the time of stopping, thereby implying that it commits no error a.s. However, we wish to remark that this is not true, and recognise the possibility of the event that $h^*(n) = h' \neq h$ and $M_{h^*(n)}(n) \geq \log((K-1)L)$ for some n, in which case the policy stops at time n and outputs h' as the index of the odd arm, thereby making error. While we shall soon demonstrate that the probability of occurrence of such an error event under our policy is small, we leverage the implication of (41) to define a version of our policy that, under the underlying configuration $C = (h, P_1, P_2)$, waits until the event $M_h(n) \geq \log((K-1)L)$ occurs, at which point it stops and declares h as the index of the odd arm. We denote this version by $\pi_h^\star(L,\delta)$. Thus, $\pi_h^\star(L,\delta)$ stops only at declaration h.

It then follows that the stopping times of policies $\pi^\star(L,\delta)$ and $\pi_h^\star(L,\delta)$ are a.s. related as $\tau(\pi_h^\star(L,\delta)) \geq \tau(\pi^\star(L,\delta))$, as a consequence of which we have the following set of inequalities holding a.s.:

$$\begin{aligned}\tau(\pi^\star(L,\delta)) \leq \tau(\pi_h^\star(L,\delta)) &= \inf\{n \geq 1 : M_h(n) \geq \log((K-1)L)\}\\ &\leq \inf\left\{n \geq 1 : M_{hh'}(n') \geq \log((K-1)L) \text{ for all } n' \geq n \text{ and for all } h' \neq h\right\}\\ &< \infty,\end{aligned} \tag{42}$$

where the last line follows as a consequence of Proposition 2. This establishes that a.s. policy $\pi^\star(L,\delta)$ stops in finite time.

2) Probability of Error of Policy $\pi^\star(L,\delta)$: We now show that for policy $\pi^\star(L,\delta)$, the threshold parameter L may be chosen to achieve any desired target error probability. This is formalised in the proposition below.

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

The proof uses Proposition 2 and the fact that policy $\pi^\star(L,\delta)$ stops a.s. in finite time. Further, the average in the numerator of the modified GLR statistic, in place of the maximum in the classical GLR statistic, plays a role. For details, see Section VII-C.

3) Upper Bound on the Expected Stopping Time of Policy $\pi^\star(L,\delta)$: We conclude this section by presenting an upper bound on the expected stopping time of the policy $\pi^\star(L,\delta)$. We show that this upper bound may be made arbitrarily close to the lower bound in (17) by tuning δ appropriately.

As a first step, we show that under the non-stopping version of policy $\pi^\star(L,\delta)$, when $C = (h, P_1, P_2)$ is the underlying configuration of the arms, the modified GLR process has an asymptotic drift that is close to $D^*(h, P_1, P_2)$ that appears in the lower bound (17).

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

$$\lim_{n\to\infty} \frac{M_{hh'}(n)}{n} = D_\delta^*(h, P_1, P_2) \quad a.s., \tag{43}$$

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

$$D^*_\delta(h, P_1, P_2) = \lambda^*_\delta\, D(P_1||P_\delta|\mu_1) + (1-\lambda^*_\delta)\frac{(K-2)}{(K-1)} D(P_2||P_\delta|\mu_2), \tag{44}$$

with $\lambda^*_\delta = \frac{\delta}{K} + (1-\delta)\lambda^* \in [0,1]$*, and for each* $i, j \in \mathcal{S}$*,* $P_\delta(j|i)$ *is as in (19) with* λ_1 *replaced by* λ^*_δ*.* □

We 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 (43) , 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 $D^*_\delta(h, P_1, P_2)$. This is the notion of certainty equivalence in control theory. The details of the proof may be found in Section VII-D.

We now state the main result of this section.

Proposition 5. *Let* $C = (h, P_1, P_2)$ *denote the underlying configuration of the arms. Fix parameters* $L \geq 1$ *and* $\delta \in (0,1)$*. Then, under policy* $\pi = \pi^\star(L,\delta)$*, we have*

$$\limsup_{L\to\infty} \frac{E^\pi[\tau(\pi)|C]}{\log L} \leq \frac{1}{D^*_\delta(h, P_1, P_2)}. \tag{45}$$

□

The proof uses Proposition 4 and involves showing that (a) the stopping time $\tau(\pi)$ satisfies an asymptotic almost sure upper bound that matches with the right-hand side of (45), and (b) the family $\{\tau(\pi^\star(L,\delta))/\log L : L \geq 1\}$ is uniformly integrable. The almost sure convergence together with uniform integrability then yields the relation (45). The details may be found in Section VII-E.

It is clear that $D^*_\delta(h, P_1, P_2)$ is a continuous function of δ, with the property that

$$\lim_{\delta\downarrow 0} D^*_\delta(h, P_1, P_2) = D^*(h, P_1, P_2), \tag{46}$$

where $D^*(h, P_1, P_2)$ on the right-hand side of (46) is the same the constant that appears in the lower bound of (17). Thus, we note that δ may be tuned to make $D^*_\delta(h, P_1, P_2)$ as close as desired to $D^*(h, P_1, P_2)$, hence establishing the near-optimality of the policy $\pi^\star(L,\delta)$.

V. The Main Result

We now present the main result of this paper, combining the lower and upper bounds stated in Section III and Section IV respectively.

Theorem 1. *Consider* $K \geq 3$ *independent Markov processes on a common finite state space that are irreducible, aperiodic and time homogeneous. Suppose that* $C = (h, P_1, P_2)$ *is the underlying configuration of the arms, where* h *denotes the index of the odd arm, and* $P_2 \neq P_1$*. Let* $(\epsilon_n)_{n\geq 1}$ *denote a sequence of error probability values with the property that* $\epsilon_n \to 0$ *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, we have*

$$\liminf_{n\to\infty} \inf_{\pi\in\Pi(\epsilon_n)} \frac{E[\tau(\pi)|C]}{\log L_n} = \lim_{\delta\downarrow 0} \lim_{n\to\infty} \frac{E[\tau(\pi^\star(L_n,\delta))|C]}{\log L_n} = \frac{1}{D^*(h, P_1, P_2)}. \tag{47}$$

□

Proof: From Proposition 1, it follows that the expected stopping time of any policy $\pi \in \Pi(\epsilon_n)$ grows as $(\log L_n)/D^*(h, P_1, P_2)$ for large values of n. Also, from Proposition 3, policy $\pi^\star(L_n,\delta)$ belongs to the family $\Pi(\epsilon_n)$ and, from Proposition 5, achieves an asymptotic growth of at most $(\log L_n)/D^*_\delta(h, P_1, P_2)$. Since $\lim_{\delta\downarrow 0} D^*_\delta(h, P_1, P_2) = D^*(h, P_1, P_2)$, we may approach the lower bound by choosing an arbitrarily small value of δ. This establishes the theorem. ■

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

VI. Simulation Results

Fix $K = 8$ and $C = (h, P_1, P_2)$, with $h = 1$ and

$$P_1 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix}.$$

Fig. 1 depicts the average stopping time of policy $\pi^\star(L,\delta)$ as a function of $\log L$, averaged over 100 rounds of iterations, for $\delta = 0.01, 0.1, 0.25$. For the aforementioned values of P_1 and P_2, numerical evaluation yields $D^*(h, P_1, P_2) \simeq 0.094$, thus

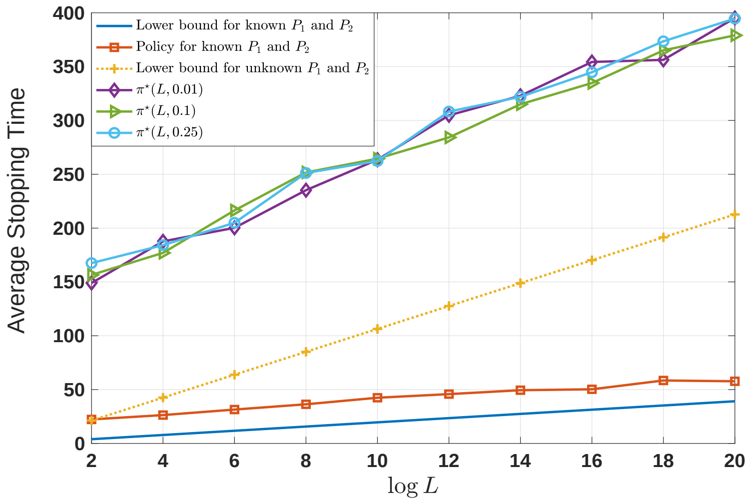

Fig. 1. Plots of average stopping time of policy $\pi^\star(L, \delta)$, as function of $\log L$, for $\delta = 0.01, 0.1, 0.25$.

resulting in a lower bound of $1/D^*(h, P_1, P_2) \simeq 10.635$. Since (17) is a statement about the slope of the growth rate of average stopping time of policy $\pi^\star(L, \delta)$ as a function of $\log L$, the top 3 plots in the figure respect the lower bound in (17), with the slopes in these plots only marginally higher than that given by the lower bound. Theory predicts that as $\delta \downarrow 0$ and $L \to \infty$, the slopes will approach the lower bound. Also included in the figure are the plots of (a) the lower bound for the case when P_1 and P_2 are known, and (b) a policy similar to that of $\pi^\star(L, \delta)$ that uses the knowledge of P_1 and P_2 to identify the index of the odd arm. Such a policy clearly takes lesser time than $\pi^\star(L, \delta)$ to identify the index of the odd arm. The figure shows that the performance of this policy also matches in slope to that given by its lower bound for large values of L.

VII. Proofs of the Main Results

A. Proof of Proposition 1

We first present below 3 lemmas that will be used in the proof of the proposition. The first of these, given below, is an analogue of the change of measure argument of Kaufmann et al. [10, Lemma 18] for the case of Markov observations from each arm.

Recall the definition of $\mathcal{F}_\tau$ in (13):

$$\mathcal{F}_\tau = \{E \in \mathcal{F} : E \cap \{\tau = n\} \in \mathcal{F}_n \text{ for all } n \geq 0\},$$

where for each n, $\mathcal{F}_n$ is as defined in (12). Further, for any $h' \neq h$, define $Z_{hh'}(\tau) \coloneqq Z_h(\tau) - Z_{h'}(\tau)$, where $Z_h(\tau) = \sum_{a=1}^{K} Z_h^a(\tau)$.

Lemma 1. *Fix $\epsilon > 0$ and transition matrices P_1 and P_2, and let τ be the stopping time of a policy $\pi \in \Pi(\epsilon)$. Then, for any event $E \in \mathcal{F}_\tau$ and configuration triplets $C = (h, P_1, P_2)$ and $C' = (h', P_1', P_2')$, with $h' \neq h$, we have*

$$P^\pi(E|C') = E^\pi[1_E \ \exp(-Z_{hh'}(\tau))|C]. \tag{48}$$

□

Proof: The proof follows the outline in [10], with crucial modifications needed for the Markov problem at hand. We use the shorthand notations $E_h[\cdot]$ and $E_{h'}[\cdot]$ to denote respectively the quantities $E^\pi[\cdot|C]$ and $E^\pi[\cdot|C']$; similarly, $P_h(\cdot)$ and $P_{h'}(\cdot)$ denote the respective probabilities. We begin by showing that for all $n \geq 0$, the following statement is true: for any measurable function $g : \mathcal{A}^{n+1} \times \mathcal{S}^{n+1} \to \mathbb{R}$, we have

$$E_{h'}[g(A^n, \bar{X}^n)] = E_h[g(A^n, \bar{X}^n) \ \exp(-Z_{hh'}(n))]. \tag{49}$$

Assuming that the above statement is true, for any $E \in \mathcal{F}_\tau$, we have

$$\begin{aligned}
P_{h'}(E) &= E_{h'}[1_E] \\
&\stackrel{(a)}{=} \sum_{n=0}^{\infty} E_{h'}[1_E 1_{\{\tau=n\}}] \\
&\stackrel{(b)}{=} \sum_{n=0}^{\infty} E_h[1_E 1_{\{\tau=n\}} \exp(-Z_{hh'}(n))] \\
&= E_h[1_E \ \exp(-Z_{hh'}(\tau))],
\end{aligned} \tag{50}$$

hence proving the lemma. In the above set of equations, (a) is due to monotone convergence theorem, and (b) follows from the application of (49) to the function $g(A^n, \bar{X}^n) = 1_E \cdot 1_{\{\tau=n\}}$ by noting that $E \in \mathcal{F}_\tau$, and therefore $E \cap \{\tau = n\} \in \mathcal{F}_n$ for all n.

We now proceed to prove (49) by induction on n. From (11) and (9), it follows that $Z_{hh'}(0) = 0$. Then, for any measurable function $g : \mathcal{A}^{n+1} \times \mathcal{S}^{n+1} \to \mathbb{R}$, the proof of (49) for $n = 0$ follows from the following set of equations.

$$\begin{aligned}
E_{h'}[g(A_0, \bar{X}_0)] &= \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} P_{h'}(A_0 = a) \cdot P_{h'}(\bar{X}_0 = i | A_0 = a) \cdot g(a, i) \\
&= \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} P_{h'}(A_0 = a) \cdot P_{h'}(X_0^a = i) \cdot g(a, i) \\
&= \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} P_{h'}(A_0 = a) \cdot \nu(i) \cdot g(a, i) \\
&\stackrel{(a)}{=} \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} P_h(A_0 = a) \cdot P_h(X_0^a = i) \cdot g(a, i) \\
&= E_h[g(A_0, \bar{X}_0)] \\
&= E_h[g(A_0, \bar{X}_0) \exp(-Z_{hh'}(0))],
\end{aligned} \tag{51}$$

where in writing (a), we use

- the fact that $P_h(A_0 = a) = P_{h'}(A_0 = a)$ since the manner in which A_0 is selected is not a function of either h or h'. For instance, we may assume that each of the arms is picked once in the first K time instants, and note that this does not affect the asymptotic performance of the policy. In such a case, $P_h(A_0 = 1) = 1 = P_{h'}(A_0 = 1)$.
- the fact that $X_0^a \sim \nu$ under hypotheses H_h and $H_{h'}$.

We now assume that (49) holds for some positive integer n, and show that it also holds for $n + 1$. We have

$$E_{h'}[g(A^{n+1}, \bar{X}^{n+1})] = E_{h'}\left[E_{h'}\left[g(A^{n+1}, \bar{X}^{n+1}) | A^n, \bar{X}^n\right]\right]. \tag{52}$$

Since the inner conditional expectation term on the right-hand side of (52) is a measurable function of $(A^n, \bar{X}^n)$, using the induction hypothesis, we get

$$\begin{aligned}
&E_{h'}[g(A^{n+1}, \bar{X}^{n+1})] \\
&= E_h\left[E_{h'}\left[g(A^{n+1}, \bar{X}^{n+1}) | A^n, \bar{X}^n\right] \exp(-Z_{hh'}(n))\right] \\
&= \sum_{a^n \in \mathcal{A}^n} \sum_{\bar{x}^n \in \mathcal{S}^{n+1}} P_h(A^n = a^n, \bar{X}^n = \bar{x}^n) \cdot \exp(-z_{hh'}(n)) \cdot E_{h'}[g(A^{n+1}, \bar{X}^{n+1}) | A^n = a^n, \bar{X}^n = \bar{x}^n],
\end{aligned} \tag{53}$$

where $z_{hh'}(n)$ denotes the value of $Z_{hh'}(n)$ when $A^n = a^n$ and $\bar{X}^n = \bar{x}^n$. Then, we have

$$\begin{aligned}
&E_{h'}[g(A^{n+1}, \bar{X}^{n+1}) | A^n = a^n, \bar{X}^n = \bar{x}^n] \\
&= \sum_{a'=1}^{K} \sum_{j \in \mathcal{S}} g(a^n, a', \bar{x}^n, j) \cdot P_{h'}(A_{n+1} = a' | A^n = a^n, \bar{X}^n = \bar{x}^n) \cdot P_{h'}^{a'}(X_{N_{a'}(n)}^{a'} = j | X_{N_{a'}(n)-1}^{a'}) \\
&= \sum_{a'=1}^{K} \sum_{j \in \mathcal{S}} g(a^n, a', \bar{x}^n, j) \cdot P_h(A_{n+1} = a' | A^n = a^n, \bar{X}^n = \bar{x}^n) \cdot P_{h'}^{a'}(X_{N_{a'}(n)}^{a'} = j | X_{N_{a'}(n)-1}^{a'}),
\end{aligned} \tag{54}$$

where in writing the last line above, we use the fact that the probability of selecting an arm at any time, based on the history of past arm selections and observations, is independent of the underlying configuration of the arms, and is thus the same under hypotheses H_h and $H_{h'}$. We now write (54) as

$$\begin{aligned}
&E_{h'}[g(A^{n+1},\bar{X}^{n+1})|A^n=a^n,\bar{X}^n=\bar{x}^n] \\
&=\sum_{a'=1}^{K}\sum_{j\in\mathcal{S}}\Bigg\{g(a^n,a',\bar{x}^n,j)\cdot P_h(A_{n+1}=a'|A^n=a^n,\bar{X}^n=\bar{x}^n) \\
&\qquad\cdot\frac{P_{h'}^{a'}(X^{a'}_{N_{a'}(n)-1}=j|X^{a'}_{N_{a'}(n)-1})}{P_h^{a'}(X^{a'}_{N_{a'}(n)}=j|X^{a'}_{N_{a'}(n)-1})}\cdot P_h^{a'}(X^{a'}_{N_{a'}(n)}=j|X^{a'}_{N_{a'}(n)-1})\Bigg\}.
\end{aligned}\tag{55}$$

Plugging back (55) in (53), and using

$$z_{hh'}(n+1)=z_{hh'}(n)+\log\frac{P_h^{a'}(X^{a'}_{N_{a'}(n)}=j|X^{a'}_{N_{a'}(n)-1})}{P_{h'}^{a'}(X^{a}_{N_{a'}(n)}=j|X^{a'}_{N_{a'}(n)-1})},\tag{56}$$

we get

$$\begin{aligned}
&E_{h'}[g(A^{n+1},\bar{X}^{n+1})] \\
&=\sum_{a^n\in\mathcal{A}^n}\sum_{\bar{x}^n\in\mathcal{S}^{n+1}}\sum_{a'=1}^{K}\sum_{j\in\mathcal{S}}\Bigg\{g(a^n,a',\bar{x}^n,j)\cdot\exp(-z_{hh'}(n+1)) \\
&\qquad\cdot P_h(A^n=a^n,\bar{X}^n=\bar{x}^n)\cdot P_h(A_{n+1}=a',\bar{X}_{n+1}=j|A^n=a^n,\bar{X}^n=\bar{x}^n)\Bigg\} \\
&=E_h[g(A^{n+1},\bar{X}^{n+1})\exp(-Z_{hh'}(n+1))],
\end{aligned}\tag{57}$$

hence proving (48) . ■

The second lemma below relates the expected number of i to j transitions $E^\pi[N_a(\tau,i,j)|C]$ observed on the Markov process of arm a to $E^\pi[N_a(\tau,i)|C]$, the expected number of exits out of state i observed on the Markov process of arm a.

Lemma 2. *Fix $\epsilon>0$, a policy $\pi\in\Pi(\epsilon)$, and a configuration $C=(h,P_1,P_2)$. For each $i,j\in\mathcal{S}$ and $a\in\mathcal{A}$, we have*

$$E^\pi[N_a(\tau,i,j)|C]=E^\pi[N_a(\tau,i)|C]\cdot P_h^a(j|i),\tag{58}$$

where $P_h^a(j|i)$ is as given in (7). □

Proof: We use the shorthand notation $E_h[\cdot]$ to denote $E^\pi[\cdot|C]$. We demonstrate that for each $i,j\in\mathcal{S}$ and $a\in\mathcal{A}$,

$$E_h[E_h[N_a(\tau,i,j)|X_0^a]|N_a(\tau)]=E_h[E_h[N_a(\tau,i)|X_0^a]|N_a(\tau)]\cdot P_h^a(j|i).\tag{59}$$

Towards this, we note that

$$E_h[E_h[N_a(\tau,i,j)|X_0^a]|N_a(\tau)]=E_h\left[\sum_{m=1}^{N_a(\tau)-1}E_h[1_{\{X^a_{m-1}=i,\,X^a_m=j\}}|X_0^a]\middle|N_a(\tau)\right].\tag{60}$$

We now simplify the inner conditional expectation term in (60) by considering the cases $m=1$ and $m\geq 2$ separately.

1) Case $m=1$: In this case, we get

$$\begin{aligned}
E_h[1_{\{X^a_0=i,\,X^a_1=j\}}|X_0^a]&=1_{\{X^a_0=i\}}\cdot E_h[1_{\{X^a_1=j\}}|X_0^a] \\
&=1_{\{X^a_0=i\}}\cdot P_h^a(X_1^a=j|X_0^a=i) \\
&=1_{\{X^a_0=i\}}\cdot P_h^a(j|i).
\end{aligned}\tag{61}$$

2) Case $m\geq 2$: Here, we get

$$\begin{aligned}
E_h[1_{\{X^a_{m-1}=i,\,X^a_m=j\}}|X_0^a=k]&=P_h^a(X^a_{m-1}=i,\,X^a_m=j|X_0^a=k) \\
&\overset{(a)}{=}P_h^a(X^a_{m-1}=i|X^a_0=k)\cdot P_h^a(X_1^a=j|X_0^a=i) \\
&=E_h[1_{\{X^a_{m-1}=i\}}|X_0^a=k]\cdot P_h^a(j|i),
\end{aligned}\tag{62}$$

from which it follows that $E_h[1_{\{X^a_{m-1}=i,\,X^a_m=j\}}|X_0^a]=E_h[1_{\{X^a_{m-1}=i\}}|X_0^a]\cdot P_h^a(j|i)$. In the above set of equations, (a) follows from the fact that the Markov process of arm a is time homogeneous.

From the aforementioned cases, it follows that the relation

$$E_h[1_{\{X^a_{m-1}=i,\,X^a_m=j\}}|X_0^a]=E_h[1_{\{X^a_{m-1}=i\}}|X_0^a]\cdot P_h^a(j|i)\tag{63}$$

holds for all $m \geq 1$. Substituting (63) in (60) and simplifying, we arrive at (59). The lemma then follows by applying expectation $E_h[\cdot]$ to both sides of (59). ∎

The third lemma lemma presented below will be used to simplify a minimisation term later in the proof of the proposition.

Lemma 3. *Denote by $\mathcal{P}(\mathcal{S})$ the set of all probability distributions on the set $\mathcal{S}$, and let ν_1 and ν_2 be any two distinct elements of $\mathcal{P}(\mathcal{S})$. Then, for all $w_1, w_2 \in [0,1]$ such that $w_1 + w_2 = 1$, we have*

$$\min_{\psi \in \mathcal{P}(\mathcal{S})} [w_1 D(\nu_1 || \psi) + w_2 D(\nu_2 || \psi)] = w_1 D(\nu_1 || \nu^*) + w_2 D(\nu_2 || \nu^*), \tag{64}$$

where $\nu^* \in \mathcal{P}(\mathcal{S})$ is given by $\nu^* = w_1 \nu_1 + w_2 \nu_2$. □

Proof: This is well known with ν^* viewed as a root of "information centre" and the right-hand side of (64) viewed as a mutual information. Here is the proof for completeness.

Let ν^* be as defined in the statement of the lemma. For any $\psi \in \mathcal{P}(\mathcal{S})$, we have

$$\begin{aligned} w_1 D(\nu_1||\psi) + w_2 D(\nu_2||\psi) &= w_1 E_{\nu_1}\left[\log \frac{d\nu_1}{d\psi}\right] + w_2 E_{\nu_2}\left[\log \frac{d\nu_2}{d\psi}\right] \\ &= w_1 E_{\nu_1}\left[\log \frac{d\nu_1}{d\nu^*}\right] + w_2 E_{\nu_2}\left[\log \frac{d\nu_2}{d\nu^*}\right] + w_1 E_{\nu_1}\left[\log \frac{d\nu^*}{d\psi}\right] + w_2 E_{\nu_2}\left[\log \frac{d\nu^*}{d\psi}\right] \\ &= w_1 E_{\nu_1}\left[\log \frac{d\nu_1}{d\nu^*}\right] + w_2 E_{\nu_2}\left[\log \frac{d\nu_2}{d\nu^*}\right] + E_{\nu^*}\left[\log \frac{d\nu^*}{d\psi}\right] \\ &= w_1 E_{\nu_1}\left[\log \frac{d\nu_1}{d\nu^*}\right] + w_2 E_{\nu_2}\left[\log \frac{d\nu_2}{d\nu^*}\right] + D(\nu^*||\psi) \\ &\geq w_1 E_{\nu_1}\left[\log \frac{d\nu_1}{d\nu^*}\right] + w_2 E_{\nu_2}\left[\log \frac{d\nu_2}{d\nu^*}\right], \end{aligned} \tag{65}$$

with equality in the last line above if and only if $\psi = \nu^*$. This completes the proof of the lemma. ∎

Proof of Proposition 1: Fix an arbitrary $\epsilon > 0$, and let $\pi \in \Pi(\epsilon)$ be a policy whose stopping is $\tau = \tau(\pi)$. Without loss of generality, we assume that $E^\pi[\tau(\pi)|C] < \infty$, for otherwise the inequality (17) holds trivially. We organise the proof of the proposition into various sections. In the first of these sections presented below, we lower bound the expected value of $Z_{hh'}(\tau)$ in terms of the error probability ϵ. This uses the above Lemma 1, Lemma 2 and the result of [10, Lemma 19].

1) A Lower Bound on The Expected Value of $Z_{hh'}(\tau)$: Let $\pi \in \Pi(\epsilon)$, with stopping time is $\tau = \tau(\pi)$. For any $h' \neq h$, let $Z_{hh'}(\tau)$ be as defined in the statement of Lemma 1. Then, Lemma 1 in conjunction with [10, Lemma 19] yields the following: conditioned on the underlying configuration $C = (h, P_1, P_2)$, for any alternative configuration $C' = (h', P_1', P_2')$, where $h' \neq h$, under the assumption that $E^\pi[\tau|C] < \infty$, we have

$$E^\pi[Z_{hh'}(\tau)|C] \geq \sup_{E \in \mathcal{F}_\tau} d(P^\pi(E|C), P^\pi(E|C')), \tag{66}$$

where

$$d(p,q) \coloneqq p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$$

denotes the binary KL divergence, with the convention that $d(0,0) = 0 = d(1,1)$. We now note the following points:

1) For each alternative configuration C', by taking $E = \{I(\pi) = h\}$ and recognising that $\pi \in \Pi(\epsilon)$, we have $P^\pi(E|C) > 1-\epsilon$ and $P^\pi(E|C') \leq \epsilon$. Using this, along with the fact that the mapping $x \mapsto d(x,y)$ is monotone increasing for $x < y$ and the mapping $y \mapsto d(x,y)$ is monotone decreasing for any fixed x, we obtain

$$\begin{aligned} d(P^\pi(E|C), P^\pi(E|C')) &\geq d(1-\epsilon, P^\pi(E|C')) \\ &\geq d(1-\epsilon, \epsilon). \end{aligned} \tag{67}$$

2) We may minimise both sides of (66) over all alternative configurations C' to obtain

$$\min_{C'=(h',P_1',P_2')} E^\pi[Z_{hh'}(\tau)|C] \geq \min_{C'=(h',P_1',P_2')} \sup_{E \in \mathcal{F}_\tau} d(P^\pi(E|C), P^\pi(E|C')). \tag{68}$$

Combining the points noted above, and using $d(1-\epsilon,\epsilon) = d(\epsilon, 1-\epsilon)$, we obtain

$$\min_{C'=(h',P_1',P_2')} E^\pi[Z_{hh'}(\tau)|C] \geq d(\epsilon, 1-\epsilon). \tag{69}$$

2) A Relation Between $E^\pi[Z_{hh'}(\tau)|C]$ and $E^\pi[\tau|C]$: As our next step, we obtain an upper bound for $E^\pi[Z_{hh'}(\tau)|C]$ in terms of $E^\pi[\tau|C]$. Towards this, we have

$$E^\pi[Z_{hh'}(\tau)|C] = \sum_{a=1}^{K} E^\pi\left[\sum_{m=1}^{N_a(\tau)-1} \log\left(\frac{P_h^a(X_m^a|X_{m-1}^a)}{P_{h'}^a(X_m^a|X_{m-1}^a)}\right)\bigg|C\right], \tag{70}$$

where we take inner summation term to be zero whenever $N_a(\tau) < 2$. Focus on the expectation term in (70). This term may be written as

$$\begin{aligned}E^\pi\left[\sum_{m=1}^{N_a(\tau)-1} \log\left(\frac{P_h^a(X_m^a|X_{m-1}^a)}{P_{h'}^a(X_m^a|X_{m-1}^a)}\right)\bigg|C\right] &\overset{(a)}{=} E^\pi\left[\sum_{m=1}^{N_a(\tau)-1} \sum_{i,j\in S} 1_{\{X_{m-1}^a=i,\, X_m^a=j\}} \log\left(\frac{P_h^a(j|i)}{P_{h'}^a(j|i)}\right)\bigg|C\right]\\ &= \sum_{i,j\in S} E^\pi[N_a(\tau,i,j)|C]\, f_{hh'}^a(j|i),\end{aligned} \tag{71}$$

where (a) above follows from the fact that the Markov process of arm a is time homogeneous, and $f_{hh'}^a(j|i) \coloneqq \log\left(\frac{P_h^a(j|i)}{P_{h'}^a(j|i)}\right)$. Using the result of Lemma 2 in (71), we get

$$\begin{aligned}E^\pi[Z_{hh'}(\tau)|C] &= \sum_{a=1}^{K} \sum_{i,j\in S} E^\pi[N_a(\tau,i)|C]\cdot P_h^a(j|i)\cdot f_{hh'}^a(j|i)\\ &= \sum_{a=1}^{K} \sum_{i\in S} E[N_a(\tau,i)|C]\, D(P_h^a(\cdot|i)||P_{h'}^a(\cdot|i)),\end{aligned} \tag{72}$$

where $D(P_h^a(\cdot|i)||P_{h'}^a(\cdot|i)) = \sum_{j\in S} P_h^a(j|i) f_{hh'}^a(j|i)$ denotes the KL divergence between the probability distributions $P_h^a(\cdot|i)$ and $P_{h'}^a(\cdot|i)$. We now express (72) by introducing some additional terms as below:

$$\begin{aligned}&E^\pi[Z_{hh'}(\tau)|C]\\ &= (E^\pi[\tau+1|C]-K)\left(\sum_{a=1}^{K}\left[\frac{E^\pi[N_a(\tau)|C]-1}{E^\pi[\tau+1|C]-K}\right]\sum_{i\in\mathcal{S}}\left[\frac{E^\pi[N_a(\tau,i)|C]}{E^\pi[N_a(\tau)|C]-1}\right]D(P_h^a(\cdot|i)||P_{h'}^a(\cdot|i))\right)\\ &= (E^\pi[\tau+1|C]-K)\left(\sum_{a=1}^{K}\left[\frac{E^\pi[N_a(\tau)|C]-1}{E^\pi[\tau+1|C]-K}\right]\sum_{i\in\mathcal{S}} p_h^a(i)\cdot D(P_h^a(\cdot|i)||P_{h'}^a(\cdot|i))\right),\end{aligned} \tag{73}$$

where $p_h^a(i) \coloneqq \frac{E^\pi[N_a(\tau,i)|C]}{E^\pi[N_a(\tau)|C]-1}$ represents the average (computed with respect to $E^\pi[\cdot|C]$) fraction of times a transition out of state i is observed on the Markov process of arm a.

3) Asymptotics of Vanishing Error Probability: Since $\sum_{i\in\mathcal{S}} p_h^a(i) = 1$, the inner summation term over i in (73) represents the average of the numbers $(D(P_h^a(\cdot|i)||P_{h'}^a(\cdot|i)))_{i\in\mathcal{S}}$ with respect to $(p_h^a(i))_{i\in\mathcal{S}}$. Suppose that at some time, arm a is selected, and it makes a transition out of a state i and into another state j, for some $i,j\in\mathcal{S}$. Then, the next time arm a is selected, it makes a transition out of state j and into state k for some $k\in\mathcal{S}$. Except for the final selection of arm a, all the previous selections result in an exit out of a state and entry into another state. The final selection contributes only to an entry into some state. For $a\in\mathcal{A}$ and $i\in\mathcal{S}$, let

$$N^a(\tau,i) \coloneqq \sum_{m=2}^{N_a(\tau)} 1_{\{X_{m-1}^a=i\}} \tag{74}$$

denote the number of times an entry *into* state i is observed on the Markov process of arm a. In conjunction with (3), it is easy to see that for each $i\in\mathcal{S}$, we have

$$N_a(\tau,i) = N^a(\tau,i) - 1_{\{X_{N_a(\tau)-1}^a=i\}} + 1_{\{X_0^a=i\}}, \tag{75}$$

which implies that $N^a(\tau,i)-1 \le N_a(\tau,i) \le N^a(\tau,i)+1$ a.s. Thus, we notice that for the Markov process of each arm, the number of times a transition *into* a state $i\in\mathcal{S}$ is observed is at most one more than the number of times a transition out of i is observed. We then have

$$\frac{E^\pi[N^a(\tau,i)|C]-1}{E^\pi[N_a(\tau)|C]-1} \le p_h^a(i) \le \frac{E^\pi[N^a(\tau,i)|C]+1}{E^\pi[N_a(\tau)|C]-1}. \tag{76}$$

Using (76) in (73), we arrive at the form

$$u-\Delta \le E^\pi[Z_{hh'}(\tau)] \le u+\Delta, \tag{77}$$

where the terms u and Δ are as below:

$$u = (E^\pi[\tau+1|C]-K)\bigg(\sum_{a=1}^{K}\left[\frac{E^\pi[N_a(\tau)|C]-1}{E^\pi[\tau+1|C]-K}\right]\sum_{i\in\mathcal{S}}\left[\frac{E^\pi[N^a(\tau,i)|C]}{E^\pi[N_a(\tau)|C]-1}\right]D(P_h^a(\cdot|i)||P_{h'}^a(\cdot|i))\bigg),$$

$$\Delta = \sum_{a=1}^{K}\sum_{i\in\mathcal{S}}D(P_h^a(\cdot|i)||P_{h'}^a(\cdot|i)) = \sum_{i\in\mathcal{S}}D(P_1(\cdot|i)||P_2'(\cdot|i)) + \sum_{i\in\mathcal{S}}D(P_2(\cdot|i)||P_1'(\cdot|i)) + \sum_{a\neq h}\sum_{i\in\mathcal{S}}D(P_2(\cdot|i)||P_2'(\cdot|i)). \tag{78}$$

We shall soon show that the regime of vanishing error probabilities, i.e., $\epsilon\downarrow 0$, necessarily means that $E^\pi[N_a(\tau)|C]\to\infty$, which in turn implies that $E^\pi[\tau|C]\to\infty$. In this asymptotic regime, the limiting probabilities of entry into and exit out of any $i\in\mathcal{S}$ are equal and invariant to the one step transitions on arm a. Since the Markov process of arm a is irreducible and positive recurrent, its transition matrix admits a unique stationary distribution. Therefore, by the Ergodic theorem, the aforementioned probabilities must converge to those given by the stationary distribution associated with arm a. We shall denote this stationary distribution by $\mu_h^a(\cdot)$ under configuration $C=(h,P_1,P_2)$, given by

$$\mu_h^a(i) = \begin{cases}\mu_1(i), & a=h,\\ \mu_2(i), & a\neq h.\end{cases} \tag{79}$$

Then, as $\epsilon\downarrow 0$, we have that both the lower and upper bounds in (76) converge to $\mu_h^a(i)$. We shall soon exploit this fact below to arrive at the lower bound. Going further, we denote by $(q_h^a(i))_{i\in\mathcal{S}}$ the probability distribution given by

$$q_h^a(i) = \frac{E^\pi[N^a(\tau,i)|C]}{E^\pi[N_a(\tau)|C]-1}, \quad i\in\mathcal{S}. \tag{80}$$

Using the upper bound in (77) in combination with (69), we have the following chain of inequalities:

$$\begin{aligned} d(\epsilon,1-\epsilon) &\leq \min_{C'=(h',P_1',P_2')} E^\pi[Z_{hh'}(\tau)|C] \\ &\leq \min_{C'=(h',P_1',P_2')}(u+\Delta) \\ &\leq \min_{C'=(h',P_1',P_2')} u + \min_{C'=(h',P_1',P_2')}\Delta. \end{aligned} \tag{81}$$

The first term in (81) may be upper bounded as follows:

$$\begin{aligned} &\min_{C'=(h',P_1',P_2')} u \\ &= (E^\pi[\tau+1|C]-K)\left\{\min_{C'=(h',P_1',P_2')}\bigg(\sum_{a=1}^{K}\left[\frac{E^\pi[N_a(\tau)|C]-1}{E^\pi[\tau+1|C]-K}\right]\sum_{i\in\mathcal{S}}\left[\frac{E^\pi[N^a(\tau,i)|C]}{E^\pi[N_a(\tau)|C]-1}\right]D(P_h^a(\cdot|i)||P_{h'}^a(\cdot|i))\bigg)\right\} \\ &= (E^\pi[\tau+1|C]-K)\left\{\min_{C'=(h',P_1',P_2')}\bigg(\sum_{a=1}^{K}\left[\frac{E^\pi[N_a(\tau)|C]-1}{E^\pi[\tau+1|C]-K}\right]\sum_{i\in\mathcal{S}}q_h^a(i)\,D(P_h^a(\cdot|i)||P_{h'}^a(\cdot|i))\bigg)\right\} \\ &\overset{(a)}{=} (E^\pi[\tau+1|C]-K)\left\{\min_{C'=(h',P_1',P_2')}\bigg(\sum_{a=1}^{K}\left[\frac{E^\pi[N_a(\tau)|C]-1}{E^\pi[\tau+1|C]-K}\right]D(P_h^a(\cdot|\cdot)||P_{h'}^a(\cdot|\cdot)|q_h^a)\bigg)\right\} \\ &\overset{(b)}{\leq} (E^\pi[\tau+1|C]-K)\left\{\max_{\lambda\in\mathcal{P}(\mathcal{A})}\min_{C'=(h',P_1',P_2')}\bigg(\sum_{a=1}^{K}\lambda(a)D(P_h^a(\cdot|\cdot)||P_{h'}^a(\cdot|\cdot)|q_h^a)\bigg)\right\}, \end{aligned} \tag{82}$$

where, in (a) above,

$$D(P_h^a(\cdot|\cdot)||P_{h'}^a(\cdot|\cdot)|q_h^a) = \sum_{i\in S}q_h^a(i)\cdot D(P_h^a(\cdot|i)||P_{h'}^a(\cdot|i)),$$

while (b) follows by noting that maximising over the set $\mathcal{P}(\mathcal{A})$ of all probability distributions on the set of arms $\mathcal{A}$ only increases the right-hand side. The second term in (81) may be simplified as

$$\begin{aligned} \min_{C'=(h',P_1',P_2')}\Delta &= \min_{P_1',P_2':P_1'\neq P_2'}\left\{\sum_{i\in\mathcal{S}}D(P_1(\cdot|i)||P_2'(\cdot|i)) + \sum_{i\in\mathcal{S}}D(P_2(\cdot|i)||P_1'(\cdot|i)) + \sum_{a\neq h}\sum_{i\in\mathcal{S}}D(P_2(\cdot|i)||P_2'(\cdot|i))\right\} \\ &\overset{(a)}{=} \min_{P_2'}\left\{\sum_{i\in\mathcal{S}}D(P_1(\cdot|i)||P_2'(\cdot|i)) + \sum_{a\neq h}\sum_{i\in\mathcal{S}}D(P_2(\cdot|i)||P_2'(\cdot|i))\right\} \\ &= \min\left\{\sum_{i\in\mathcal{S}}D(P_1(\cdot|i)||P_2(\cdot|i)),\ (K-1)\sum_{i\in\mathcal{S}}D(P_2(\cdot|i)||P_1(\cdot|i))\right\}, \end{aligned} \tag{83}$$

where (a) above follows by noting that P_1' appears only in the term $D(P_2(\cdot|i)||P_1'(\cdot|i))$, and that for the choice $P_1' = P_2$, we get $D(P_2(\cdot|i)||P_1'(\cdot|i)) = 0$ for all $i \in \mathcal{S}$. For ease of notation, we shall denote the quantity in (83) by Δ', which we note is a constant.

Combining (82) with (81), we get the following relation after rearrangement:

$$d(\epsilon, 1-\epsilon) \leq \Delta' + (E^{\pi}[\tau+1|C] - K) \left\{ \max_{\lambda \in \mathcal{P}(\mathcal{A})} \min_{C'=(h', P_1', P_2')} \left[\sum_{a=1}^{K} \lambda(a) D(P_h^a(\cdot|\cdot)||P_{h'}^a(\cdot|\cdot)|q_h^a) \right] \right\}. \tag{84}$$

Since (84) is valid for any arbitrary choice of $\epsilon > 0$ and for all $\pi \in \Pi(\epsilon)$, letting $\epsilon \downarrow 0$ and using $d(\epsilon, 1-\epsilon)/\log\frac{1}{\epsilon} \to 1$ as $\epsilon \downarrow 0$, along with the fact that $q_h^a(i) \to \mu_h^a(i)$ for all $i \in \mathcal{S}$ in the regime of vanishing error probabilities, we get

$$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^{\pi}[\tau(\pi)|C]}{\log\frac{1}{\epsilon}} \geq \frac{1}{D^*(h, P_1, P_2)}, \tag{85}$$

where the quantity $D^*(h, P_1, P_2)$ depends on the underlying configuration of the arms, and is given by

$$D^*(h, P_1, P_2) = \max_{\lambda \in \mathcal{P}(\mathcal{A})} \min_{C'=(h', P_1', P_2')} \left(\sum_{a=1}^{K} \lambda(a)\, D(P_h^a(\cdot|\cdot)||P_{h'}^a(\cdot|\cdot)|\mu_h^a) \right). \tag{86}$$

We now show that the quantities in (86) and (18) are the same.

4) The Final Steps: Using (7) and (79), and using the shorthand notation $D(P_h^a||P_{h'}^a|\mu_h^a)$ to denote the KL divergence term inside the summation in (86), we get

$$D^*(h, P_1, P_2) \quad = \max_{\lambda \in \mathcal{P}(\mathcal{A})} \min_{h' \neq h,\, P_1',\, P_2'} \Big(\lambda(h)\, D(P_1||P_2'|\mu_1) + \lambda(h')\, D(P_2||P_1'|\mu_2) + (1 - \lambda(h) - \lambda(h'))D(P_2||P_2'|\mu_2) \Big). \tag{87}$$

Since P_1' appears only in the second term on right-hand side of the above expression, the minimum over all P_1' of the quantity $D(P_2||P_1'|\mu_2)$ is equal to zero, which is attained for $P_1' = P_2$. Thus, we have

$$D^*(h, P_1, P_2) = \max_{\lambda \in \mathcal{P}(\mathcal{A})} \min_{h' \neq h,\, P_2'} \Big(\lambda(h)\, D(P_1||P_2'|\mu_1) + (1 - \lambda(h) - \lambda(h'))D(P_2||P_2'|\mu_2) \Big). \tag{88}$$

We now note that

$$\begin{aligned} \min_{h' \neq h} (1 - \lambda(h) - \lambda(h')) &= 1 - \lambda(h) - \max_{h' \neq h} \lambda(h') \\ &\overset{(a)}{\leq} 1 - \lambda(h) - \frac{1-\lambda(h)}{K-1} \\ &= (1 - \lambda(h)) \frac{(K-2)}{(K-1)}, \end{aligned} \tag{89}$$

where (a) above follows by lower bounding the maximum of a set of numbers by their arithmetic mean. We then have

$$D^*(h, P_1, P_2) = \max_{0 \leq \lambda(h) \leq 1} \min_{P_2'} \left(\lambda(h)\, D(P_1||P_2'|\mu_1) + (1 - \lambda(h)) \frac{(K-2)}{(K-1)} D(P_2||P_2'|\mu_2) \right). \tag{90}$$

Using Lemma 3 in (90), and recognising that the hand side of (90) is not a function of h, we write

$$D^*(h, P_1, P_2) = \max_{0 \leq \lambda_1 \leq 1} \left(\lambda_1\, D(P_1||P|\mu_1) + (1 - \lambda_1) \frac{(K-2)}{(K-1)} D(P_2||P|\mu_2) \right), \tag{91}$$

where P is a transition probability matrix whose entry in the ith row and jth column is given by

$$P(j|i) = \frac{\lambda_1 \mu_1(i) P_1(j|i) + (1-\lambda_1)\frac{(K-2)}{(K-1)} \mu_2(i) P_2(j|i)}{\lambda_1 \mu_1(i) + (1-\lambda_1)\frac{(K-2)}{(K-1)} \mu_2(i)}. \tag{92}$$

Noting that the right-hand sides of (91) and (18) are identical, this completes the proof of the proposition. ■

B. Proof of Proposition 2

Let $C = (h, P_1, P_2)$ be the underlying configuration of the arms. We first show in the following lemma that under the non-stopping version of policy $\pi^\star(L, \delta)$, the maximum likelihood estimates $\hat{P}^n_{1,h}$ and $\hat{P}^n_{h,2}$ converge to their respective true values P_1 and P_2.

Lemma 4. *Let $C = (h, P_1, P_2)$ denote the underlying configuration of the arms. Then, under the non-stopping version of policy $\pi^\star(L, \delta)$, as $n \to \infty$, the following convergences hold a.s. for all $i, j \in \mathcal{S}$:*

$$\frac{N_a(n,i,j)}{N_a(n,i)} \longrightarrow \begin{cases} P_1(j|i), & a = h, \\ P_2(j|i), & a \neq h, \end{cases}$$
$$\frac{\sum\limits_{a \neq h} N_a(n,i,j)}{\sum\limits_{a \neq h} N_a(n,i)} \longrightarrow P_2(j|i). \tag{93}$$

□

Proof: Fix $i, j \in \mathcal{S}$ and $a \in \mathcal{A}$. Let $S_a(n)$ denote the quantity

$$S_a(n) = \sum_{t=0}^{n-1} \left(1_{\{A_{t+1}=a\}} - P(A_{t+1} = a | A^t, \bar{X}^t)\right), \tag{94}$$

where $P(A_{t+1} = a | A^t, \bar{X}^t)$ is given by

$$P(A_{t+1} = a | A^t, \bar{X}^t) = \frac{\delta}{K} + (1-\delta)\, \lambda^*(h^*(t), \hat{P}^t_{h^*(t),1}, \hat{P}^t_{h^*(t),2})(a). \tag{95}$$

Noting that (94) is a sum of a bounded martingale difference sequence, it follows from [18, Th. 1.2A] that $\frac{S_a(n)}{n} \to 0$ a.s. This implies that the following is true a.s. for sufficiently large values of n:

$$\frac{\delta}{2K} < \frac{N_a(n) - 1}{n} < 1 + \frac{\delta}{2K}. \tag{96}$$

Thus, we have $\liminf\limits_{n\to\infty} \frac{N_a(n)}{n} > \frac{\delta}{2K} > 0$ a.s.. By the ergodic theorem, it then follows that as $n \to \infty$, the following convergences hold a.s.:

$$\frac{N_a(n,i)}{N_a(n)} \longrightarrow \mu^a_h(i),$$
$$\frac{N_a(n,i,j)/N_a(n)}{N_a(n,i)/N_a(n)} \longrightarrow P^a_h(j|i); \tag{97}$$

here, $\mu^a_h(i)$ and $P^a_h(j|i)$ are as defined in (79) and (7) respectively. This establishes the convergence in the first line of (93) under the assumption that $C = (h, P_1, P_2)$ is the underlying configuration of the arms.

We then note that a.s.,

$$\frac{\sum\limits_{a \neq h} N_a(n,i,j)}{\sum\limits_{a \neq h} N_a(n,i)} = \frac{\sum\limits_{a \neq h} \frac{N_a(n,i,j)}{N^a_h(n,i)} \frac{N^a_h(n,i)}{N^a_h(n)} \frac{N^a_h(n)}{n}}{\sum\limits_{a \neq h} \frac{N_a(n,i)}{N^a_h(n)} \frac{N^a_h(n)}{n}}$$
$$\stackrel{n\to\infty}{\longrightarrow} P_2(j|i), \tag{98}$$

where the convergence in the last line above follows from (97) by noting that for $a \neq h$, when $C = (h, P_1, P_2)$ is the underlying configuration of the arms, $\mu^a_h(i) = \mu_2(i)$ and $P^a_h(j|i) = P_2(j|i)$. This establishes the convergence in the second line of (93), thus completing the proof of the lemma. ■

Proof of Proposition 2: We now use Lemma 4 to show that (38) holds for any $h' \neq h$. Towards this, we show that the quantity on the right-hand side of (29) is strictly positive.

For any choice of $\epsilon' > 0$, we have the following:

1) Since T_1 is a constant that does not grow with n, we have

$$\lim_{n\to\infty} \frac{T_1}{n} = 0, \tag{99}$$

and therefore it follows that there exists a positive integer $M_1 = M_1(\epsilon')$ such that $T_1/n \geq -\epsilon'$ for all $n \geq M_1$.

2) From (31), we have

$$\frac{T_2(n)}{n} = \frac{1}{n}\sum_{i\in\mathcal{S}} \log B((N_h(n,i,j)+1)_{j\in\mathcal{S}}). \tag{100}$$

Fix $i \in \mathcal{S}$. Then, we have

$$\log B((N_h(n,i,j)+1)_{j\in\mathcal{S}}) = \log E\left[\prod_{j\in\mathcal{S}} X_{ij}^{N_h(n,i,j)}\right], \tag{101}$$

where the random vector $(X_{ij})_{j\in\mathcal{S}}$ follows Dirichlet distribution with parameters $\alpha_j = 1$ for all $j \in \mathcal{S}$. We now write (101) as follows:

$$\frac{1}{N_h(n)}\log B((N_h(n,i,j)+1)_{j\in\mathcal{S}}) = \frac{1}{N_h(n)}\log E\left[\exp\left(N_h(n)\sum_{j\in\mathcal{S}}\frac{N_h(n,i,j)}{N_h(n)}\log X_{ij}\right)\right]. \tag{102}$$

When $C = (h, P_1, P_2)$ is the underlying configuration of the arms, from Lemma 4, we have that $N_h(n,i,j)/N_h(n)$ converges a.s. as $n \to \infty$ to $\mu_1(i)P_1(j|i)$. Thus, there exists a positive integer $M_{21} = M_{21}(\epsilon')$ such that for all $n \geq M_{21}$, we have

$$\frac{1}{N_h(n)}\log B((N_h(n,i,j)+1)_{j\in\mathcal{S}}) \geq \frac{1}{N_h(n)}\log E\left[\exp\left(N_h(n)\sum_{j\in\mathcal{S}}(\mu_1(i)P_1(j|i)+\epsilon')\log X_{ij}\right)\right]. \tag{103}$$

Noting that $N_h(n)$ converges a.s. to $+\infty$ as $n \to \infty$, by Varadhan's integral lemma [19, Theorem 4.3.1], there exists a positive integer $M_{22} = M_{22}(\epsilon')$ such that for all $n \geq M_2 = \max\{M_{21}, M_{22}\}$, we have

$$\begin{aligned}\frac{1}{N_h(n)}\log B((N_h(n,i,j)+1)_{j\in\mathcal{S}}) &\overset{(a)}{\geq} \sup_{\{z_j\geq 0,\, \sum_{j\in\mathcal{S}} z_j = 1\}} \sum_{j\in\mathcal{S}}(\mu_1(i)P_1(j|i)+\epsilon')\log z_j - \frac{\epsilon'}{|\mathcal{S}|} \\ &= \sum_{j\in\mathcal{S}}(\mu_1(i)P_1(j|i)+\epsilon')\log\frac{\mu_1(i)P_1(j|i)+\epsilon'}{\mu_1(i)+\epsilon'|\mathcal{S}|} - \frac{\epsilon'}{|\mathcal{S}|},\end{aligned} \tag{104}$$

where the supremum on the right-hand side of (a) above is computed over all vectors $(z_j)_{j\in\mathcal{S}}$ such that $z_j \geq 0$ for all $j \in \mathcal{S}$, and $\sum_{j\in\mathcal{S}} z_j = 1$. Plugging (104) into (100), we get

$$\frac{T_2(n)}{n} \geq \frac{N_h(n)}{n}\left\{\left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}}(\mu_1(i)P_1(j|i)+\epsilon')\log\frac{\mu_1(i)P_1(j|i)+\epsilon'}{\mu_1(i)+\epsilon'|\mathcal{S}|}\right] - \epsilon'\right\} \tag{105}$$

for all $n \geq M_2$.

3) From (32), we have

$$\frac{T_3(n)}{n} = \frac{1}{n}\sum_{i\in\mathcal{S}}\log B\left(\left(\sum_{a\neq h} N_a(n,i,j)+1\right)_{j\in\mathcal{S}}\right). \tag{106}$$

Using the same arguments as those used to simplify (100), we obtain the following: there exists a positive integer $M_3 = M_3(\epsilon')$ such that for all $n \geq M_3$, we have

$$\frac{T_3(n)}{n} \geq \frac{\sum_{a\neq h} N_a(n)}{n}\left\{\left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}}(\mu_2(i)P_2(j|i)+\epsilon')\log\frac{\mu_2(i)P_2(j|i)+\epsilon'}{\mu_2(i)+\epsilon'|\mathcal{S}|}\right] - \epsilon'\right\}. \tag{107}$$

4) From (33), we have

$$\frac{T_4(n)}{n} = -\frac{1}{n}\sum_{i,j\in\mathcal{S}} N_{h'}(n,i,j)\log\frac{N_{h'}(n,i,j)}{N_{h'}(n,i)}. \tag{108}$$

If $N_h(n,i) = 0$ for some state $i \in \mathcal{S}$ (in which case it follows that $N_h(n,i,j) = 0$ for all $j \in \mathcal{S}$), or if $N_h(n,i,j) = 0$[1] for some pair of states $i, j \in \mathcal{S}$, then the corresponding terms in the summation in (108) will be of the form $0\log\frac{0}{0}$ or $0\log 0$ respectively, which we treat as zero by convention. Thus, without loss of generality, we assume that $N_h(n,i,j) > 0$ for all $i, j \in \mathcal{S}$.

[1]This may be the case if, for instance, $P_2(j|i) = 0$ for some pair of states $i, j \in \mathcal{S}$.

Noting that $h' \neq h$, when the underlying configuration is $C = (h, P_1, P_2)$, from Lemma 4, we have the following almost sure convergences (as $n \to \infty$):

$$\frac{N_{h'}(n,i,j)}{n} \to \mu_2(i)P_2(j|i),$$
$$\frac{N_{h'}(n,i,j)}{N_{h'}(n,i)} \to P_2(j|i). \tag{109}$$

Using these in (108), we get that there exists a positive integer $M_4 = M_4(\epsilon')$ such that for all $n \geq M_4$, we have

$$\frac{T_4(n)}{n} \geq \sum_{i,j \in \mathcal{S}} (\mu_2(i)P_2(j|i) - \epsilon') \log \frac{1}{P_2(j|i) + \epsilon'}. \tag{110}$$

5) Lastly, we present a simplification of the term $T_5(n)/n$. From (34), we have

$$\frac{T_5(n)}{n} = -\frac{1}{n} \sum_{i,j \in \mathcal{S}} \sum_{a \neq h'} N_a(n,i,j) \log \frac{\sum\limits_{a \neq h'} N_a(n,i,j)}{\sum\limits_{a \neq h'} N_a(n,i)}. \tag{111}$$

For each n and each $i, j \in \mathcal{S}$, we define $P_n(j|i)$ as the following quantity:

$$P_n(j|i) = \frac{\sum\limits_{a \neq h'} N_a(n,i,j)}{\sum\limits_{a \neq h'} N_a(n,i)}. \tag{112}$$

Note that $P_n = (P_n(j|i))_{i,j \in \mathcal{S}}$ constitutes a valid transition matrix. From Lemma 4, under the underlying configuration $C = (h, P_1, P_2)$, we note the following almost convergences as $n \to \infty$:

$$\frac{\sum\limits_{a \neq h,h'} N_a(n,i,j)}{\sum\limits_{a \neq h,h'} N_a(n,i)} \overset{n\to\infty}{\longrightarrow} P_2(j|i),$$
$$\frac{\sum\limits_{a \neq h,h'} N_a(n,i)}{\sum\limits_{a \neq h,h'} N_a(n)} \overset{n\to\infty}{\longrightarrow} \mu_2(i). \tag{113}$$

The above convergences then imply that there exists a positive integer $M_5 = M_5(\epsilon')$ such that for all $n \geq M_5$, we have

$$\frac{T_5(n)}{n} \geq \frac{N_h(n)}{n} \sum_{i,j \in \mathcal{S}} (\mu_1(i)P_1(j|i) - \epsilon') \log \frac{1}{P_n(j|i)} + \frac{\sum\limits_{a \neq h,h'} N_a(n)}{n} \sum_{i,j \in \mathcal{S}} (\mu_2(i)P_2(j|i) - \epsilon') \log \frac{1}{P_n(j|i)}. \tag{114}$$

Combining the results in (99), (105), (107), (110) and (114), we get that for all $n \geq M(\epsilon') = \max\{M_1, \ldots, M_5\}$, we have

$$\frac{M_{hh'}(n)}{n} \geq f_n(\epsilon'), \tag{115}$$

where $f_n(\epsilon')$ denotes the sum of the terms of the right-hand sides of (99), (105), (107), (110) and (114).

We now define $f_n(0)$ as the following quantity:

$$f_n(0) \coloneqq \frac{N_h(n)}{n} D(P_1||P_n|\mu_1) + \frac{\sum\limits_{a \neq h,h'} N_a(n)}{n} D(P_2||P_n||\mu_2). \tag{116}$$

Then, by continuity, we have that for any choice of $\epsilon > 0$, there exists $\epsilon' > 0$ such that $f_n(\epsilon') > f_n(0) - \epsilon$ for all sufficiently large values of n. From (115), this implies that

$$\frac{M_{hh'}(n)}{n} > f_n(0) - \epsilon \tag{117}$$

for all sufficiently large values of n, from which it follows that

$$\liminf_{n\to\infty} \left[\frac{M_{hh'}(n)}{n} - f_n(0) \right] \geq -\epsilon. \tag{118}$$

Since the above equation is true for an arbitrary choice of ϵ, letting $\epsilon \downarrow 0$, we get

$$\liminf_{n\to\infty} \frac{M_{hh'}(n)}{n} - \limsup_{n\to\infty} f_n(0) \geq 0, \tag{119}$$

from which it follows that

$$\begin{aligned}\liminf_{n\to\infty}\frac{M_{hh'}(n)}{n}&\geq \limsup_{n\to\infty} f_n(0)\\&\geq \liminf_{n\to\infty} f_n(0)\\&\geq \liminf_{n\to\infty}\left\{\frac{N_h(n)}{n}D(P_1||P_n|\mu_1)+\frac{\sum\limits_{a\neq h,h'}N_a(n)}{n}D(P_2||P_n|\mu_2)\right\}\\&\geq \liminf_{n\to\infty}\left\{\frac{N_h(n)}{n}\,D(P_1||P_n|\mu_1)\right\}+\liminf_{n\to\infty}\left\{\frac{\sum\limits_{a\neq h,h'}N_a(n)}{n}\,D(P_2||P_n|\mu_2)\right\}\end{aligned} \tag{120}$$

We now claim that $\sup\limits_{n\geq 0} D(P_1||P_n|\mu_1)<\infty$ a.s.. Indeed, we note that

$$\begin{aligned}P_n(j|i)&=\frac{\sum\limits_{a\neq h'}N_a(n,i,j)}{\sum\limits_{a\neq h'}N_a(n,i)}\\&\geq\frac{\sum\limits_{a\neq h'}N_a(n,i,j)}{n}\\&\geq\left(\frac{N_h(n)}{n}\right)\left(\frac{N_h(n,i)}{N_h(n)}\right)\left(\frac{N_h(n,i,j)}{N_h(n,i)}\right)+\left(\frac{\sum\limits_{a\neq h,h'}N_a(n)}{n}\right)\left(\frac{\sum\limits_{a\neq h,h'}N_a(n,i)}{\sum\limits_{a\neq h,h'}N_a(n)}\right)\left(\frac{\sum\limits_{a\neq h,h'}N_a(n,i,j)}{\sum\limits_{a\neq h,h'}N_a(n,i)}\right)\\&\overset{(a)}{\geq}\left(\frac{\delta}{2K}\right)\left(\frac{\mu_1(i)\,P_1(j|i)}{2}\right)+(K-2)\left(\frac{\delta}{2K}\right)\left(\frac{\mu_2(i)\,P_2(j|i)}{2}\right)\\&\overset{(b)}{\geq}\left(\frac{\delta}{2K}\right)\left(\frac{\mu_1(i)P_1(j|i)+\mu_2(i)P_2(j|i)}{2}\right)\\&\geq\left(\frac{\delta}{2K}\right)\left(\min\left\{\min_{i\in\mathcal{S}}\mu_1(i),\,\min_{i\in\mathcal{S}}\mu_2(i)\right\}\right)\left(\frac{P_1(j|i)+P_2(j|i)}{2}\right)\ a.s.\end{aligned} \tag{121}$$

for all sufficiently large values of n, where (a) follows from (96) and Lemma 4, and (b) follows by using the fact that the number of arms $K\geq 3$. It then follows that

$$\begin{aligned}D(P_1||P_n|\mu_1)&=\sum_{i\in\mathcal{S}}\mu_1(i)\sum_{j\in\mathcal{S}}P_1(j|i)\log\frac{P_1(j|i)}{P_n(j|i)}\\&\leq\sum_{i,j\in\mathcal{S}}\mu_1(i)\,P_1(j|i)\log\frac{P_1(j|i)}{\frac{P_1(j|i)+P_2(j|i)}{2}}+\sum_{i,j\in\mathcal{S}}\mu_1(i)P_1(j|i)\log P_1(j|i)+\log\frac{1}{\left(\frac{\delta}{2K}\right)\left(\min\left\{\min\limits_{i\in\mathcal{S}}\mu_1(i),\,\min\limits_{i\in\mathcal{S}}\mu_2(i)\right\}\right)}\\&=D\left(P_1\middle|\middle|\frac{P_1+P_2}{2}\middle|\mu_1\right)+\sum_{i\in\mathcal{S}}\mu_1(i)(-H(P_1(\cdot|i))+\log\frac{1}{\left(\frac{\delta}{2K}\right)\left(\min\left\{\min\limits_{i\in\mathcal{S}}\mu_1(i),\,\min\limits_{i\in\mathcal{S}}\mu_2(i)\right\}\right)}\\&<\infty\ a.s..\end{aligned} \tag{122}$$

On similar lines, it can be shown that $D(P_2||P_n|\mu_1)$ is bounded uniformly a.s. for all $n\geq 0$. Using the uniform boundedness property just proved, we may express (120) as

$$\begin{aligned}\liminf_{n\to\infty}\frac{M_{hh'}(n)}{n}&\geq\left\{\liminf_{n\to\infty}\frac{N_h(n)}{n}\right\}\left\{\liminf_{n\to\infty}D(P_1||P_n|\mu_1)\right\}+\left\{\liminf_{n\to\infty}\frac{\sum\limits_{a\neq h,h'}N_a(n)}{n}\right\}\left\{\liminf_{n\to\infty}D(P_2||P_n|\mu_2)\right\}\\&\geq\left(\frac{\delta}{2K}\right)\left(\liminf_{n\to\infty}D(P_1||P_n|\mu_1)+(K-2)\,\liminf_{n\to\infty}D(P_2||P_n|\mu_2)\right)\ a.s.,\end{aligned} \tag{123}$$

where the last line follows from (96).

Finally, we show that the first limit infimum term in (123) is strictly positive, and note that an exactly parallel argument may be used to show that the second limit infimum term is also strictly positive. Suppose that $\liminf\limits_{n\to\infty}D(P_1||P_n|\mu_1)=0$ a.s.. By the property that KL divergence is zero if and only if the argument probability distributions are identical, it follows that there exists a subsequence $(n_k)_{k\geq 1}$ such that $P_{n_k}(j|i)\to P_1(j|i)$ as $k\to\infty$ a.s. for all $i,j\in\mathcal{S}$. We now fix attention to this subsequence, and note that by the property that the sequences $(N_h(n_k)/n_k)_{k\geq 1}$ and $(\sum\limits_{a\neq h,h'}N_a(n_k)/n_k)_{k\geq 1}$ are bounded, there exists a further subsequence $(n_{k_l})_{l\geq 1}$ of $(n_k)_{k\geq 1}$ such that the aforementioned bounded sequences admit limits, say α and β respectively. From Lemma 4, we then have the following convergence a.s. as $l \to \infty$:

$$P_{n_{k_l}}(j|i) \to \frac{\alpha\,\mu_1(i)\,P_1(j|i) + \beta\,\mu_2(i)\,P_2(j|i)}{\alpha\,\mu_1(i) + \beta\,\mu_2(i)}. \tag{124}$$

However, we note that the right-hand side of (124) is not equal to $P_1(j|i)$ whenever $P_2(j|i) > 0$, thus resulting in a contradiction. This completes the proof of the proposition. ■

C. Proof of Proposition 3

The policy $\pi^\star(L,\delta)$ commits error if one of the following events is true:

1) The policy never stops in finite time.
2) The policy stops in finite time and declares $h' \neq h$ as the true index of the odd arm.

The event in item 1 above has zero probability as a consequence of Proposition 2. Thus, the probability of error of policy $\pi = \pi^\star(L,\delta)$, which we denote by P_e^π, may be evaluated as follows: suppose $C = (h, P_1, P_2)$ is the underlying configuration of the arms. Then,

$$P_e^\pi = P^\pi(I(\pi) \neq h|C) = P^\pi\left(\exists\; n \text{ and } \; h' \neq h \text{ such that } I(\pi) = h' \text{ and } \tau(\pi) = n\middle|C\right). \tag{125}$$

We now let

$$\mathcal{R}_{h'}(n) \coloneqq \{\omega : \tau(\pi)(\omega) = n,\, I(\pi)(\omega) = h'\} \tag{126}$$

denote the set of all sample paths for which the policy stops at time n and declares h' as the true index of the odd arm. Clearly, the collection $\{\mathcal{R}_{h'}(n) : h' \neq h,\, n \geq 0\}$ is a collection of mutually disjoint sets. Therefore, we have

$$\begin{aligned}
P_e^\pi &= P^\pi\left(\bigcup_{h'\neq h}\bigcup_{n=0}^{\infty}\mathcal{R}_{h'}(n)\middle|C\right)\\
&= \sum_{h'\neq h}\sum_{n=0}^{\infty} P^\pi(\tau(\pi) = n, I(\pi) = h'|C)\\
&= \sum_{h'\neq h}\sum_{n=0}^{\infty} \int_{\mathcal{R}_{h'}(n)} dP^\pi(\omega|C)\\
&\overset{(a)}{=} \sum_{h'\neq h}\sum_{n=0}^{\infty} \int_{\mathcal{R}_{h'}(n)} f(A^n(\omega), \bar{X}^n(\omega)|H_h)\left[\prod_{t=0}^{n} P_h(A_t|A^{t-1}, \bar{X}^{t-1})\right] d(A^n(\omega), \bar{X}^n(\omega))\\
&\overset{(b)}{\leq} \sum_{h'\neq h}\sum_{n=0}^{\infty} \int_{\mathcal{R}_{h'}(n)} \hat{f}(A^n(\omega), \bar{X}^n(\omega)|H_h)\left[\prod_{t=0}^{n} P_h(A_t|A^{t-1}, \bar{X}^{t-1})\right] d(A^n(\omega), \bar{X}^n(\omega))\\
&\overset{(c)}{=} \sum_{h'\neq h}\sum_{n=0}^{\infty} \left\{\int_{\mathcal{R}_{h'}(n)} e^{-M_{h'h}(n)}\; f(A^n(\omega), \bar{X}^n(\omega)|H_{h'})\left[\prod_{t=0}^{n} P_{h'}(A_t|A^{t-1}, \bar{X}^{t-1})\right] d(A^n(\omega), \bar{X}^n(\omega))\right\}\\
&\leq \sum_{h'\neq h}\sum_{n=0}^{\infty} \left\{\int_{\mathcal{R}_{h'}(n)} \frac{1}{(K-1)L}\; dP^\pi(\omega|C')\right\}\\
&= \sum_{h'\neq h} \frac{1}{(K-1)L}\; P^\pi\left(\bigcup_{n=0}^{\infty}\mathcal{R}_{h'}(n)\middle|C'\right) \leq \frac{1}{L},
\end{aligned} \tag{127}$$

where in (a) above, $P_h(A_t|A^{t-1}, \bar{X}^{t-1})$ denotes the probability of selecting arm A_t at time t when the index of the odd arm is h, with the convention that at time $t = 0$, this term represents $P_h(A_0)$; (b) above follows by the definition of $\hat{f}$ in (28), and (c) follows by using the fact that the probability of selecting an arm at any time t, based on the history of past arm selections and observations, is independent of the odd arm index, and is thus the same when the arm indexed by either h or h' is the odd arm. Setting $L = 1/\epsilon$ gives $P_e^\pi \leq \epsilon$, thus proving that $\pi = \pi^\star(L,\delta) \in \Pi(\epsilon)$. This completes the proof of the proposition. □

D. Proof of Proposition 4

Before we present the proof of Proposition 4, we show that the odd arm chosen by the non-stopping version of policy $\pi^\star(L,\delta)$ is indeed the correct one. Further, we show that the arm selection frequencies under the same policy converge to the respective optimal values given in (21).

Proposition 6. *Let $C=(h,P_1,P_2)$ denote the underlying configuration of the arms. Fix $L\geq 1$ and $\delta\in(0,1)$, and consider the non-stopping version of policy $\pi^\star(L,\delta)$. For any $h'\neq h$ and $i,j\in\mathcal{S}$, let*

$$P_n(j|i) \coloneqq \frac{\sum\limits_{a\neq h'} N_a(n,i,j)}{\sum\limits_{a\neq h'} N_a(n,i)}. \tag{128}$$

Then, the following convergences hold a.s. as $n\to\infty$.

$$h^*(n)\to h, \tag{129}$$

$$\lambda_{opt}(h^*(n),\hat{P}^n_{h^*(n),1},\hat{P}^n_{h^*(n),2})\to\lambda_{opt}(h,P_1,P_2), \tag{130}$$

$$\frac{N_a(n)}{n}\to\lambda^*_\delta(h,P_1,P_2)(a) \text{ for all } a\in\mathcal{A}, \tag{131}$$

$$P_n(j|i)\to P_\delta(j|i) \text{ for all } i,j\in\mathcal{S}, \tag{132}$$

where for each $a\in\mathcal{A}$ and each $i,j\in\mathcal{S}$, the quantity $\lambda^*_\delta(h,P_1,P_2)(a)$ and the term $P_\delta(j|i)$ in (132) are as defined in the statement of Proposition 4. □

Proof: Fix an arbitrary $h'\neq h$. Then, as a consequence of Proposition 2, we have the following set of inequalities holding a.s.:

$$\begin{aligned}\limsup_{n\to\infty} M_{h'}(n) &= \limsup_{n\to\infty}\min_{a\neq h'} M_{h'a}(n)\\ &\leq \limsup_{n\to\infty} M_{h'h}(n)\\ &= \limsup_{n\to\infty} -M_{hh'}(n)\\ &= -\liminf_{n\to\infty} M_{hh'}(n)\\ &\leq -\liminf_{n\to\infty} M_h(n)\\ &< 0.\end{aligned} \tag{133}$$

From the above set of inequalities, it follows that under policy $\pi^\star(L,\delta)$,

$$h^*(n)=\arg\max_{h\in\mathcal{A}} M_h(n)=h \text{ a.s.} \tag{134}$$

for all sufficiently large values of n. This establishes (129), which in turn implies that

$$\lambda_{opt}(h^*(n),\hat{P}^n_{h^*(n),1},\hat{P}^n_{h^*(n),2})\to\lambda_{opt}(h,P_1,P_2), \tag{135}$$

because of the convergence of the maximum likelihood estimates shown in (93), and the fact that $\lambda^*(h,P,Q)$ is jointly continuous in the pair (P,Q), a fact that follows from Berge's Maximum Theorem [20]. This establishes (130).

We now proceed to show (131). Towards this, we observe that from (37) and the convergence in (130), we have

$$\begin{aligned}P(A_{n+1}=a|A^n,\bar{X}^n) &= \frac{\delta}{K}+(1-\delta)\,\lambda_{opt}(h^*(n),\hat{P}^n_{h^*(n),1},\hat{P}^n_{h^*(n),2})(a)\\ &\to \frac{\delta}{K}+(1-\delta)\lambda_{opt}(h,P_1,P_2)(a).\end{aligned} \tag{136}$$

We revisit the quantity $S_a(n)$ defined in (94), and use the fact that $\frac{S_a(n)}{n}\to 0$ a.s. as $n\to\infty$ to obtain

$$\begin{aligned}\frac{N_a(n)}{n} &\to \frac{1}{n}\sum_{t=0}^{n-1} P(A_{t+1}=a|A^t,\bar{X}^t)\\ &\to \frac{\delta}{K}+(1-\delta)\lambda_{opt}(h,P_1,P_2)(a).\end{aligned} \tag{137}$$

This establishes (131).

Defining

$$\alpha_n\coloneqq\frac{N_h(n)}{n},\quad \beta_n\coloneqq\frac{\sum\limits_{a\neq h,h'} N_a(n)}{n}, \tag{138}$$

we note that the convergence in (131) implies in particular that

$$\begin{aligned}\alpha_n &\to \lambda_\delta^*(h, P_1, P_2)(h) = \frac{\delta}{K} + (1-\delta)\lambda^* = \lambda_\delta^*,\\ \beta_n &\to (K-2)\left(\frac{\delta}{K} + (1-\delta)\frac{1-\lambda^*}{K-1}\right)\\ &= \frac{(K-2)}{(K-1)}\left\{1 - \left(\frac{\delta}{K} + (1-\delta)\lambda^*\right)\right\}\\ &= \frac{(K-2)}{(K-1)}(1-\lambda_\delta^*).\end{aligned} \tag{139}$$

Taking limits as $n \to \infty$ on both sides of (128), and using the above limits for α_n and β_n, we get the convergence in (132), hence completing the proof of the proposition. ■

Proof of Proposition 4: We recall from (120) and (116) that

$$\begin{aligned}\liminf_{n\to\infty} \frac{M_{hh'}(n)}{n} &\geq \liminf_{n\to\infty} \alpha_n D(P_1||P_n|\mu_1) + \liminf_{n\to\infty} \beta_n D(P_2||P_n|\mu_2)\\ &= \lambda_\delta^* D(P_1||P_\delta|\mu_1) + \frac{(K-2)}{(K-1)}(1-\lambda_\delta^*) D(P_2||P_\delta||\mu_2),\end{aligned} \tag{140}$$

where the terms α_n and β_n are as given in (138). Using Varadhan's integral lemma [19, Theorem 4.3.1] to write

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log B((N_h(n,i,j)+1)_{j\in\mathcal{S}}) &\leq \limsup_{n\to\infty} \frac{N_h(n)}{n}\mu_1(i) \sup_{\{z_j\geq 0,\, \sum_{j\in\mathcal{S}} z_j = 1\}} \sum_{j\in\mathcal{S}} P_1(j|i)\log z_j\\ &= \lim_{n\to\infty} \frac{N_h(n)}{n}\mu_1(i)(-H(P_1(\cdot|i))),\end{aligned} \tag{141}$$

and following similar steps leading to (105), we obtain

$$\begin{aligned}\limsup_{n\to\infty} \frac{M_{hh'}(n)}{n} &\leq \lim_{n\to\infty} \alpha_n D(P_1||P_n|\mu_1) + \lim_{n\to\infty} \beta_n D(P_2||P_n|\mu_2)\\ &= \lambda_\delta^* D(P_1||P_\delta|\mu_1) + \frac{(K-2)}{(K-1)}(1-\lambda_\delta^*) D(P_2||P_\delta||\mu_2).\end{aligned} \tag{142}$$

Combining (140) and (142), we get the desired result. ■

E. Proof of Proposition 5

This section is organised as follows. We first show in Lemma 5 that the stopping time of policy $\pi^\star(L,\delta)$ goes to infinity as the error probability vanishes (or as $L \to \infty$). We then exploit this to show that under policy $\pi^\star(L,\delta)$, the modified GLR statistic has the correct drift (see Lemma 6). That is, we build on the result of Proposition 2 and obtain the explicit limit for the modified GLR statistic for the regime of vanishing error probability. We then use the result of Lemma 6 to show in Lemma 7 that the stopping time of policy $\pi^*(L,\delta)$ satisfies an asymptotic almost sure upper bound that matches with the right-hand side of (45). Finally, we establish that for any fixed $\delta \in (0,1)$, the family $\{\tau(\pi^\star(L,\delta))/\log L : L \geq 1\}$ is uniformly integrable, and as an intermediate step towards this, we establish in Lemma 8 an exponential upper bound for a certain probability term. Combining the almost sure limit of Lemma 7 along with the uniform integrability result then yields the desired upper bound in (45).

Lemma 5. *Let $C = (h, P_1, P_2)$ denote the underlying configuration of the arms. Fix $\delta \in (0,1)$. Then, under policy $\pi^\star(L,\delta)$, we have*

$$\liminf_{L\to\infty} \tau(\pi^\star(L,\delta)) = \infty \text{ a.s.} \tag{143}$$

□

Proof: Since policy $\pi = \pi^\star(L,\delta)$ selects each of the K arms in the first K slots, in order to prove the lemma, we note that it suffices to prove the following statement:

$$\text{for each } m \geq K, \quad \lim_{L\to\infty} P^\pi(\tau(\pi) \leq m|C) = 0. \tag{144}$$

Fix $m \geq K$, and note that

$$\begin{aligned}\limsup_{L\to\infty} P^{\pi}(\tau(\pi) \leq m|C) &= \limsup_{L\to\infty} P^{\pi}\left(\exists\, K \leq n \leq m \text{ and } \tilde{h} \in \mathcal{A} \text{ such that } M_{\tilde{h}}(n) > \log((K-1)L) \middle| C\right)\\ &\leq \limsup_{L\to\infty} \sum_{\tilde{h}\in\mathcal{A}} \sum_{n=K}^{m} P^{\pi}(M_{\tilde{h}}(n) > \log((K-1)L)|C)\\ &\leq \limsup_{L\to\infty} \frac{1}{\log((K-1)L)} \sum_{\tilde{h}\in\mathcal{A}} \sum_{n=K}^{m} E^{\pi}[M_{\tilde{h}}(n)|C], \end{aligned} \tag{145}$$

where the first inequality above follows from the union bound, and the second inequality follows from Markov's inequality.

We now show that for each $m \in \{K, \ldots, n\}$, the expectation term inside the summation in (145) is finite. Towards this, we have

$$\begin{aligned} M_{\tilde{h}}(n) &= \log\left(\frac{f(A^n, \bar{X}^n|H_{\tilde{h}})}{\max\limits_{h'\neq\tilde{h}} \hat{f}(A^n, \bar{X}^n|H_{h'})}\right)\\ &\leq \log\left(\frac{\hat{f}(A^n, \bar{X}^n|H_{\tilde{h}})}{\hat{f}(A^n, \bar{X}^n|H_{h'})}\right) \text{ for all } h' \neq \tilde{h}. \end{aligned} \tag{146}$$

Fix an arbitrary $h' \neq \tilde{h}$. We recognise that the logarithmic term in (146) is the classical GLR test statistic of hypothesis $H_{\tilde{h}}$ with respect to hypothesis $H_{h'}$, given by

$$\log\left(\frac{\hat{f}(A^n, \bar{X}^n|H_{\tilde{h}})}{\hat{f}(A^n, \bar{X}^n|H_{h'})}\right) = S_1(n) + S_2(n) + S_3(n) + S_4(n), \tag{147}$$

where the terms $S_1(n), \ldots, S_4(n)$ appearing in (147) are as below.

1) The term $S_1(n)$ is given by

$$S_1(n) = \sum_{i,j\in\mathcal{S}} N_{\tilde{h}}(n,i,j) \log \frac{N_{\tilde{h}}(n,i,j)}{N_{\tilde{h}}(n,i)}. \tag{148}$$

2) The term $S_2(n)$ is given by

$$S_2(n) = \sum_{i,j\in\mathcal{S}} \sum_{a\neq\tilde{h}} N_a(n,i,j) \log \frac{\sum\limits_{a\neq\tilde{h}} N_a(n,i,j)}{\sum\limits_{a\neq\tilde{h}} N_a(n,i)}. \tag{149}$$

3) The term $S_3(n)$ is given by

$$S_3(n) = -\sum_{i,j\in\mathcal{S}} N_{h'}(n,i,j) \log \frac{N_{h'}(n,i,j)}{N_{h'}(n,i)}. \tag{150}$$

4) The term $S_4(n)$ is given by

$$S_4(n) = -\sum_{i,j\in\mathcal{S}} \sum_{a\neq h'} N_a(n,i,j) \log \frac{\sum\limits_{a\neq h'} N_a(n,i,j)}{\sum\limits_{a\neq h'} N_a(n,i)}. \tag{151}$$

We now obtain an a.s. upper bound for (147). We recognise that $S_1(n)$ and $S_2(n)$ are non-positive, and thus upper bound each of these terms by zero. Let

$$A(i) = (N_{h'}(n,i,j)/N_{h'}(n,i))_{j\in\mathcal{S}}$$

denote the probability vector corresponding to state i. Then, denoting the Shannon entropy of $A(i)$ by $H(A(i))$, we may express $S_3(n)$ as

$$\begin{aligned} S_3(n) &= (N_{h'}(n)-1) \sum_{i\in\mathcal{S}} \left[\frac{N_{h'}(n,i)}{N_{h'}(n)-1}\right] H(A(i))\\ &\leq (N_{h'}(n)-1)\ H\left(\sum_{i\in\mathcal{S}} \left[\frac{N_{h'}(n,i)}{N_{h'}(n)-1}\right] A(i)\right)\\ &\leq N_{h'}(n) \log|\mathcal{S}|, \end{aligned} \tag{152}$$

where the first inequality above follows from the concavity of the entropy function $H(\cdot)$, and the second inequality follows by noting that the Shannon entropy of a probability distribution on an alphabet of size R is upper bounded by $\log R$. On similar lines, we get

$$S_4(n) \leq \left(\sum_{a\neq h'} N_a(n)\right) \log|\mathcal{S}|. \tag{153}$$

Using in (147) the results of (152) and (153), along with the zero upper bound for the non-positive terms in (148) and (149) and the relation (5c), we get

$$M_{\tilde{h}}(n) \leq (n+1)\log|\mathcal{S}| \text{ a.s.}, \tag{154}$$

from which it follows that

$$\begin{aligned}\limsup_{L\to\infty} P^{\pi}(\tau(\pi)\leq m|C) &\leq \limsup_{L\to\infty} \frac{1}{\log((K-1)L)} \sum_{\tilde{h}\in\mathcal{A}} \sum_{n=K}^{m} (n+1)\log|\mathcal{S}| \\ &= 0. \end{aligned} \tag{155}$$

This completes the proof of the lemma. ■

Lemma 6. *Let $C=(h,P_1,P_2)$ denote the underlying configuration of the arms. Fix $\delta\in(0,1)$. Then, under policy $\pi=\pi^\star(L,\delta)$, for any $h'\neq h$, we have*

$$\lim_{L\to\infty} \frac{M_{hh'}(\tau(\pi))}{\tau(\pi)} = D^*_{\delta}(h,P_1,P_2) \quad a.s. \tag{156}$$

□

Proof: The proof follows as a consequence of Proposition 4 and Lemma 5. ■

Lemma 7. *Let $C=(h,P_1,P_2)$ denote the underlying configuration of the arms. Fix $\delta\in(0,1)$. Then, under policy $\pi=\pi^*(L,\delta)$, we have*

$$\limsup_{L\to\infty} \frac{\tau(\pi)}{\log L} \leq \frac{1}{D^*_{\delta}(h,P_1,P_2)} \quad a.s. \tag{157}$$

Proof: We first show that for any $h'\neq h$ and $n\geq 1$, the increment $M_{hh'}(n)-M_{hh'}(n-1)$ is bounded. Fix an arbitrary $h'\neq h$, and consider the following cases.

1) Case 1: Suppose that arm h is selected at time n. Then, noting that in the expression for $M_{hh'}(n)$, the only terms that depend on the arm index h are those in (31) and (34), we have

$$M_{hh'}(n)-M_{hh'}(n-1) = \Big[T_2(n)-T_2(n-1)\Big] + \Big[T_5(n)-T_5(n-1)\Big]. \tag{158}$$

Suppose that at time n, the Markov process of arm h undergoes a transition from state i to state j, where $i,j\in\mathcal{S}$ are such that $\max\{P_1(j|i),P_2(j|i)\}>0$[2]. Then, noting that

$$\begin{aligned} N_a(n,i',j') &= N_a(n-1,i',j') \quad \text{for all } a\in\mathcal{A},\ i'\neq i,\ j'\neq j, \\ N_h(n,i,j) &= N_h(n-1,i,j)+1, \\ N_a(n,i') &= N_a(n-1,i') \quad \text{for all } a\in\mathcal{A},\ i'\neq i, \\ N_h(n,i) &= N_h(n-1,i)+1, \end{aligned} \tag{159}$$

it can be shown after some simplification that

$$\begin{aligned} T_2(n)-T_2(n-1) &= \log\frac{B(N_h(n-1,i,j)+2,(N_h(n-1,i,j')+1)_{j'\neq j})}{B(N_h(n-1,i,j')+1)_{j'\in\mathcal{S}}} \\ &\overset{(a)}{=} \frac{N_h(n-1,i,j)}{\sum_{j'\in\mathcal{S}} N_h(n-1,i,j')} \\ &\leq 1 \quad a.s., \end{aligned} \tag{160}$$

[2]Otherwise, a jump from i to j is not observed on arm h.

where (a) above follows by using the relation

$$B(\alpha_1,\ldots,\alpha_{|\mathcal{S}|}) = \left(\prod_{k=1}^{|\mathcal{S}|}\Gamma(\alpha_k)\right)\Big/\Gamma\left(\sum_{k=1}^{|\mathcal{S}|}\alpha_k\right). \tag{161}$$

Also, we have

$$\begin{aligned} T_5(n) - T_5(n-1) = &\left(\sum_{a\neq h'} N_a(n-1,i,j)\right)\log\frac{\sum\limits_{a\neq h'} N_a(n-1,i,j)}{\sum\limits_{a\neq h'} N_a(n-1,i)} \\ &- \left(1+\sum_{a\neq h'} N_a(n-1,i,j)\right)\log\frac{1+\sum\limits_{a\neq h'} N_a(n-1,i,j)}{1+\sum\limits_{a\neq h'} N_a(n-1,i)} \\ &\leq \log\frac{\sum\limits_{a\neq h'} N_a(n-1,i)}{\sum\limits_{a\neq h'} N_a(n,i,j)} \\ &\to \log\frac{1}{P_\delta(j|i)} \quad a.s., \end{aligned} \tag{162}$$

where the convergence in the last line follows from (132). Thus, it follows that the increment $M_{hh'}(n) - M_{hh'}(n-1)$ is bounded for all $n \geq 1$.

2) Case 2: Suppose that arm h' is sampled at time n. Noting that the only terms that depend on the arm index h' are those in (32) and (33), the analysis for this case proceeds on the exactly same lines as that of Case 1 presented above, and is omitted.
3) Case 3: Suppose that arm a' is sampled at time n, where $a' \in \mathcal{A}\setminus\{h,h'\}$. Noting that the only terms that depend on the arm index a' are those in (32) and (34), the analysis for this case proceeds on the exactly same lines as that of Case 1 presented above, and is omitted.

This establishes that the increments of the modified GLR process are bounded at all times.

Fix an arbitrary $h' \neq h$. By the definition of stopping time $\tau(\pi)$, we have that $M_{hh'}(\tau(\pi)-1) < \log((K-1)L)$. Using this, we have

$$\begin{aligned} \limsup_{L\to\infty}\frac{M_{hh'}(\tau(\pi))}{\log L} &\overset{(a)}{=} \limsup_{L\to\infty}\frac{M_{hh'}(\tau(\pi)-1)}{\log L} \\ &\leq \limsup_{L\to\infty}\frac{\log((K-1)L)}{\log L} \\ &= 1 \quad a.s., \end{aligned} \tag{163}$$

where (a) above is due to boundedness of the increments of the modified GLR process established above. Then, using Lemma 6 along with the relation (163) yields

$$\begin{aligned} \limsup_{L\to\infty}\frac{\tau(\pi)}{\log L} &= \limsup_{L\to\infty}\left\{\left(\frac{\tau(\pi)}{M_{hh'}(\tau(\pi))}\right)\left(\frac{M_{hh'}(\tau(\pi))}{\log L}\right)\right\} \\ &= \left(\lim_{L\to\infty}\frac{\tau(\pi)}{M_{hh'}(\tau(\pi))}\right)\left(\limsup_{L\to\infty}\frac{M_{hh'}(\tau(\pi))}{\log L}\right) \\ &\leq \frac{1}{D^*_\delta(h,P_1,P_2)} \quad a.s., \end{aligned} \tag{164}$$

thus completing the proof of the lemma. ■

Proof of Proposition 5: For any fixed $\delta \in (0,1)$, we now establish that under policy $\pi = \pi^\star(L,\delta)$, the family $\{\tau(\pi)/\log L : L \geq 1\}$ is uniformly integrable. In order to do so, we note that it suffices to show that

$$\limsup_{L\to\infty} E^\pi\left[\exp\left(\frac{\tau(\pi)}{\log L}\right)\middle| C\right] < \infty. \tag{165}$$

Towards this, let $l(L,\delta)$ denote the quantity

$$l(L,\delta) \coloneqq \frac{3\log((K-1)L)}{\frac{\delta}{2K}\Big(D(P_1||P_\delta|\mu_1) + D(P_2||P_\delta|\mu_2)\Big)}. \tag{166}$$

Let $C = (h, P_1, P_2)$ be the underlying configuration of the arms. Further, let $\pi_h^\star = \pi_h^\star(L,\delta)$ denote the version of policy $\pi^\star(L,\delta)$ that stops only upon declaring h as the index of the odd arm. Let

$$u(L) \coloneqq \exp\left(\frac{1 + l(L,\delta)}{\log L}\right) \tag{167}$$

Clearly, we have $\tau(\pi_h^\star) \geq \tau(\pi)$ a.s.. Then,

$$\begin{aligned}
\limsup_{L\to\infty} E^\pi\left[\exp\left(\frac{\tau(\pi)}{\log L}\right)\middle| C\right] &= \limsup_{L\to\infty} \int_0^\infty P^\pi\left(\frac{\tau(\pi)}{\log L} > \log x \middle| C\right) dx \\
&\leq \limsup_{L\to\infty} \int_0^\infty P^\pi\left(\tau(\pi_h^\star) \geq \lceil (\log x)(\log L)\rceil \middle| C\right) dx \\
&\overset{(a)}{\leq} \limsup_{L\to\infty} \left\{u(L) + \int_{u(L)}^\infty P^\pi\left(\tau(\pi_h^\star) \geq \lceil (\log x)(\log L)\rceil \middle| C\right) dx\right\} \\
&\leq \exp\left(\frac{3}{\frac{\delta}{2K}(D(P_1||P_\delta|\mu_1) + D(P_2||P_\delta|\mu_2))}\right) \\
&\qquad + \limsup_{L\to\infty} \sum_{n\geq l(L,\delta)} \exp\left(\frac{n+1}{\log L}\right) P^\pi(M_h(n) < \log((K-1)L)|C),
\end{aligned} \tag{168}$$

where (a) above follows by upper bounding the probability term by 1 for all $x \leq u(L)$.

We now show that for all $n \geq l(L,\delta)$, the probability term in (168) decays exponentially in n. This is a strengthening of the result in Proposition 2 which only establishes that when $C = (h, P_1, P_2)$ is the underlying configuration of the arms, $M_h(n) \to \infty$ as $n \to \infty$.

Lemma 8. *Let $C = (h, P_1, P_2)$ denote the underlying configuration of the arms. Fix $L \geq 1$, $\delta \in (0,1)$, and consider the policy $\pi = \pi^\star(L,\delta)$. There exist constants $\theta > 0$ and $0 < B < \infty$ independent of L such that for all sufficiently large values of n, we have*

$$P^\pi(M_h(n) < \log((K-1)L)|C) \leq Be^{-\theta n}. \tag{169}$$

□

Proof: Since

$$\begin{aligned}
P^\pi(M_h(n) < \log((K-1)L)|C) &= P^\pi\left(\min_{h'\neq h} M_{hh'}(n) < \log((K-1)L)\middle| C\right) \\
&\leq \sum_{h'\neq h} P^\pi\left(M_{hh'}(n) < \log((K-1)L)\middle| C\right),
\end{aligned} \tag{170}$$

in order to prove the lemma, it suffices to show that each term inside the summation in (170) is exponentially bounded. Going further, we drop the superscript π and the conditioning on configuration C in $P^\pi(\cdot|C)$ for ease of notation. For all $i, j \in \mathcal{S}$, let

$$\tilde{P}_n(j|i) \coloneqq \frac{\alpha_n \mu_1(i) P_1(j|i) + \beta_n \mu_2(i) P_2(j|i)}{\alpha_n \mu_1(i) + \beta_n \mu_2(i)}, \tag{171}$$

where α_n and β_n are as in (138). Fix $h' \neq h$ and $\epsilon > 0$ arbitrarily. Then, using (29) and triangle inequality, we have

$$P(M_{hh'}(n) < \log((K-1)L)) \leq U_1 + U_2 + U_3 + U_4 + U_5 + U_6 + U_7, \tag{172}$$

where the terms $U_1, \ldots, U_7$ in (172) are as below.

1) The term U_1 is given by

$$U_1 = P\left(\frac{T_1(n)}{n} < -\epsilon\right), \tag{173}$$

where T_1 is given by (30).

2) The term U_2 is given by

$$U_2 = P\left(\frac{T_2(n)}{n} - \frac{N_h(n)}{n}\sum_{i\in\mathcal{S}} \mu_1(i)(-H(P_1(\cdot|i))) < -\epsilon\right), \tag{174}$$

where $T_2(n)$ is given by (31).

3) The term U_3 is given by

$$U_3 = P\left(\frac{T_3(n)}{n} - \frac{\sum\limits_{a\neq h} N_a(n)}{n}\sum_{i\in\mathcal{S}}\mu_2(i)(-H(P_2(\cdot|i))) < -\epsilon\right), \tag{175}$$

where $T_3(n)$ is given by (32).

4) The term U_4 is given by

$$U_4 = P\left(\frac{T_4(n)}{n} - \frac{N_{h'}(n)}{n}\sum_{i\in\mathcal{S}}\mu_2(i)H(P_2(\cdot|i)) < -\epsilon\right), \tag{176}$$

where $T_4(n)$ is given by (33).

5) The term U_5 is given by

$$U_5 = P\left(\frac{T_5(n)}{n} - \sum_{i\in\mathcal{S}}(\alpha_n\mu_1(i) + \beta_n\mu_2(i))H(\tilde{P}_n(\cdot|i)) < -\epsilon\right), \tag{177}$$

where $T_5(n)$ is given by (34).

6) The term U_6 is given by

$$U_6 = P\bigg(\alpha_n\bigg[D(P_1||\tilde{P}_n|\mu_1) - D(P_1||P_\delta|\mu_1)\bigg] + \beta_n\bigg[D(P_2||\tilde{P}_n|\mu_2) - D(P_2||P_\delta|\mu_2)\bigg] < -\epsilon\bigg), \tag{178}$$

where P_δ is the transition matrix described in the statement of Proposition 4.

7) The term U_7 is given by

$$U_7 = P\bigg(\alpha_n D(P_1||P_\delta|\mu_1) + \beta_n D(P_2||P_\delta|\mu_2) - 6\epsilon < \frac{\log((K-1)L)}{n}\bigg). \tag{179}$$

In (174), the term $H(P_1(\cdot|i))$ refers to the Shannon entropy of the probability distribution $(P_1(j|i))_{j\in\mathcal{S}}$ on set $\mathcal{S}$; the terms $H(P_2(\cdot|i))$ and $H(\tilde{P}_n(\cdot|i))$ are defined similarly.

We now obtain a bound for the terms in (173)-(179).

1) We begin by showing an exponential upper bound for (179). We choose $0 < \epsilon' < \frac{2}{3}$, and then select $\epsilon > 0$ such that the following holds:

$$\frac{\delta}{2K}(1-\epsilon')\bigg(D(P_1||P_\delta|\mu_1) + D(P_2||P_\delta|\mu_2)\bigg) - 6\epsilon > \frac{1}{3}\cdot\frac{\delta}{2K}\bigg(D(P_1||P_\delta|\mu_1) + D(P_2||P_\delta|\mu_2)\bigg). \tag{180}$$

Then, for all $n \geq l(L,\delta)$, we have

$$P\bigg(\alpha_n D(P_1||P_\delta|\mu_1) + \beta_n D(P_2||P_\delta|\mu_2) - 6\epsilon < \frac{\log((K-1)L)}{n},\ \frac{N_a(n)}{n} > \frac{\delta}{2K}(1-\epsilon') \text{ for all } a\in\mathcal{A}\bigg) = 0. \tag{181}$$

Writing the probability term in (179) as a sum of the probability term in (181) and a second probability term given by

$$P\bigg(\alpha_n D(P_1||P_\delta|\mu_1) + \beta_n D(P_2||P_\delta|\mu_2) - 6\epsilon < \frac{\log((K-1)L)}{n},\ \frac{N_a(n)}{n} \leq \frac{\delta}{2K}(1-\epsilon') \text{ for some } a\in\mathcal{A}\bigg), \tag{182}$$

and upper bounding (182) by $P(N_a(n)/n \leq (\delta/2K)(1-\epsilon')$ for some $a\in\mathcal{A})$, an application of the union bound yields

$$P\bigg(\alpha_n D(P_1||P_\delta|\mu_1) + \beta_n D(P_2||P_\delta|\mu_2) - 6\epsilon < \frac{\log((K-1)L)}{n}\bigg) \leq \sum_{a=1}^{K} P\left(\frac{N_a(n)}{n} \leq \frac{\delta}{2K}(1-\epsilon')\right). \tag{183}$$

Noting that for each $a\in\mathcal{A}$, the sequence $\left(N_a(n) - n\frac{\delta}{2K}\right)_{n\geq 0}$ is a submartingale, with the absolute value of the difference between any two successive terms of the submartingale sequence being of value at most 1, we use the Azuma-Hoeffding inequality to obtain

$$\begin{aligned}
P\left(\frac{N_a(n)}{n} \leq \frac{\delta}{2K}(1-\epsilon')\right) &= P\left(N_a(n) - n\frac{\delta}{2K} \leq -n\epsilon'\frac{\delta}{2K}\right)\\
&= P\bigg(\bigg[N_a(n) - n\frac{\delta}{2K}\bigg] - N_a(0) \leq -n\epsilon'\frac{\delta}{2K} - N_a(0)\bigg)\\
&\leq P\bigg(\bigg[N_a(n) - n\frac{\delta}{2K}\bigg] - N_a(0) \leq -n\epsilon'\frac{\delta}{2K}\bigg)\\
&\leq \exp\left(-\frac{n(\epsilon')^2\delta^2}{8K^2}\right).
\end{aligned} \tag{184}$$

Plugging (184) back in (183), we arrive at

$$P\bigg(\alpha_n D(P_1||P_\delta|\mu_1) + \beta_n D(P_2||P_\delta|\mu_2) - 6\epsilon < \frac{\log((K-1)L)}{n}\bigg) \leq K \exp\left(-\frac{n(\epsilon')^2\delta^2}{8K^2}\right). \tag{185}$$

2) We now turn attention to (176), which we upper bound as follows:

$$\begin{aligned}
&P\left(\frac{T_4(n)}{n} - \frac{N_{h'}(n)}{n}\sum_{i\in\mathcal{S}}\mu_2(i)H(P_2(\cdot|i)) < -\epsilon\right)\\
&= P\bigg(\frac{N_{h'}(n)}{n}\bigg\{\sum_{i\in\mathcal{S}}\frac{N_{h'}(n,i)}{N_{h'}(n)}H\left(\frac{N_{h'}(n,i,\cdot)}{N_{h'}(n,i)}\right) - \mu_2(i)H(P_2(\cdot|i))\bigg\} < -\epsilon\bigg)\\
&\leq P\bigg(\frac{N_{h'}(n)}{n}\bigg\{\sum_{i\in\mathcal{S}}\frac{N_{h'}(n,i)}{N_{h'}(n)}H\left(\frac{N_{h'}(n,i,\cdot)}{N_{h'}(n,i)}\right) - \sum_{i\in\mathcal{S}}\mu_2(i)H(P_2(\cdot|i))\bigg\} < -\epsilon, \frac{N_a(n)}{n} > \frac{\delta}{2K}(1-\epsilon') \text{ for all } a\in\mathcal{A}\bigg)\\
&+\sum_{a=1}^{K}P\left(\frac{N_a(n)}{n} \leq \frac{\delta}{2K}(1-\epsilon')\right).
\end{aligned} \tag{186}$$

From the analysis using the Azuma-Hoeffding inequality for bounded difference submartingales presented earlier, we know that each term inside the summation in (186) is exponentially bounded. The first term in (186) may be written as

$$\begin{aligned}
&P\bigg(\frac{N_{h'}(n)}{n}\bigg\{\sum_{i\in\mathcal{S}}\frac{N_{h'}(n,i)}{N_{h'}(n)}H\left(\frac{N_{h'}(n,i,\cdot)}{N_{h'}(n,i)}\right) - \sum_{i\in\mathcal{S}}\mu_2(i)H(P_2(\cdot|i))\bigg\} < -\epsilon,\ \frac{N_a(n)}{n} > \frac{\delta}{2K}(1-\epsilon') \text{ for all } a\in\mathcal{A}\bigg)\\
&\leq P\bigg(\bigg\{\sum_{i\in\mathcal{S}}\frac{N_{h'}(n,i)}{N_{h'}(n)}H\left(\frac{N_{h'}(n,i,\cdot)}{N_{h'}(n,i)}\right) - \sum_{i\in\mathcal{S}}\mu_2(i)H(P_2(\cdot|i))\bigg\} < -\epsilon,\ \frac{N_a(n)}{n} > \frac{\delta}{2K}(1-\epsilon') \text{ for all } a\in\mathcal{A}\bigg).
\end{aligned} \tag{187}$$

From Lemma 4, we have the following almost sure convergences as $n\to\infty$:

$$\begin{aligned}
\frac{N_{h'}(n,i,j)}{N_{h'}(n,i)} &\to P_2(j|i), \text{ for all } i,j\in\mathcal{S},\\
\frac{N_{h'}(n,i)}{N_{h'}(n)} &\to \mu_2(i), \text{ for all } i\in\mathcal{S}.
\end{aligned} \tag{188}$$

Using the above convergences and the continuity of the Shannon entropy functional $H(\cdot)$, we get that there exist constants $\delta_1 = \delta_1(\epsilon)$ and $\delta_2 = \delta_2(\epsilon)$ such that the probability in (187) may be upper bounded by the probability

$$P\bigg(\exists\ i,j\in\mathcal{S} \text{ such that } \left|\frac{N_{h'}(n,i,j)}{N_{h'}(n,i)} - P_2(j|i)\right| > \delta_1, \left|\frac{N_{h'}(n,i)}{N_{h'}(n)} - \mu_2(i)\right| > \delta_2, \frac{N_a(n)}{n} > \frac{\delta}{2K}(1-\epsilon') \text{ for all } a\in\mathcal{A}\bigg). \tag{189}$$

Noting that $(N_{h'}(n,i,j) - N_{h'}(n,i)P_2(j|i))_{n\geq 0}$ and $(N_{h'}(n,i) - N_{h'}(n)\mu_2(j|i))_{n\geq 0}$ are martingale sequences for all $i,j\in\mathcal{S}$, we may then express (189) as a probability of deviation of martingale sequences from zero, which may be exponentially bounded by using results from [18, Theorem 1.2A].

3) We now upper bound the term in (174). Towards this, we first pick $\epsilon_1 > 0$ satisfying

$$0 < \epsilon_1 \leq \frac{\epsilon}{1 + 2\sum_{i\in\mathcal{S}}\mu_1(i)H(P_1(\cdot|i))}. \tag{190}$$

Then, the following almost sure convergences hold for all $i,j\in\mathcal{S}$:

$$\begin{aligned}
\frac{N_h(n)}{n} &\to \lambda_\delta^*,\\
\frac{N_h(n,i,j)}{N_h(n)} &\to \mu_1(i)P_1(j|i).
\end{aligned} \tag{191}$$

Following the steps leading up to (105), we note that for every choice of $\epsilon' > 0$, there exists $M = M(\epsilon')$ such that (105) holds. We now choose ϵ' such that

$$\begin{aligned}
\frac{T_2(n)}{n} &\geq \frac{N_h(n)}{n}\bigg\{\bigg[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}}(\mu_1(i)P_1(j|i)+\epsilon')\log\frac{\mu_1(i)P_1(j|i)+\epsilon'}{\mu_1(i)+\epsilon'|\mathcal{S}|}\bigg] - \epsilon'\bigg\}\\
&\geq \frac{N_h(n)}{n}\bigg(\sum_{i\in\mathcal{S}}\mu_1(i)(-H(P_1(\cdot|i)))\bigg) - \epsilon_1
\end{aligned} \tag{192}$$

holds for all sufficiently large values of n, where the last line above follows from the continuity of the term within braces as a function of ϵ'. We then have

$$
\begin{aligned}
&P\left(\frac{T_2(n)}{n} - \frac{N_h(n)}{n}\sum_{i\in\mathcal{S}}\mu_1(i)(-H(P_1(\cdot|i))) < -\epsilon\right) \\
&\leq P\Bigg(\frac{T_2(n)}{n} - \frac{N_h(n)}{n}\sum_{i\in\mathcal{S}}\mu_1(i)(-H(P_1(\cdot|i))) < -\epsilon, \left|\frac{N_h(n)}{n} - \lambda_\delta^*\right| \leq \epsilon_1, \\
&\qquad\qquad \left|\frac{N_h(n,i,j)}{N_h(n)} - \mu_1(i)P_1(j|i)\right| \leq \epsilon' \text{ for all } i,j\in\mathcal{S}\Bigg) \\
&+ P\left(\left|\frac{N_h(n)}{n} - \lambda_\delta^*\right| > \epsilon_1\right) + \sum_{i,j\in\mathcal{S}} P\left(\left|\frac{N_h(n,i,j)}{N_h(n)} - \mu_1(i)P_1(j|i)\right| > \epsilon'\right).
\end{aligned} \tag{193}
$$

We now focus on the first term in (193), and notice that for all sufficiently large values of n, this term may be upper bounded as

$$
\begin{aligned}
&P\Bigg((\lambda_\delta^* + \epsilon_1)\sum_{i\in\mathcal{S}}\mu_1(i)(-H(P_1(\cdot|i))) - \epsilon_1 < -\epsilon + (\lambda_\delta^* - \epsilon_1)\sum_{i\in\mathcal{S}}\mu_1(i)(-H(P_1(\cdot|i)))\Bigg) \\
&\leq P\left(\epsilon_1 > \frac{\epsilon}{1 + 2\sum_{i\in\mathcal{S}}\mu_1(i)H(P_1(\cdot|i))}\right) \\
&= 0,
\end{aligned} \tag{194}
$$

where the last line follows from the choice of ϵ_1 in (190). Exponential bounds for the remaining terms in (193) can be obtained similarly as in the analysis of the first term in (186).

Lastly, for the terms in (173), (175), (177) and (178), noting that the left-hand sides of the inequality inside the probability expression in all the three terms converge to zero a.s., similar procedures as used above for (174) and (176) may be used to obtain exponential upper bounds.

This completes the proof of the lemma. ■

Using the result of Lemma 8 in (168), we get that there exist constants $\theta > 0$ and $0 < B < \infty$ independent of L such that the following holds:

$$
\begin{aligned}
\limsup_{L\to\infty} E^\pi\left[\exp\left(\frac{\tau(\pi)}{\log L}\right)\middle| C\right] &\leq \exp\left(\frac{3}{\frac{\delta}{2K}(D(P_1||P_\delta|\mu_1) + D(P_2||P_\delta|\mu_2))}\right) + \limsup_{L\to\infty}\sum_{n\geq l(L,\delta)} B\exp\left(\frac{n+1}{\log L} - n\theta\right) \\
&< \infty,
\end{aligned} \tag{195}
$$

thus establishing that the family $\{\tau(\pi^\star(L,\delta))/\log L : L \geq 1\}$ is uniformly integrable.

Combining the above result on uniform integrability along with the asymptotic bound in (157) yields the desired upper bound in (45), thus completing the proof of the proposition. ■

VIII. Summary

We analysed 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 derived an asymptotic lower bound on the expected stopping time of any policy as a function of error probability. We identified an explicit configuration-dependent constant in the lower bound. Furthermore, we proposed a scheme that (a) is a modification of the classical GLRT, and (b) uses an idea of "forced exploration" from [8]. This scheme takes as inputs two parameters: $L \geq 1$ and $\delta \in (0,1)$. We showed 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 highlighted how to overcome some of the key challenges that the Markov setting offers in the analysis. To the best of our knowledge, the odd arm identification problem (or variants like the best arm identification) in the Markov rewards setting have not been analysed in the literature. Our analysis of the rested Markov setting is a key first step in understanding the different case of restless Markov setting, which is still open.

References

[1] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[2] J. C. Gittins, "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 148–177, 1979.

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

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

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

[6] G. R. Prabhu, S. Bhashyam, A. Gopalan, and R. Sundaresan, "Learning to detect an oddball target with observations from an exponential family," 2017.

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

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

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

[10] 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.

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

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

[13] M. Naghshvar and T. Javidi, "Active m-ary sequential hypothesis testing," in *2010 IEEE International Symposium on Information Theory*. IEEE, 2010, pp. 1623–1627.

[14] ——, "Information utility in active sequential hypothesis testing," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2010, pp. 123–129.

[15] ——, "Performance bounds for active sequential hypothesis testing," in *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 2666–2670.

[16] M. Naghshvar, T. Javidi *et al.*, "Active sequential hypothesis testing," *The Annals of Statistics*, vol. 41, no. 6, pp. 2703–2738, 2013.

[17] A. Wald, "On cumulative sums of random variables," *The Annals of Mathematical Statistics*, vol. 15, no. 3, pp. 283–296, 1944.

[18] H. Victor, "A general class of exponential inequalities for martingales and ratios," *The Annals of Probability*, vol. 27, no. 1, pp. 537–564, 1999.

[19] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*. Springer Science & Business Media,, 2009, vol. 38.

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