Detecting an Odd Restless Markov Arm with a Trembling Hand

P. N. Karthik and Rajesh Sundaresan

Abstract

In this paper, we consider a multi-armed bandit in which each arm is a Markov process evolving on a finite state space. The state space is common across the arms, and the arms are independent of each other. The transition probability matrix of one of the arms (the odd arm) is different from the common transition probability matrix of all the other arms. A decision maker, who knows these transition probability matrices, wishes to identify the odd arm as quickly as possible, while keeping the probability of decision error small. To do so, the decision maker collects observations from the arms by pulling the arms in a sequential manner, one at each discrete time instant. However, the decision maker has a trembling hand, and the arm that is actually pulled at any given time differs, with a small probability, from the one he intended to pull. The observation at any given time is the arm that is actually pulled and its current state. The Markov processes of the unobserved arms continue to evolve. This makes the arms restless.

For the above setting, we derive the first known asymptotic lower bound on the expected time required to identify the odd arm, where the asymptotics is of vanishing error probability. The continued evolution of each arm adds a new dimension to the problem, leading to a family of Markov decision problems (MDPs) on a countable state space. We then stitch together certain parameterised solutions to these MDPs and obtain a sequence of strategies whose expected times to identify the odd arm come arbitrarily close to the lower bound in the regime of vanishing error probability. Prior works dealt with independent and identically distributed (across time) arms and rested Markov arms, whereas our work deals with restless Markov arms.

Index Terms

Multi-armed bandits, restless bandits, odd arm identification, Markov decision process, trembling hand.

I. Introduction

The problem of odd arm identification deals with identifying an anomalous (or *odd*) arm in a multi-armed bandit as quickly as possible, while keeping the probability of decision error small. Here, the term *anomaly* simply means that the law, say ψ_1, of one of the arms is different from the common law, say ψ_2, of each of the other arms. We assume that the arms are independent of each other. A decision maker, who may or may not have prior knowledge of ψ_1 and ψ_2, and whose goal it is to identify the index of the odd arm, samples the arms in a sequential manner, one at a time. The process of sampling the arms continues until the decision maker is sufficiently confident of which arm is odd, at which time he stops further sampling and declares the index of the odd arm. In forming his decision about the odd arm, it is important for the decision maker to ensure

P. N. Karthik is with the Department of Electrical Communication Engineering at the Indian Institute of Science, Bangalore 560012, Karnataka, India. Rajesh Sundaresan is with the Department of Electrical Communication Engineering and the Robert Bosch Centre for Cyber Physical Systems at the Indian Institute of Science, Bangalore 560012, Karnataka, India. Email: (periyapatna, rajeshs)@iisc.ac.in.

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

A shorter version of this paper was presented at the 2020 IEEE International Symposium on Information Theory (ISIT).

that his error probability is low (below a pre-specified threshold). It is natural to expect that smaller the pre-specified error probability threshold, longer the decision maker will have to wait before declaring the odd arm location. The main objective of this paper is to identify the asymptotic growth rate of the decision time as a function of the error probability, where the asymptotics is as the error probability goes to zero.

Prior works on odd arm identification consider the cases when either each arm yields independent and identically distributed (iid) observations [1]–[3], or when each arm yields Markov observations from a common finite state space [4]. When each arm yields iid observations, ψ_1 refers to the law of a random observation coming from the odd arm, while ψ_2 refers to the law of a random observation coming from any of the non-odd arms. When each arm yields Markov observations, ψ_1 refers to the transition law of the Markov process of the odd arm, while ψ_2 refers to the transition law of the Markov process of each of the non-odd arms. When the state space is discrete, the transition laws ψ_1 and ψ_2 may be specified equivalently by the respective transition probability matrices, say P_1 and P_2, where $P_1 \neq P_2$. We use the term 'observation' in place of the commonly used term 'reward' because our focus is on early identification of the odd arm in contrast to reward maximisation or regret minimisation.

An important feature of the setting in [4] is that the Markov process of any given arm evolves by one time step only when the arm is selected, and does not evolve otherwise; this is known as the setting of *rested* arms. In this paper, we partially extend the results of [4] to the more difficult *restless* arms setting in which the Markov process of each arm continues to evolve whether or not the arm is selected. The continued evolution of the Markov process of each arm makes it necessary for the decision maker to keep a record of (a) the time elapsed since each arm was previously selected (called the arm's *delay*), and (b) the state of each arm as observed at its previous selection time (called the *last observed state* of the arm). Notice that the notion of arm delays is superfluous when the arms are rested as in [4] since the unobserved arms remain frozen at their previously observed states. It is also superfluous in the special case of the restless setting when each arm yields iid observations (as in [1]–[3]) because the last observed state of each arm is independent of the arm's current state. Therefore, the notions of arm delays and last observed states are strikingly new features of the setting of general restless Markov arms.

For the rest of this paper, we assume that the transition matrices P_1 and P_2 of the odd arm and the non-odd arm Markov processes are known to the decision maker. Further, we assume that the common state space of the Markov process of each arm is finite as in [4]. All the essential conceptual difficulties related to restless arms remain despite these simplifications. New tools are needed to overcome the difficulties, and these are highlighted in Section I-C. The case when P_1 and P_2 are unknown is beyond the scope of this paper and is currently under study.

A. Motivation and The Notion of a Trembling Hand

Our motivation to study the restless odd Markov arm problem comes from the desire to extend, to more general settings, the decision theoretic formulation of a certain visual search experiment conducted by Sripati and Olson [5] and analysed in Vaidhiyan et al. [1]. In this experiment, human subjects were shown a number of images at once, with one *oddball* image in a sea of *distracter* images. The goal of the experiment was to understand the relationship between (a) the average time taken by the human subject to identify the oddball image, and (b) the dissimilarity between the oddball and distracter images as perceived by the human subject. The images used in the above experiment were static images. Vaidhiyan et al. also conducted experiments with dynamic drifting-dots images (movies), similar to the ones conducted by Krueger et al. [6], in which the dots in each movie location executed Brownian motions with fixed drifts. Further, the drifts were identical in all the distracter movie locations, and were different from the drift in the oddball movie location. In this context, what are optimal strategies

to identify the oddball movie? A systematic analysis of this question, along the lines of [1], requires an understanding of the restless odd Markov arm problem which forms the main subject of this paper.

It is often the case in such visual search experiments that though the subject (or decision maker) intends to focus his attention at a certain location, the actual focus location differs from the intended focus location with a small probability. We model this in our multi-armed bandit setting as a *trembling hand* for the decision maker: with probability $1-\eta$, the decision maker pulls the intended arm, but with probability η, the decision maker pulls a uniformly randomly chosen arm (we use the phrases 'arm pulls' and 'arm selections' interchangeably). Up to Section VI, we assume that $\eta > 0$, as is often the case in visual search experiments such as those described above. The case when $\eta = 0$ is dealt with separately in Section VII.

B. Prior Works on Restless Markov Arms

The topic of restless Markov arms has been studied extensively in the literature in the context of reward maximisation (or equivalently, regret minimisation). In such works, each arm is assumed to yield, upon being sampled, an immediate 'reward' based on the arm's current state. Regret is then defined as the difference between the expected sum of rewards obtained under a particular arm selection scheme and that obtained by a scheme that knows which arm yields the highest expected reward. Whittle [7] refined and extended the results of Gittins [8] on the optimality, in the setting of rested arms, of a certain index-based policy. Whittle [7] demonstrated that Gittins's policy in [8] is not necessarily optimal in the context of restless arms, introduced a new index (now called *Whittle's index*) which could be computed if each arm satisfied an *indexability* condition, and demonstrated that the new index coincides with Gittins's index in the rested setting. Yet, as Whittle showed, the new index-based policy is not necessarily optimal for the general setting of restless arms.

Whittle's results require the Markov transition laws of each of the arms to be known beforehand. Extensions of Whittle's results to the case when the laws are not known beforehand appear in Liu et al. [9]. Ortner et al. [10] provide a policy that, when the transition laws of the arms are unknown, gives a regret of the order $O(\sqrt{T})$ after T time steps in relation to a policy that knows the Markov transition laws of all the arms. As Ortner et al. show in [10], an optimal policy for the restless bandit problem does not necessarily pick the arm with the largest stationary mean at each time instant[1], but instead switches between the arms in an optimal fashion. Working on this key idea, Grünelwalder et al. [12] provide conditions under which the problem of finding the arm with the largest stationary mean serves as a "good" approximation to the original problem of finding the optimal arm switching strategy when each arm is a stationary ϕ-mixing process and the arms are restless. The works [10] and [12] deal with general state spaces (i.e., not necessarily finite or countable) and address the associated technical challenges.

In contrast to all the works mentioned above, this paper focuses on the *stopping* problem of identifying the index of the odd arm as quickly as possible. It is worth noting here, as also noted in [13], that policies which are optimal for the problem of minimising regret may not necessarily be optimal in the stopping problem context.

For applications of the restless odd Markov arm problem, see [4]. For a related problem of best arm identification instead of odd arm identification, see [14], [15].

C. An Overview of the Results and Our Contributions

We now provide an overview of our results and highlight our contributions.

1) We show that given a pre-specified error probability threshold $\epsilon > 0$, the expected time taken by the decision maker to identify the index of the odd arm with probability of error at most ϵ grows as $\Theta(\log(1/\epsilon))$. We give a precise

[1]This is indeed the case in a multi-armed bandit problem with iid observations from each arm, as was shown in [11].

characterisation of the best (smallest) constant multiplying $\log(1/\epsilon)$, which we call $R^*(P_1, P_2)$, in terms of the Markov transition probability matrices P_1 and P_2. This is the first known characterisation of this constant for the setting of restless Markov arms. See Section IV for an exact mathematical expression. We prove this by first showing a lower bound in Section IV and then a matching asymptotic upper bound in Section V.

2) An examination of the lower bounds in the prior works [1]–[4] reveals that the best constant multiplier in these works is the solution to an optimisation problem having an outer supremum over all (unconditional) probability distributions on the arms, followed by an inner minimum over all alternative odd arm locations (i.e., a sup-min optimisation problem). A further examination reveals that when arm h is the odd arm, there exists a probability distribution λ_h^* on the arms, possibly depending on the odd arm location h, that (a) attains the outer supremum, and (b) puts equal mass on each of the non-odd arm locations.

   Along lines similar to those of the prior works, we show that the best constant multiplier $R^*(P_1, P_2)$ is the solution to a sup-min optimisation problem in which the supremum is over all *conditional* probability distributions on the arms, conditioned on arm delays and last observed states, and the minimum is over all alternative odd arm locations. We also show that the constant $R^*(P_1, P_2)$ is not a function of the actual odd arm location; this is due to symmetry in the structure of the arms. The constant $R^*(P_1, P_2)$ represents the amount of effort required to identify the true odd arm location by guarding against identifying the nearest, incorrect alternative odd arm location.

   However, given an odd arm location h, the question of whether there exists a conditional probability distribution that attains the supremum in the expression for $R^*(P_1, P_2)$ is still under study.

3) In order to derive the constant $R^*(P_1, P_2)$, we use the fact that the arm delays and the last observed states form a *controlled Markov process*, with the arm selections playing the role of *controls*. This approach of ours takes into account the delays and the last observed states of *all* the arms jointly. In contrast, the approaches of [1]–[4] suggest dealing with the delays and the last observed states of each of the arms separately, which we view as a 'local' perspective of the arm delays and the last observed states. In Section IV-A, we show that this local perspective of arm delays and last observed states leads to an infinite dimensional, constrained, linear programming problem (LPP). The drawback of this approach is that it is not easy to find the tightest set of constraints for the LPP. As a consequence, the constant multiplier obtained as the solution to the LPP may not necessarily be the best (smallest).

   On the other hand, our 'lift' approach, which considers the delays and the last observed states of all the arms jointly, leads us naturally to a family of Markov decision problems (MDPs) and, in turn, provides the necessary perspective to arrive at the best constant multiplier $R^*(P_1, P_2)$.

4) We show that under a *stationary* arm selection policy (in which at each time, the arms are selected according to a certain conditional probability distribution on the arms, conditioned on the delays and last observed states at that time), the aforementioned controlled Markov process is, in fact, a Markov process. Additionally, we show that under every stationary arm selection policy, this Markov process is *ergodic* when the trembling hand parameter $\eta > 0$ (Lemma 1). It is this ergodicity property, together with the strict positivity of the trembling hand parameter η, that plays a crucial role in our analysis of the lower and the upper bounds. The case $\eta = 0$ demands a careful examination since, in this case, such an ergodicity property is not readily available for every stationary arm selection policy.

5) We show that for every arm selection policy of the decision maker, stationary or otherwise, what enters into the analyses of the lower and the upper bounds is the following statistic: for each possible value of arm delays $\underline{d}$, last observed states $\underline{i}$ and arm a, the long-term fraction of times the aforementioned controlled Markov process visits the state $(\underline{d}, \underline{i})$ and

arm a is selected. This fact, together with [16, Theorem 8.8.2], enables us to restrict attention only to stationary arm selection policies in arriving at the best constant multiplier $R^*(P_1, P_2)$.

In spite of the above simplification, the computability of $R^*(P_1, P_2)$ remains an issue since it involves a search over the space of all stationary arm selection policies. One must resort to Q-learning in the context of restless Markov arms [17] to compute $R^*(P_1, P_2)$.

6) The question of whether the supremum in the expression for $R^*(P_1, P_2)$ is attainable is still under study, as mentioned in point 2 above. The arm delays, being positive and integer-valued, introduce a countably infinite dimension to the problem. As a consequence, it is not clear if the space of all conditional distributions on the arms, conditioned on the arm delays and the last observed states, is compact. In the iid and the rested Markov settings of the prior works, only unconditional distributions on the arms appear in the analysis of the lower and the upper bounds, and because of the finite nature of the number of arms, it follows immediately that the space of all unconditional distributions on the arms is compact. Such a compactness property plays a key role in showing that the supremum is attained.

   Notwithstanding the additional technical difficulty encountered in the setting of restless arms due to the presence of the countably infinite-valued arm delays, we show that the supremum in the expression for $R^*(P_1, P_2)$ may be approached arbitrarily closely by stitching together certain parameterised solutions to the MDPs mentioned in point 3 above. We present the details in Section V.

7) The trembling hand model (with $\eta > 0$) may be viewed as a regularisation that ensures stability of the aforementioned controlled Markov process (of arm delays and last observed states) for free. If $\eta = 0$, one could deliberately add some regularisation parameterised by η, re-label the constant $R^*(P_1, P_2)$ in this case as $R^*_\eta(P_1, P_2)$ for each $\eta > 0$, and analyse the limiting value of $R^*_\eta(P_1, P_2)$ as $\eta \downarrow 0$. We show that in this case, (a) the limit of $R^*_\eta(P_1, P_2)$ as $\eta \downarrow 0$ exists, and (b) the upper bound is governed by $\lim\limits_{\eta \downarrow 0} R^*_\eta(P_1, P_2)$, while the lower bound is governed by $R^*_0(P_1, P_2)$ (which is obtained by plugging $\eta = 0$ in the expression for $R^*_\eta(P_1, P_2)$). So, the question then is, do these lower and the upper bounds match? In Section VII, we are only able to establish that $\lim\limits_{\eta \downarrow 0} R^*_\eta(P_1, P_2) \leq R^*_0(P_1, P_2)$. A key tool needed to establish equality in this inequality is the "envelope theorem" [18, Theorem 2]. A verification of the hypotheses of the envelope theorem for the setting of restless arms still remains open.

8) We verify that the envelope theorem holds in the iid and rested Markov settings of the prior works [1]–[4], thus leading to matching upper and lower bounds in these works. Thus, sufficient conditions for the upper and the lower bounds to match are either (a) $\eta > 0$, or (b) $\eta = 0$ and the observations come from either iid or rested Markov arms.

D. Organisation of the Paper

The rest of this paper is organised as follows. In Section II, we set up the notations that we use throughout the paper. In Section III, we provide some preliminaries on MDPs. In Section IV, we present the lower bound on the expected time to identify the odd arm as a function of the error probability for the setting of restless Markov arms. In the same section, we also show that by following the conventional approaches available in the prior works, we arrive at an infinite-dimensional linear programming problem (LPP) with countably infinitely many constraints. In Section V, we present a sequence of strategies whose expected times to identify the odd arm approach that of the lower bound in the limit of vanishing error probabilities, following which we state the main result of this paper in Section VI. We discuss the no trembling hand case in Section VII. The proofs of all the results are contained in Appendices A-E. We conclude the paper in Section VIII.

II. Notations and Problem Formulation

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

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

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

A. Policy

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

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

We use π to denote a generic policy, and let $\tau(\pi)$ denote the stopping time of policy π. Throughout this paper, all stopping times are defined with respect to the filtration $\mathcal{F}_t \coloneqq \sigma(B^{t-1}, A^{t-1}, \bar{X}^{t-1})$, $t \geq 1$ and $\mathcal{F}_0 \coloneqq \{\Omega, \emptyset\}$. Let $\theta(\tau(\pi))$ denote the index of the odd arm declared by the policy π at its stopping time $\tau(\pi)$.

Let $P_h^\pi(\cdot)$ and $E_h^\pi[\cdot]$ denote probabilities and expectations computed under policy π. For ease of notation, we drop the superscript π, and beg the gentle reader to bear the dependence on π in mind. Given a target probability of error $\epsilon > 0$, we define $\Pi(\epsilon)$ as the set

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

of all policies whose probability of error at stoppage is below ϵ for all possible odd arm locations. We emphasise that policies in $\Pi(\epsilon)$ work for all possible odd arm locations. We anticipate from similar results in the prior works that

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

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

B. Delays and Last Observed States

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

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

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

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

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

where $\bar{X}_t$ is the state of the arm $A_t = a'$ at time t.

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

C. Controlled Markov Process and the Resulting Markov Decision Problem

From the update rule in (2), it is clear that the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ takes values in a subset $\mathbb{S}$ of the countable set $\mathbb{N}^K \times \mathcal{S}^K$, where $\mathbb{N} = \{1, 2, \ldots\}$ denotes the set of natural numbers. The subset $\mathbb{S}$ is formed based on the constraint that at

[2]Let us note here that this may not always be the case, and the intended arm selections may differ from the actual arms selected. If this is the case, the decision maker may exercise arm selections uniformly at random until he observes that each arm is selected as indicated. By virtue of the fact that the trembling hand parameter $\eta > 0$, the probability of selecting any arm is strictly positive at each time instant. As a result, it can be shown that the above exercise of pulling arms randomly till each arm is selected as indicated will take only finite time almost surely.

any time $t \geq K$, exactly one of the components of $\underline{d}(t)$ is equal to 1, and all the other components are strictly greater than 1. Note that for all $(\underline{d}, \underline{i}) \in \mathbb{S}$ and $t \geq K$,

$$P(\underline{d}(t+1) = \underline{d}, \underline{i}(t+1) = \underline{i} \mid (\underline{d}(s), \underline{i}(s)), B_s,\ K \leq s \leq t) = P(\underline{d}(t+1) = \underline{d}, \underline{i}(t+1) = \underline{i} \mid (\underline{d}(t), \underline{i}(t)), B_t). \tag{3}$$

On account of (3) being satisfied, we say that under any policy π, the evolution of the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is *controlled* by the sequence $\{B_t\}_{t \geq 0}$ of intended arm selections under policy π. Alternatively, we say that $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is a controlled Markov process, with $\{B_t\}_{t \geq 0}$ as the sequence of controls; the terminology used here follows that of Borkar [19]. Thus, we are in a Markov decision problem (MDP) setting. We now make precise the state space, the action space, the transition probabilities and our objective.

The state space of the MDP is $\mathbb{S}$, with the state at time t denoted $(\underline{d}(t), \underline{i}(t))$. The action space of the MDP is $\mathcal{A}$, with action B_t at time t possibly depending on the previous actions B^{t-1} and the previous states $\{(\underline{d}(s), \underline{i}(s)), K \leq s \leq t\}$. (It is easy to see that this is equivalent to taking an action based on $(B^{t-1}, A^{t-1}, \bar{X}^{t-1})$.) The transition probabilities for the MDP are given by

1) the trembling hand rule
$$P(A_t = a | B_t) = \frac{\eta}{K} + (1 - \eta)\ \mathbb{I}_{\{B_t = a\}}, \quad \forall a \in \mathcal{A}, \tag{4}$$
2) the law associated with arm A_t, and
3) the update rule (2).

In (4), $\mathbb{I}$ denotes the indicator function. In order to write the transition probabilities of the MDP precisely, let us introduce some notations. Given $h, a \in \mathcal{A}$, let P_h^a denote the transition probability matrix of the Markov process of arm a under the hypothesis $\mathcal{H}_h$. That is,

$$P_h^a = \begin{cases} P_1, & a = h, \\ P_2, & a \neq h. \end{cases} \tag{5}$$

Furthermore, for any integer $d \geq 1$, let $(P_h^a)^d$ denote the transition probability matrix obtained by multiplying P_h^a with itself d times. Then, given any $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$ and $b \in \mathcal{A}$, the transition probabilities for the MDP are given by

$$\begin{aligned} &P(\underline{d}(t+1) = \underline{d}', \underline{i}(t+1) = \underline{i}' \mid \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, B_t = b) \\ &\qquad = \begin{cases} \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b\}}(a)\right)\ (P_h^a)^{d_a}(i'_a | i_a), & \text{if } d'_a = 1 \text{ and } d'_{\tilde{a}} = d_{\tilde{a}} + 1 \text{ for all } \tilde{a} \neq a, \\ & i'_{\tilde{a}} = i_{\tilde{a}} \text{ for all } \tilde{a} \neq a, \\ 0, & \text{otherwise}, \end{cases} \end{aligned} \tag{6}$$

where d'_a and i'_a in (6) denote the component corresponding to arm a in $\underline{d}'$ and $\underline{i}'$ respectively. Note that the transition probabilities defined in (6) are stationary and independent of time. Also, for $a \in \mathcal{A}$, we have

$$\begin{aligned} &P(\underline{d}(t+1) = \underline{d}', \underline{i}(t+1) = \underline{i}' \mid \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, A_t = a) \\ &\qquad = \begin{cases} (P_h^a)^{d_a}(i'_a | i_a), & \text{if } d'_a = 1 \text{ and } d'_{\tilde{a}} = d_{\tilde{a}} + 1 \text{ for all } \tilde{a} \neq a, \\ & i'_{\tilde{a}} = i_{\tilde{a}} \text{ for all } \tilde{a} \neq a, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{7}$$

The left-hand sides of (6) and (7) differ in that B_t in (6) is replaced by A_t in (7). We shall write $Q(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, a)$ to denote the quantity in (7).

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

$$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)}. \tag{8}$$

In the next section, we provide some preliminaries on MDPs. The terminologies used follow Borkar [19].

III. Preliminaries on MDPs

Let π be an arbitrary policy. Consider the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$, with the corresponding sequence of controls $\{B_t\}$, under the policy π. Note that for all $t \geq K$,

$$\begin{aligned}
&P(\underline{d}(t+1) = \underline{d}, \underline{i}(t+1) = \underline{i} \mid B^{t-1}, \{(\underline{d}(s), \underline{i}(s)),\ K \leq s \leq t\}) \\
&= \sum_{b=1}^{K} P(B_t = b \mid B^{t-1}, \{(\underline{d}(s), \underline{i}(s)),\ K \leq s \leq t\})\, P(\underline{d}(t+1) = \underline{d}, \underline{i}(t+1) = \underline{i} \mid B_t = b,\ B^{t-1}, \{(\underline{d}(s), \underline{i}(s)),\ K \leq s \leq t\}) \\
&= \sum_{b=1}^{K} P(B_t = b \mid B^{t-1}, \{(\underline{d}(s), \underline{i}(s)),\ K \leq s \leq t\}) \cdot P(\underline{d}(t+1) = \underline{d}, \underline{i}(t+1) = \underline{i} \mid (\underline{d}(t), \underline{i}(t)), B_t = b),
\end{aligned} \tag{9}$$

where the last line above follows from (3). From (9), it is evident that the policy π may be described completely by specifying $P(B_t | B^{t-1}, \{(\underline{d}(s), \underline{i}(s)), K \leq s \leq t\})$ for all $t \geq K$. We say that a policy π is a *stationary randomised strategy* (SRS) if there exists a Cartesian product λ of the form

$$\lambda = \bigotimes_{(\underline{d}, \underline{i}) \in \mathbb{S}} \lambda_{(\underline{d}, \underline{i})}, \tag{10}$$

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

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

Such an SRS π will be denoted π^λ. Note that $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is indeed a *Markov process* under the SRS π^λ. This follows from the relation (9) where the first probability term inside the summation in (9) is now a function only of $(\underline{d}(t), \underline{i}(t))$. Let Π_{SRS} denote the set of all SRS policies.

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

An immediate and important property of any $\pi^\lambda \in \Pi_{\mathsf{SRS}}$ is the following.

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

Proof: See Appendix A. ∎

The proof of Lemma 1 relies on the hypothesis that the trembling hand parameter $\eta > 0$.

As a consequence of Lemma 1, it follows that under every SRS policy, a unique stationary distribution exists for the Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$. Let us call this stationary distribution μ^λ corresponding to the SRS policy π^λ.

With the above ingredients in place, we state in the next section the first main result of this paper – an asymptotic lower bound on the expected time to identify the odd arm.

IV. Lower Bound

We now present a lower bound for (8). Given two probability distributions μ and ν on the finite state space $\mathcal{S}$, the Kullback-Leibler (KL) divergence (also called the relative entropy) between μ and ν is defined as

$$D(\mu\|\nu) \coloneqq \sum_{i\in\mathcal{S}} \mu(i)\log\frac{\mu(i)}{\nu(i)}, \tag{11}$$

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

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

$$\liminf_{\epsilon\downarrow 0}\ \inf_{\pi\in\Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} \geq \frac{1}{R^*(P_1,P_2)}, \tag{12}$$

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

$$R^*(P_1,P_2) \coloneqq \sup_{\pi^\lambda\in\Pi_{SRS}} \min_{h'\neq h} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \nu^\lambda(\underline{d},\underline{i},a)\, k(\underline{d},\underline{i},a), \tag{13}$$

with

$$k(\underline{d},\underline{i},a) \coloneqq \begin{cases} D(P_1^{d_a}(\cdot|i_a)\|P_2^{d_a}(\cdot|i_a)), & a=h,\\ D(P_2^{d_a}(\cdot|i_a)\|P_1^{d_a}(\cdot|i_a)), & a=h',\\ 0, & a\neq h,h', \end{cases} \tag{14}$$

and

$$\nu^\lambda(\underline{d},\underline{i},a) \coloneqq \mu^\lambda(\underline{d},\underline{i})\left(\frac{\eta}{K}+(1-\eta)\,\lambda(a|\underline{d},\underline{i})\right), \quad \forall(\underline{d},\underline{i},a)\in\mathbb{S}\times\mathcal{A}. \tag{15}$$

Proof: See Appendix B. ∎

The proof of the lower bound follows the outline in [4], with necessary modifications for the setting of restless arms. The key ingredients are the data processing inequality for relative entropies, a Wald-type Lemma for Markov processes, and a recognition that, for any $(\underline{d},\underline{i})$, the long-term fraction of exits from the state $(\underline{d},\underline{i})$ matches the long-term fraction of entries into the state $(\underline{d},\underline{i})$. This forces the long-term probability of seeing the controlled Markov process in the state $(\underline{d},\underline{i})$ to be that under its unique stationary distribution, by ergodicity (Lemma 1). These observations lead to (12).

Observe that the left-hand side of (12) is evaluated by taking into consideration *all* policies, including those that are not necessarily SRS policies, whereas the supremum in (13) is only over SRS policies. This is a consequence of [16, Theorem 8.8.2]; see Appendix B for more details. Also, the constant $R^*(P_1,P_2)$ in (13) does not depend on h. This is due to symmetry in the structure of the arms.

A. An Infinite-Dimensional Linear Programming Problem

It may be a little surprising to the reader as to why the summation on the right-hand side of $R^*(P_1,P_2)$ in (13) is over the delays and the last observed states of *all* the arms when the function $k(\underline{d},\underline{i},a)$, as given in (14), is a function only of d_a and i_a, the delay and the last observed state of arm a. In order to better appreciate the usefulness of taking into account the arm delays and last observed states of all the arms in deriving the lower bound, we present below a sketch proof of a possibly

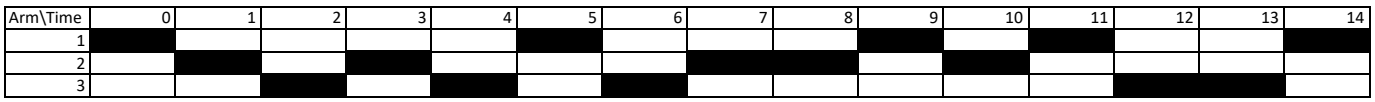

Fig. 1. A schematic representation of arm selections over time for $K = 3$ arms. In this schematic, an arm selected at any given time is indicated by a black box. Note that arm 1 is selected at time $t = 0$, arm 2 at time $t = 1$ and arm 3 at time $t = 2$. Thereafter, for $t \geq 3$, arm 1 is selected at certain time instants and is not selected at certain other time instants. Whenever arm 1 is not selected, *some* other arm is selected, as a consequence of which the delay of arm 1 increases, and it is this fact that must be captured as a constraint on the delays of arm 1. Similar constraints apply for each of the other arms.

weaker lower bound in which we first fix an arm a and consider only its delay and last observed state in the subsequent calculations. Fix arm $a \in \mathcal{A}$. Given an integer $d \geq 1$, $i, j \in \mathcal{S}$ and a policy π, let

$$N(\tau(\pi), d, i, a, j) \coloneqq \sum_{t=K}^{\tau(\pi)} \mathbb{I}_{\{d_a(t)=d, i_a(t)=i, A_t=a, X_t^a=j\}}. \tag{16}$$

Recall that $\tau(\pi)$ denotes the stopping time of policy π. Following the earlier approaches of [1]–[4], and using the data processing inequality, one arrives at[3]

$$\sum_{a=1}^{K} \sum_{d=1}^{\infty} \sum_{i \in \mathcal{S}} E_h[N(\tau(\pi), d, i, a)]\, D((P_h^a)^d(\cdot|i) \| (P_{h'}^a)^d(\cdot|i)), \tag{17}$$

where $N(\tau(\pi), d, i, a)$ in (17) is simply the summation over all $j \in \mathcal{S}$ of the right-hand side of (16).

From the exposition in Section II, we know that at any given time $t \geq K$, the vector $\underline{d}(t)$ must satisfy the following constraint: exactly one component of $\underline{d}(t)$ is equal to 1, and all the other components are strictly greater than 1. Let us now express this constraint mathematically. Recall the assumption that the policy π selects, without loss of generality, arm 1 at time $t = 0$, arm 2 at time $t = 1$ and so on until arm K at time $t = K - 1$. From time $t = K$ onwards, arm a may or may not be selected at all time instants, and whenever it is not selected, *some* arm $b \neq a$ is selected. It is this observation (that some arm is selected at every time instant until the stopping time of the policy) that must be modelled as a constraint mathematically. Figure 1 depicts the selection of arms at various time instants for the case when $K = 3$.

Assume without loss of generality that under the policy π, arm a is selected at time $t = \tau(\pi)$. Then, it follows that

$$(a-1) + \sum_{i \in \mathcal{S}} \sum_{d=1}^{\infty} d\, N(\tau(\pi), d, i, a) + 1 = \tau(\pi) + 1; \tag{18}$$

in (18), the term $(a - 1)$ on the left-hand side denotes the number of time instants that have passed before arm a is selected for the first time. The second term on the left-hand side of (18) denotes the total number of time instants that have passed, starting from time $t = K$, until the final selection time instant of arm a. The last term on the left-hand side of (18) counts the final selection instant of arm a. Thus, the total value of the left-hand side of (18) is equal to the total number of time instants that have passed from $t = 0$ to $t = \tau(\pi)$ (both inclusive), which is precisely the quantity on the right-hand side of (18). Applying $E_h[\cdot]$ to both sides of (18), and using the monotone convergence theorem, we arrive at the following relation after some rearrangement:

$$\sum_{i \in \mathcal{S}} \sum_{d=1}^{\infty} d\, \frac{E_h[N(\tau(\pi), d, i, a)]}{E_h[\tau(\pi)]} + \frac{a-1}{E_h[\tau(\pi)]} = 1. \tag{19}$$

[3]For the gentle reader interested in the details, this can be obtained by following the chain of equalities leading up to (88) in Appendix B, with the inner summation over $(\underline{d}, \underline{i})$ now replaced by a summation over $(d, i) \in \{1, 2, \ldots\} \times \mathcal{S}$.

In fact, it is easy to see that (18), and therefore (19), holds for every arm, whether or not the arm is selected at time $t = \tau(\pi)$. Mimicking the steps in Appendix B, and using (17) in place of (88) in Appendix B along with the constraint in (19), we arrive at the following relation in place of (93):

$$d(\epsilon, 1-\epsilon) \leq \sup_{\kappa} \min_{h' \neq h} \Bigg\{ E_h \left[\sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)} \right] + \Big(E_h[\tau(\pi) - K + 1] \Big) \cdot \sum_{a=1}^{K} \sum_{d=1}^{\infty} \sum_{i \in \mathcal{S}} \kappa(d,i,a)\, D((P_h^a)^d(\cdot|i) \| (P_{h'}^a)^d(\cdot|i)) \Bigg\}, \tag{20}$$

where $d(\epsilon, 1-\epsilon)$ is the relative entropy between a Bernoulli distribution with parameter ϵ and a Bernoulli distribution with parameter $1-\epsilon$, and the supremum in (20) is over all probability distributions κ on $\{1, 2, \ldots\} \times \mathcal{S} \times \mathcal{A}$ that satisfy the constraint

$$\sum_{i \in \mathcal{S}} \sum_{d=1}^{\infty} d\ \kappa(d,i,a) = 1 \quad \text{ for all } a \in \mathcal{A}. \tag{21}$$

The constraint in (21) may be obtained from (19) by letting $E_h[\tau(\pi)] \to \infty$ (which is the same as $\epsilon \downarrow 0$) and replacing the fractional term on the left-hand side of (19) by $\kappa(d,i,a)$; here, $\kappa(d,i,a)$ represents the long-term joint probability of observing arm a to have a delay d and last observed state i, and subsequently selecting arm a.

Dividing both sides of (20) by $d(\epsilon, 1-\epsilon)$, and using the fact that $d(\epsilon, 1-\epsilon)/\log(1/\epsilon) \to 1$ as $\epsilon \downarrow 0$, we arrive at

$$\liminf_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} \geq \frac{1}{R_1^*(P_1, P_2)}, \tag{22}$$

where $R_1^*(P_1, P_2)$ is the solution to the following constrained optimisation problem:

$$R_1^*(P_1, P_2) = \sup_{\kappa} \min_{h' \neq h} \sum_{a=1}^{K} \sum_{d=1}^{\infty} \sum_{i \in \mathcal{S}} \kappa(d,i,a)\, D((P_h^a)^d(\cdot|i) \| (P_{h'}^a)^d(\cdot|i))$$

subject to

$$\begin{aligned} &\sum_{i \in \mathcal{S}} \sum_{d=1}^{\infty} d\ \kappa(d,i,a) = 1 \quad \text{ for all } a \in \mathcal{A}, \\ &\sum_{d=1}^{\infty} \sum_{i \in \mathcal{S}} \sum_{a=1}^{K} \kappa(d,i,a) = 1, \\ &\kappa(d,i,a) \geq 0 \quad \text{ for all } d \in \{1, 2, \ldots\},\ i \in \mathcal{S},\ a \in \mathcal{A}. \end{aligned} \tag{23}$$

Notice that (23) constitutes an infinite-dimensional linear programming problem with linear constraints. It is not clear if a solution to (23) exists. Also, it is not clear if the constraints in (23) constitute the tightest set of constraints. From Proposition 1, we must of course have $R_1^*(P_1, P_2) \geq R^*(P_1, P_2)$.

We end this section with a remark that by taking into account the delays and the last observed states of all the arms in deriving the lower bound, as done in Appendix B, the constraint in (18) is automatically captured since any vector $\underline{d} = (d_1, \ldots, d_K)$ of arm delays belongs, by definition, to the subset $\mathbb{S}$ which obeys the constraint in (18). Thus, the viewpoint of controlled Markov processes greatly simplifies the analysis of the lower bound. The key insight of this paper is that our 'lift' approach of considering the arm delays and the last observed states of all the arms jointly, instead of dealing with the delays and last observed states of each arm separately, makes the problem amenable to analysis.

V. Achievability

The question of whether the supremum in (13) is attainable is still under study. Recall that this supremum is over all $\pi^\lambda \in \Pi_{\mathsf{SRS}}$ for $\lambda(\cdot|\cdot)$ which are conditional probability distributions on the arms, conditioned on the arm delays and the

last observed states. This is in contrast to the works [1]–[4], where the corresponding supremum is over all *unconditional* probability distributions on the arms. This is because, in those works, the arm delays are superfluous. The unconditional probability measures are elements of the probability simplex on $\mathcal{A}$, whereas the conditional probability measures are more complex due to the countably many possible values for the arm delays. In spite of this added complexity, we can come arbitrarily close to the supremum in (13). We shall use this fact in our achievability result, which is the topic of this section.

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

$$\begin{aligned} Z_{hh'}(n) &= \log \frac{P_h(B^n, A^n, \bar{X}^n)}{P_{h'}(B^n, A^n, \bar{X}^n)} \\ &= \log \frac{P_h(B_0)}{P_{h'}(B_0)} + \log \frac{P_h(A_0|B_0)}{P_{h'}(A_0|B_0)} + \log \frac{P_h(\bar{X}_0|B_0, A_0)}{P_{h'}(\bar{X}_0|B_0, A_0)} && (24) \\ &\quad + \sum_{t=1}^{n} \log \left(\frac{P_h(B_t|B^{t-1}, A^{t-1}, \bar{X}^{t-1})}{P_{h'}(B_t|B^{t-1}, A^{t-1}, \bar{X}^{t-1})} \right) && (25) \\ &\quad + \sum_{t=1}^{n} \log \left(\frac{P_h(A_t|B^{t}, A^{t-1}, \bar{X}^{t-1})}{P_{h'}(A_t|B^{t}, A^{t-1}, \bar{X}^{t-1})} \right) && (26) \\ &\quad + \sum_{t=1}^{n} \log \left(\frac{P_h(\bar{X}_t|A_t, B^{t}, A^{t-1}, \bar{X}^{t-1})}{P_{h'}(\bar{X}_t|A_t, B^{t}, A^{t-1}, \bar{X}^{t-1})} \right). && (27) \end{aligned}$$

We now note that under the policy π, the probability of choosing arm B_t at time t, based on the history up to time t, cannot be a function of the underlying odd arm location (which is unknown to π), and must therefore be the same under hypotheses $\mathcal{H}_h$ and $\mathcal{H}_{h'}$. Thus, the first term in (24) and the expression in (25) are 0. Also, we note that $P_h(A_0|B_0) = P_{h'}(A_0|B_0)$, and for each t,

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

since A_t, the arm that is actually pulled at time t, is a function only of B_t and is related to B_t through (4). Therefore, given the history, the choice of A_t is not a function of the odd arm location, and is the same under hypotheses $\mathcal{H}_h$ and $\mathcal{H}_{h'}$, implying that the second term in (24) and the expression in (26) are 0. Finally, the probabilities in (27) do not depend on the intended arm pulls $\{B_t\}$ since the state $\bar{X}_t$ observed on arm A_t is a function only of the delay and the last observed state of arm A_t. Letting X_t^a denote the state of arm $A_t = a$, and defining

$$N(n, \underline{d}, \underline{i}, a) \coloneqq \sum_{t=K}^{n} \mathbb{I}_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a\}}, \tag{28}$$

$$N(n, \underline{d}, \underline{i}, a, j) \coloneqq \sum_{t=K}^{n} \mathbb{I}_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a, X_t^a=j\}}, \tag{29}$$

for all $(\underline{d}, \underline{i}, a) \in \mathbb{S} \times \mathcal{A}$, it can be shown that

$$\begin{aligned} Z_{hh'}(n) &= \sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)} + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \sum_{a=1}^{K} N(n, \underline{d}, \underline{i}, a, j) \log \frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)} && (30) \\ &= \sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)} + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \left[N(n, \underline{d}, \underline{i}, h, j) \log \frac{P_1^{d_h}(j|i_h)}{P_2^{d_h}(j|i_h)} + N(n, \underline{d}, \underline{i}, h', j) \log \frac{P_2^{d_{h'}}(j|i_{h'})}{P_1^{d_{h'}}(j|i_{h'})} \right]. && (31) \end{aligned}$$

Eq. (31) follows by noting that

$$P_h^a = \begin{cases} P_1, & a = h, \\ P_2, & a \neq h, \end{cases} \qquad P_{h'}^a = \begin{cases} P_1, & a = h', \\ P_2, & a \neq h', \end{cases} \tag{32}$$

and thus the only nonzero terms in the summation over the arms in (30) are those corresponding to $a = h$ and $a = h'$.

To describe our policy, we first fix constants $\delta > 0$ and $L > 1$. These will be the parameters of our policy. Recall that the supremum in (13) is over all SRS policies. By the definition of this supremum, we know that for any fixed hypothesis $\mathcal{H}_h$ and given $\delta > 0$, there exists $\lambda(\cdot \mid \cdot) = \lambda_{h,\delta}(\cdot \mid \cdot)$ such that under the corresponding SRS policy $\pi^{\lambda_{h,\delta}}$, we have

$$\min_{h' \neq h} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu^{\lambda_{h,\delta}}(\underline{d}, \underline{i}, a)\ k(\underline{d}, \underline{i}, a) \geq \frac{R^*(P_1, P_2)}{1+\delta}. \tag{33}$$

Notice that $\lambda_{h,\delta}$ is, in general, a function of δ and the hypothesis $\mathcal{H}_h$ (the hypothesis that arm h is the odd arm), although $R^*(P_1, P_2)$ itself is not a function h.

Our policy, which we call $\pi^\star(L, \delta)$, is then as below.

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

Fix $L > 1$ and $\delta > 0$. Let the parameter of the trembling hand be $\eta \in (0, 1]$. Assume[4] that $A_0 = 1$, $A_1 = 2$, and so on until $A_{K-1} = K$. Let $M_h(n) = \min_{h' \neq h} Z_{hh'}(n)$. Follow the below mentioned steps for each $n \geq K$.

(1) Let $\theta(n) = \arg\max_{h \in \mathcal{A}} M_h(n)$; resolve ties at random.

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

(3) If $M_{\theta(n)}(n) < \log((K-1)L)$, decide to pull arm B_n according to the distribution $\lambda_{\theta(n),\delta}(\cdot \mid \underline{d}(n), \underline{i}(n))$.

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

Recall that the supremum in (13) is only over SRS policies. However, the policy $\pi^\star(L, \delta)$ described above is *not* an SRS policy since the distribution in item (3) is a function of $\theta(n)$ that could potentially depend on the entire history of arm selections and observations up to time n. Yet, as we show below, its performance comes arbitrarily close to that of the lower bound.

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

We now present results on the performance of our policy.

Lemma 2. *Fix $L > 1$, $\delta > 0$ and $h \in \mathcal{A}$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Consider the non-stopping version of the policy $\pi^\star(L, \delta)$ which runs indefinitely (i.e., even if item (2) is true, it moves to item (3)). Under this policy, for every $h' \neq h$,*

$$\liminf_{n \to \infty} \frac{Z_{hh'}(n)}{n} > 0 \quad \textit{almost surely}. \tag{34}$$

Proof: See Appendix C. ∎

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

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

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

[4]If this is not the case, exercise arm pulls uniformly at random until each arm is selected at least once. It can be shown that this will only take finite time almost surely, and does not affect the asymptotic analysis of our policy.

Proof: The proof uses the fact that the policy stops in finite time with probability 1. See Appendix D for the details. ■

With the above ingredients in place, we state the main result of this section, which is that the expected stopping time of our policy satisfies an asymptotic upper bound that comes arbitrarily close to the lower bound in (12).

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

$$\limsup_{L\to\infty} \frac{E_h[\tau(\pi^\star(L,\delta))]}{\log L} \leq \frac{1+\delta}{R^*(P_1, P_2)}. \tag{35}$$

Proof: In the proof, which we provide in Appendix E, we first show that as $L \to \infty$ (equivalently $\epsilon \downarrow 0$), the ratio $\tau(\pi^\star(L,\delta))/\log L$ satisfies an almost sure upper bound that matches with the right-hand side of (35). We then show that the family $\{\tau(\pi^\star(L,\delta))/\log L : L > 1\}$ is uniformly integrable. Combining the almost sure upper bound with the uniform integrability result yields (35). ■

VI. Main Result

We are now ready to state the main result of this paper.

Theorem 1. *Consider a multi-armed bandit with $K \geq 3$ arms in which each arm is a time homogeneous and ergodic Markov process on the finite state space $\mathcal{S}$. Fix $h \in \mathcal{A}$, and suppose that h is the odd arm. Let P_1 be the transition probability matrix of the Markov process of arm h. Further, for all $a \neq h$, let the transition probability matrix of arm a be P_2, where $P_2 \neq P_1$. Fix $\eta \in (0, 1]$, and suppose that a decision maker who wishes to identify the odd arm has a trembling hand with parameter η. Assuming that P_1 and P_2 are known to the decision maker, the expected time required by the decision maker to identify the odd arm satisfies the asymptotic relation*

$$\lim_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} = \lim_{\delta\downarrow 0} \lim_{L\to\infty} \frac{E_h[\tau(\pi^\star(L,\delta))]}{\log L} = \frac{1}{R^*(P_1, P_2)}. \tag{36}$$

Proof: From Lemma 3, we see that given any error tolerance parameter $\epsilon > 0$, by setting $L = 1/\epsilon$, we have $\pi^\star(L,\delta) \in \Pi(\epsilon)$ for all $\delta > 0$. Therefore, it follows that for all $\epsilon, \delta > 0$,

$$\inf_{\pi\in\Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log\left(\frac{1}{\epsilon}\right)} \leq \frac{E_h[\tau(\pi^\star(L,\delta))]}{\log L}. \tag{37}$$

Fixing $\delta > 0$ and letting $\epsilon \downarrow 0$ (which is identical to letting $L \to \infty$) in (37), and using the upper bound in (35), we get

$$\limsup_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} \leq \limsup_{L\to\infty} \frac{E_h[\tau(\pi^\star(L,\delta))]}{\log L} \leq \frac{1+\delta}{R^*(P_1, P_2)}. \tag{38}$$

Letting $\delta \downarrow 0$ in (38) and noting that the leftmost term in (38) does not depend on δ, we get

$$\limsup_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} \leq \lim_{\delta\downarrow 0} \limsup_{L\to\infty} \frac{E_h[\tau(\pi^\star(L,\delta))]}{\log L} \leq \frac{1}{R^*(P_1, P_2)}. \tag{39}$$

Combining the result in (39) with the lower bound in (12), we get

$$\frac{1}{R^*(P_1, P_2)} \leq \liminf_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} \leq \limsup_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} \leq \lim_{\delta\downarrow 0} \limsup_{L\to\infty} \frac{E_h[\tau(\pi^\star(L,\delta))]}{\log L} \leq \frac{1}{R^*(P_1, P_2)}. \tag{40}$$

Thus, it follows that the limit infimum and the limit suprema in (40) are indeed limits, thereby yielding (36). This completes the proof of the theorem. ■

We thus see that the policy $\pi^\star(L,\delta)$ is asymptotically optimal. As noted in Lemma 3, the parameter L may be set appropriately so as to ensure that the policy meets the desired error probability at stoppage. Furthermore, the parameter δ may be set so as to ensure that the upper bound in (35) is within a desired accuracy from the lower bound in (12). Finally, we emphasise here that our analysis of the lower and upper bounds crucially relies on the trembling hand parameter η being strictly positive.

VII. The Case $\eta = 0$

We now investigate the question of whether the results of this paper extend directly to the case when the decision maker does not suffer from a trembling hand, i.e., $\eta = 0$. While, in principle, we may consider plugging $\eta = 0$ in (12) and treating the resulting expression as the lower bound for the case when $\eta = 0$, it is not clear if this new lower bound can be approached asymptotically through a sequence of strategies (policies) in the sense of (35). It is worth noting here that the settings of (a) iid observations from the arms studied in [1]–[3], and (b) rested Markov arms studied in [4], are special cases of the setting of restless Markov arms. Indeed, the iid process of each arm may be treated as a Markov process, and in this case, the notions of arm delays and last observed states are redundant, as pointed out in Section I. The setting of rested Markov arms may be realised by fixing $d_a(t) \equiv 1$ for all $a \in \mathcal{A}$ and for all $t \geq K$, which is akin to saying that the unobserved arms do not undergo state evolution.

Thus, an affirmative answer to the question of whether the results of this paper extend to the case when $\eta = 0$ will imply that the results (lower bound and asymptotically optimal policies) of the prior works may be recovered as special cases of the answer to the above question (albeit some technicalities such as the countably infinite alphabet of the Poisson random variables in [1], [3] versus the finite state space of the Markov process of each arm as considered in this paper).

In what follows, we bring to light the following observations.

1) Writing $R^*(P_1, P_2)$ of (13) more explicitly as $R^*_\eta(P_1, P_2)$ for $\eta \in (0, 1]$, we show that $\lim\limits_{\eta \downarrow 0} R^*_\eta(P_1, P_2)$ exists. This is based on a key monotonicity property which we bring out in Section VII-A.
2) Writing $R^*_0(P_1, P_2)$ to denote the constant obtained by plugging $\eta = 0$ in (13), we demonstrate that
$$\lim_{\eta \downarrow 0} R^*_\eta(P_1, P_2) \leq R^*_0(P_1, P_2). \tag{41}$$
It is not clear if, in general, the inequality in (41) is an equality.
3) We show in Section VII-B and Section VII-C that the lower bounds for the setting when either (a) each arm yields iid observations from a common finite alphabet, or (b) each arm yields Markov observations from a common finite state space and the arms are rested, may be recovered from (13) by plugging $\eta = 0$ in (13). In other words, we show that for each of the above settings, the inequality in (41) is an equality, thus implying that the lower bounds for these settings may be approached asymptotically through a sequence of "trembling-hand" based policies similar to that presented in this paper; the policies of [1, Section II.B] and [4, Section IV] are example cases in point. A key ingredient that goes into the proofs of these results is the envelope theorem [18, Theorem 2].

A. A Key Monotonicity Property

Fix $\eta \in (0, 1]$, and assume that the decision maker possesses a trembling hand with parameter η. Let $\lambda = \lambda(\cdot \mid \cdot)$ be any conditional probability distribution on the arms, conditioned on the arm delays and the last observed states, as described in Section III, and let Λ denote the set of all such conditional distributions. Define
$$\Lambda^\eta \coloneqq \left\{ \frac{\eta}{K} + (1 - \eta)\, \lambda(\cdot \mid \cdot) : \lambda(\cdot \mid \cdot) \in \Lambda \right\}. \tag{42}$$
Note that for any $\lambda(\cdot \mid \cdot) \in \Lambda$, the corresponding element of Λ^η is the probability distribution according to which arms are *actually* selected, when the decision maker *intends* to pull the arms according to $\lambda(\cdot \mid \cdot)$. Notice that $\Lambda^\eta \subset \Lambda$ for all $\eta \in (0, 1]$.

The following Lemma shows that Λ^η is non-decreasing as η decreases.

Lemma 4. *$\Lambda^\eta \subset \Lambda^{\eta'}$ for all $0 < \eta' < \eta \leq 1$.*

Proof: Fix $0 < \eta' < \eta \leq 1$, and consider $\frac{\eta}{K} + (1-\eta)\,\lambda(\cdot \mid \cdot) \in \Lambda^{\eta}$ for some $\lambda(\cdot \mid \cdot) \in \Lambda$. Then, for all $(\underline{d}, \underline{i}, a) \in \mathbb{S} \times \mathcal{A}$,

$$\begin{aligned}
\frac{\eta}{K} + (1-\eta)\ \lambda(a|\underline{d}, \underline{i}) &= \frac{\eta'}{K} + \frac{\eta - \eta'}{K} + (1-\eta)\ \lambda(a|\underline{d}, \underline{i}) \\
&= \frac{\eta'}{K} + (1-\eta') \left[\frac{\eta - \eta'}{1-\eta'} \cdot \frac{1}{K} + \frac{1-\eta}{1-\eta'}\ \lambda(a|\underline{d}, \underline{i}) \right] \\
&= \frac{\eta'}{K} + (1-\eta') \left[\frac{\eta''}{K} + (1-\eta'')\ \lambda(a|\underline{d}, \underline{i}) \right] \qquad (43) \\
&\in \Lambda^{\eta'}, \qquad (44)
\end{aligned}$$

where in (43), $\eta'' = \frac{\eta - \eta'}{1-\eta'} \in (0, 1]$, and (44) follows by noting that the term inside the square brackets in (43) is a valid element of Λ. The relation in (44) implies that every element of Λ^{η} is also an element of $\Lambda^{\eta'}$ whenever $\eta' < \eta$. This completes the proof. ∎

Plugging $\eta = 0$ in (42), and denoting the resulting set as Λ^0, we see that $\Lambda^0 = \Lambda$. Thus, it follows from Lemma 4 that

$$\bigcup_{\eta \downarrow 0} \Lambda^{\eta} \subset \Lambda. \qquad (45)$$

Let us now turn our attention to (15), and note that the right-hand side of (15) represents the long-term probability of seeing the state $(\underline{d}, \underline{i})$ and selecting arm a subsequently with probability $\frac{\eta}{K} + (1-\eta)\,\lambda(a|\underline{d}, \underline{i})$. Defining $\lambda^{\eta}(\cdot \mid \cdot) \coloneqq \frac{\eta}{K} + (1-\eta)\,\lambda(\cdot \mid \cdot)$, and writing ν^{λ} in (15) as $\nu^{\lambda^{\eta}}$, we may express the right-hand side of (13) equivalently as

$$R^*_{\eta}(P_1, P_2) \coloneqq \sup_{\lambda^{\eta}(\cdot|\cdot) \in \Lambda^{\eta}}\ \min_{h' \neq h}\ \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}}\ \sum_{a=1}^{K} \nu^{\lambda^{\eta}}(\underline{d}, \underline{i}, a)\, k(\underline{d}, \underline{i}, a). \qquad (46)$$

It follows from Lemma 4 that $R^*_{\eta}(P_1, P_2)$ is non-decreasing in η; thus, $\lim_{\eta \downarrow 0} R^*_{\eta}(P_1, P_2)$ exists.

Finally, denoting by $R^*_0(P_1, P_2)$ the quantity obtained by plugging $\eta = 0$ in (46), it follows from (45) that (41) holds.

B. IID Observations From The Arms

We now show that when each arm yields iid observations coming from a finite alphabet common across the arms, the inequality in (41) is indeed an equality. Fix $h \in \mathcal{A}$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Let arm h be associated with an iid process whose underlying law is ν_1. Further, for all $h' \neq h$, let arm h' be associated with an iid process whose law is ν_2, where $\nu_2 \neq \nu_1$. Assume that the iid process of any given arm is independent of the iid process of each of the remaining arms. Let ν_h^a denote the marginal law of the iid process of arm a under the hypothesis $\mathcal{H}_h$, i.e.,

$$\nu_h^a = \begin{cases} \nu_1, & a = h, \\ \nu_2, & a \neq h. \end{cases} \qquad (47)$$

Since any iid process is trivially a Markov process, with the state space of the Markov process being the alphabet of the iid process, we may let P_1 denote the transition probability matrix of arm h and P_2 the transition probability matrix of each of the non-odd arms $h' \neq h$. Then, for all $i, j \in \mathcal{S}$ and $d \geq 1$, we have

$$P_1^d(j|i) = \nu_1(j), \quad P_2^d(j|i) = \nu_2(j). \qquad (48)$$

Thus, when each arm yields iid observations, the function $k(\underline{d}, \underline{i}, a)$ in (14) may be expressed as

$$k(\underline{d}, \underline{i}, a) = \begin{cases} D(\nu_1 \| \nu_2), & a = h, \\ D(\nu_2 \| \nu_1), & a = h', \\ 0, & \text{otherwise.} \end{cases} \qquad (49)$$

In other words, the function k does not depend on either the arm delays or the last observed states. Noting that the right-hand side of (49) may be written compactly as $D(\nu_h^a \| \nu_{h'}^a)$, and plugging this in (46), we get

$$\begin{aligned}
R_\eta^*(P_1, P_2) &= \sup_{\lambda^\eta(\cdot|\cdot)\in\Lambda^\eta} \min_{h'\neq h} \sum_{a=1}^{K} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \nu^{\lambda^\eta}(\underline{d},\underline{i},a)\, D(\nu_h^a \| \nu_{h'}^a) \\
&\overset{(a)}{=} \sup_{\lambda^\eta(\cdot|\cdot)\in\Lambda^\eta} \min_{h'\neq h} \sum_{a=1}^{K} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \mu^{\lambda^\eta}(\underline{d},\underline{i})\, \lambda^\eta(a|\underline{d},\underline{i})\, D(\nu_h^a \| \nu_{h'}^a) \\
&= \sup_{\lambda(\cdot|\cdot)\in\Lambda} \min_{h'\neq h} \sum_{a=1}^{K} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \mu^{\lambda^\eta}(\underline{d},\underline{i}) \left[\frac{\eta}{K} + (1-\eta)\,\lambda(a|\underline{d},\underline{i})\right] D(\nu_h^a \| \nu_{h'}^a) \\
&\overset{(b)}{=} \sup_{\lambda(\cdot|\cdot)\in\Lambda} \min_{h'\neq h} \frac{\eta}{K} \sum_{a=1}^{K} D(\nu_h^a \| \nu_{h'}^a) + (1-\eta) \sum_{a=1}^{K} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \mu^{\lambda^\eta}(\underline{d},\underline{i})\, \lambda(a|\underline{d},\underline{i})\, D(\nu_h^a \| \nu_{h'}^a) \\
&= \sup_{\lambda\in\mathcal{P}(\mathcal{A})} \min_{h'\neq h} \frac{\eta}{K} \sum_{a=1}^{K} D(\nu_h^a \| \nu_{h'}^a) + (1-\eta) \sum_{a=1}^{K} \lambda(a)\, D(\nu_h^a \| \nu_{h'}^a) \qquad (50) \\
&= \sup_{\lambda\in\mathcal{P}(\mathcal{A})} \frac{\eta}{K} \left[D(\nu_1\|\nu_2) + D(\nu_2\|\nu_1)\right] + (1-\eta) \left[\lambda(h)\, D(\nu_1\|\nu_2) + \left(\min_{h'\neq h}\ \lambda(h')\right) D(\nu_2\|\nu_1)\right], \qquad (51)
\end{aligned}$$

where in (a) above, μ^{λ^η} is the long-term probability of observing the state $(\underline{d},\underline{i})$ when the arms are selected according to the distribution $\lambda^\eta(\cdot \mid \cdot)$, (b) above follows by using the fact that ν^{λ^η} is a probability distribution on $\mathbb{S}\times\mathcal{A}$, and the term $\lambda(a)$ in (50) is given by

$$\lambda(a) = \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \mu^{\lambda^\eta}(\underline{d},\underline{i})\, \lambda(a|\underline{d},\underline{i}), \quad a\in\mathcal{A},$$

with $\mathcal{P}(\mathcal{A})$ in (50) denoting the set of all probability distributions on the set $\mathcal{A}$. Lastly, (51) follows by noting that

$$\nu_h^a = \begin{cases} \nu_1, & a = h, \\ \nu_2, & a \neq h, \end{cases} \qquad\qquad \nu_{h'}^a = \begin{cases} \nu_1, & a = h', \\ \nu_2, & a \neq h', \end{cases} \qquad (52)$$

and therefore the only non-zero terms in the summation over the arms in (50) are those corresponding to $a = h$ and $a = h'$.

We now note that for each $\lambda\in\mathcal{P}(\mathcal{A})$, the mapping

$$\eta \longmapsto \frac{\eta}{K}\left[D(\nu_1\|\nu_2) + D(\nu_2\|\nu_1)\right] + (1-\eta)\left[\lambda(h)\, D(\nu_1\|\nu_2) + \left(\min_{h'\neq h}\ \lambda(h')\right) D(\nu_2\|\nu_1)\right]$$

is bounded and linear (hence absolutely continuous) for all $\eta\in[0,1]$. Using the envelope theorem [18, Theorem 2], we get that the mapping $\eta\mapsto R_\eta^*(P_1,P_2)$ is absolutely continuous for all $\eta\in[0,1]$, thereby implying that $\lim_{\eta\downarrow 0} R_\eta^*(P_1,P_2) = R_0^*(P_1,P_2)$. This establishes that the inequality in (41) holds with equality.

C. Rested Markov Arms

We now show that when each arm is a Markov process on a finite state space that is common across the arms, and the arms are rested, the inequality in (41) is indeed an equality. Fix $h\in\mathcal{A}$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Let each arm be associated with a time-homogeneous and ergodic discrete-time Markov process on a common, finite state space $\mathcal{S}$. Let P_1 be the transition probability matrix of the odd arm, and let P_2 be the transition probability matrix of each of the non-odd arms. Let μ_1 and μ_2 denote the unique stationary distributions of P_1 and P_2 respectively. Assume that the Markov process of any given arm is independent of the Markov process of each of the remaining arms.

Let P_h^a denote the transition probability matrix of arm a under the hypothesis $\mathcal{H}_h$, and let μ_h^a be the stationary distribution of P_h^a. It then follows that

$$P_h^a = \begin{cases} P_1, & a = h, \\ P_2, & a \neq h, \end{cases} \qquad \mu_h^a = \begin{cases} \mu_1, & a = h, \\ \mu_2, & a \neq h. \end{cases} \tag{53}$$

When the arms are rested, as noted at the beginning of this section, the delay parameter for every arm is identically equal to 1, i.e., $d_a(t) \equiv 1$ for all $a \in \mathcal{A}$ and $t \geq K$. Thus, we may omit the summation over $\underline{d}$ in (46). Writing $\lambda(a|\underline{i})$ in place of $\lambda(a|\underline{d}, \underline{i})$, writing $\nu^{\lambda^\eta}(\underline{i})$ in place of $\nu^{\lambda^\eta}(\underline{d}, \underline{i})$, and the following the steps presented earlier for the case of iid observations, we have

$$\begin{aligned}
R_\eta^*(P_1, P_2) &= \sup_{\lambda^\eta(\cdot|\cdot)\in\Lambda^\eta} \min_{h'\neq h} \sum_{a=1}^{K} \sum_{\underline{i}\in\mathcal{S}^K} \nu^{\lambda^\eta}(\underline{i}, a)\, D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a)) \\
&= \sup_{\lambda(\cdot|\cdot)\in\Lambda} \min_{h'\neq h} \sum_{a=1}^{K} \sum_{\underline{i}\in\mathcal{S}^K} \mu^{\lambda^\eta}(\underline{i}) \left[\frac{\eta}{K} + (1-\eta)\,\lambda(a|\underline{i})\right] D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a)) \\
&= \sup_{\lambda(\cdot|\cdot)\in\Lambda} \min_{h'\neq h} \Bigg[\frac{\eta}{K} \sum_{a=1}^{K} \sum_{\underline{i}\in\mathcal{S}^K} \mu^{\lambda^\eta}(\underline{i})\, D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a)) \\
&\qquad\qquad + (1-\eta) \sum_{a=1}^{K} \sum_{\underline{i}\in\mathcal{S}^K} \mu^{\lambda^\eta}(\underline{i})\,\lambda(a|\underline{i})\, D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a))\Bigg] \\
&\overset{(a)}{=} \sup_{\lambda(\cdot|\cdot)\in\Lambda} \min_{h'\neq h} \Bigg[\frac{\eta}{K} \sum_{a=1}^{K} \sum_{\underline{i}\in\mathcal{S}^K} \mu^{\lambda^\eta}(\underline{i})\, D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a)) \\
&\qquad\qquad + (1-\eta) \sum_{a=1}^{K} \sum_{i_a\in\mathcal{S}} \left(\sum_{\underline{i}^{-a}\in\mathcal{S}^{K-1}} \mu^{\lambda^\eta}(\underline{i})\,\lambda(a|\underline{i})\right) D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a))\Bigg] \\
&\overset{(b)}{=} \sup_{\lambda(\cdot|\cdot)\in\Lambda} \min_{h'\neq h} \Bigg[\frac{\eta}{K} \sum_{a=1}^{K} \sum_{i_a\in\mathcal{S}} \mu^{\lambda^\eta}(i_a)\, D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a)) \\
&\qquad\qquad + (1-\eta) \sum_{a=1}^{K} \sum_{i_a\in\mathcal{S}} \mu^{\lambda^\eta}(i_a)\,\lambda(a|i_a)\, D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a))\Bigg],
\end{aligned} \tag{54}$$

where in (a) above, $\underline{i}^{-a}$ denotes the vector of last observed states excluding the component corresponding to arm a, and in (b) above, $\mu^{\lambda^\eta}(i_a)$ denotes the marginal of $\mu^{\lambda^\eta}(\underline{i})$ corresponding to arm a. Further, in writing (b), we use the simplification

$$\sum_{\underline{i}^{-a}\in\mathcal{S}^{K-1}} \mu^{\lambda^\eta}(\underline{i})\,\lambda(a|\underline{i}) = \mu^{\lambda^\eta}(i_a)\,\lambda(a|i_a). \tag{55}$$

We now note that the product $\mu^{\lambda^\eta}(i_a)\,\lambda(a|i_a)$ represents the long-term probability of observing arm a in state i_a and subsequently selecting arm a according to the conditional distribution $\lambda(a|i_a)$. This may be interpreted as long-term probability of first seeing a transition *from* the state i_a on arm a and subsequently selecting arm a based on the observed transition. Since the arms are rested, long-term probability of seeing a transition *from* the state i_a on arm a is equal to the long-term probability of seeing a transition *to* the state i_a on arm a. Due to the ergodic nature of each of the arms, these probabilities are in turn equal to the probability of observing the state i_a on arm a under its stationary distribution (see [4] for a discussion on this).

Hence, under the hypothesis $\mathcal{H}_h$, we may write

$$\mu^{\lambda^\eta}(i_a)\,\lambda(a|i_a) = \lambda(a)\cdot\mu_h^a(i_a), \tag{56}$$

where in (56), $\lambda(a) = \sum_{i_a \in \mathcal{S}} \mu^{\lambda^\eta}(i_a)\, \lambda(a|i_a)$. Using (56) in (54), we have

$$\begin{aligned} R^*_\eta(P_1, P_2) = \sup_{\lambda(\cdot|\cdot)\in\Lambda} \min_{h'\neq h} \Bigg[& \frac{\eta}{K} \sum_{a=1}^{K} \sum_{i_a \in \mathcal{S}} \mu^{\lambda^\eta}(i_a)\, D(P^a_h(\cdot|i_a) \| P^a_{h'}(\cdot|i_a)) \\ & + (1-\eta) \sum_{a=1}^{K} \sum_{i_a \in \mathcal{S}} \lambda(a)\, \mu^a_h(i_a)\, D(P^a_h(\cdot|i_a) \| P^a_{h'}(\cdot|i_a)) \Bigg] \\ \overset{(a)}{=} \sup_{\lambda\in\mathcal{P}(\mathcal{A})} \min_{h'\neq h} & \left[\frac{\eta}{K} \sum_{a=1}^{K} D(P^a_h(\cdot|\cdot) \| P^a_{h'}(\cdot|\cdot) | \mu^a_h) + (1-\eta) \sum_{a=1}^{K} \lambda(a)\, D(P^a_h(\cdot|\cdot) \| P^a_{h'}(\cdot|\cdot) | \mu^a_h) \right] \\ = \sup_{\lambda\in\mathcal{P}(\mathcal{A})} & \left[\frac{\eta}{K} \Big(D(P_1(\cdot|\cdot) \| P_2(\cdot|\cdot) \| \mu_1) + D(P_2(\cdot|\cdot) \| P_1(\cdot|\cdot) \| \mu_2) \Big) \right. \\ & \left. + (1-\eta) \Big(\lambda(h)\, D(P_1(\cdot|\cdot) \| P_2(\cdot|\cdot) \| \mu_1) + \Big(\min_{h'\neq h} \lambda(h') \Big) D(P_2(\cdot|\cdot) \| P_1(\cdot|\cdot) \| \mu_2) \Big) \right] \end{aligned} \tag{57}$$

where in (a) above,

$$D(P^a_h(\cdot|\cdot) \| P^a_{h'}(\cdot|\cdot) | \mu^a_h) \coloneqq \sum_{i_a \in \mathcal{S}} \mu^a_h(i_a)\, D(P^a_h(\cdot|i_a) \| P^a_{h'}(\cdot|i_a)),$$

and (57) follows by noting that

$$P^a_h = \begin{cases} P_1, & a = h, \\ P_2, & a \neq h, \end{cases} \qquad \mu^a_h = \begin{cases} \mu_1, & a = h, \\ \mu_2, & a \neq h, \end{cases} \qquad P^a_{h'} = \begin{cases} P_1, & a = h', \\ P_2, & a \neq h', \end{cases} \qquad \mu^a_{h'} = \begin{cases} \mu_1, & a = h', \\ \mu_2, & a \neq h', \end{cases} \tag{58}$$

hence, the only non-zero terms in the summation over the arms in (a) above are those corresponding to $a = h$ and $a = h'$.

Finally, we note that for each $\lambda \in \mathcal{P}(\mathcal{A})$, the mapping

$$\begin{aligned} \eta \longmapsto & \frac{\eta}{K} \Big(D(P_1(\cdot|\cdot) \| P_2(\cdot|\cdot) \| \mu_1) + D(P_2(\cdot|\cdot) \| P_1(\cdot|\cdot) \| \mu_2) \Big) \\ & + (1-\eta) \Big(\lambda(h)\, D(P_1(\cdot|\cdot) \| P_2(\cdot|\cdot) \| \mu_1) + \Big(\min_{h'\neq h} \lambda(h') \Big) D(P_2(\cdot|\cdot) \| P_1(\cdot|\cdot) \| \mu_2) \Big) \end{aligned}$$

is bounded and linear (hence absolutely continuous) for all $\eta \in [0,1]$. Using the envelope theorem [18, Theorem 2], we get that the mapping $\eta \mapsto R^*_\eta(P_1, P_2)$ is absolutely continuous for all $\eta \in [0,1]$, thereby implying that $\lim_{\eta\downarrow 0} R^*_\eta(P_1, P_2) = R^*_0(P_1, P_2)$. This establishes that the inequality in (41) holds with equality.

VIII. Concluding Remarks

We make several concluding remarks to end the paper.

1) From (36), when the trembling hand parameter $\eta > 0$, we see that

$$\lim_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} = \frac{1}{R^*_\eta(P_1, P_2)}. \tag{59}$$

We have thus provided an answer to (8) on the minimum growth rate of the expected time to identify the odd arm location as $\epsilon \downarrow 0$.

2) The asymptotically optimal $\lambda(\cdot|\cdot)$ in the restless case may depend on history unlike that in the prior works [1]–[4] where $\lambda(\cdot)$ did not depend on history, even in the rested Markov case. At first glance, this is surprising for the rested Markov case, but in retrospect, these features are apparent from an examination of the optimisation problem (13) in these special cases.

3) Computability of $R^*_\eta(P_1, P_2)$ may be an issue, and one must usually resort to Q-learning for restless Markov arms [17] to arrive at good policies. The fact that $D(P^{d_a}_k(\cdot|i_a) || P^{d_a}_l(\cdot|i_a))$, $k, l \in \{1, 2\}$, converges as $d_a \to \infty$ could enable restriction of the countable state space $\mathbb{S}$ to a finite set, and could lead to good approximations.

4) When the trembling hand parameter $\eta > 0$, the ergodicity of the Markov process $(\underline{d}(t), \underline{i}(t))$ under any SRS policy ensures that time averages approach the ensemble averages. This is crucial to show achievability. Note also the use of uniqueness of the stationary distribution to show the converse. The trembling hand model may be viewed as a *regularisation* that gives stability of the aforementioned Markov process for free. If the trembling hand parameter η were 0, one could deliberately add some regularisation parameterised by η, and let this parameter $\eta \downarrow 0$. $R_0^*(P_1, P_2)$ governs the lower bound, whereas $\lim_{\eta \downarrow 0} R_\eta^*(P_1, P_2)$ governs the upper bound. The resulting lower and upper bounds on the growth rate may have a gap.
5) Open questions: The key difficulties when $\eta = 0$ are (a) absence of ergodicity property, and (b) a formal verification of the envelope theorem. It would be interesting to study these. Another open question is the setting when P_1 and P_2 are unknown and have to be learnt along the way.

APPENDIX A
PROOF OF LEMMA 1

Let $\eta \in (0, 1]$ be the parameter of the trembling hand. Fix $\pi^\lambda \in \Pi_{\mathsf{SRS}}$ and $h \in \mathcal{A}$, and let $\mathcal{H}_h$ be the true hypothesis. Recall that the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is indeed a Markov process under the SRS policy π^λ.

Proof of Irreducibility: Consider any two states $(\underline{d}, \underline{i}) \in \mathbb{S}$ and $(\underline{d}', \underline{i}') \in \mathbb{S}$, and suppose that the Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is in the state $(\underline{d}, \underline{i})$ at some time $t = T_0$. We shall now demonstrate that there exists N such that the state $(\underline{d}', \underline{i}')$ may be reached starting from the state $(\underline{d}, \underline{i})$ in N steps under the SRS policy π^λ. Recall that at any time t, the arm that is intended to be pulled under the policy π^λ is B_t, while the actual arm that is pulled at time t is its trembled version A_t; the arms A_t and B_t are related through the trembling hand relation in (4). In particular, for any $a \in \mathcal{A}$, we have

$$\begin{aligned}
P(A_t = a \mid B^{t-1}, A^{t-1}, \bar{X}^{t-1}) &= \sum_{b=1}^{K} P(B_t = b,\, A_t = a | B^{t-1}, A^{t-1}, \bar{X}^{t-1}) \\
&= \sum_{b=1}^{K} P(B_t = b \mid B^{t-1}, A^{t-1}, \bar{X}^{t-1}) \cdot P(A_t = a \mid B_t = b, B^{t-1}, A^{t-1}, \bar{X}^{t-1}) \\
&\overset{(a)}{=} \sum_{b=1}^{K} \lambda(b \mid \underline{d}(t), \underline{i}(t)) \cdot \Big(\frac{\eta}{K} + (1-\eta)\, \mathbb{I}_{\{a=b\}} \Big) \\
&= \frac{\eta}{K} + (1-\eta)\, \lambda(a \mid \underline{d}(t), \underline{i}(t)) \\
&\geq \frac{\eta}{K}, \qquad (60)
\end{aligned}$$

where (a) above follows from (4) and the fact that under the policy $\pi^\lambda \in \Pi_{\mathsf{SRS}}$, the intended arm B_t is selected based on the history $(B^{t-1}, A^{t-1}, \bar{X}^{t-1})$ according to the distribution $\lambda(\cdot \mid \cdot)$.

Assume without loss of generality that the vector $\underline{d}'$ of arm delays is such that $d_1' > d_2' > \cdots > d_K = 1$. From [20, Proposition 1.7] for finite state Markov processes, we get that there exists an integer M such that for all $m \geq M$,

$$P_1^m(j|i) > 0 \text{ for all } i, j \in \mathcal{S}, \quad P_2^m(j|i) > 0 \text{ for all } i, j \in \mathcal{S}. \qquad (61)$$

Consider the sequence of actions and observations as follows: starting from the state $(\underline{d}, \underline{i})$ at time $t = T_0$, let the Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ evolve for $M - 1$ time instants. Thereafter, let arm 1 be selected at the $(T_0 + M)$th time instant and let the state observed on arm 1 be i_1'; let arm 2 be selected at the $(T_0 + M + d_1' - d_2')$th time instant and let the state observed on arm 2 be i_2', and so on. Finally, let arm K be observed at the $(T_0 + M + d_1' - d_K')$th time instant, and let the

state observed on arm K be i'_K. Additionally, let arm 1 not be selected for all $T_0+M<t<T_0+M+d'_1$; let arm 2 not be selected for all $T_0+M+d'_1-d'_2<t<T_0+M+d'_1$ and so on.

Clearly, the above sequence of actions and observations leads to the state $(\underline{d}',\underline{i}')$ after $M+d'_1-d'_K$ time instants. Thus, the probability of starting from the state $(\underline{d},\underline{i})$ and reaching the state $(\underline{d}',\underline{i}')$ may be lower bounded by the probability that the above sequence of actions and observations occur under the policy π^λ which, under the hypothesis $\mathcal{H}_h$, is given by

$$\begin{aligned}
&\left(\prod_{a=1}^{K-1} P(A_{T_0+M+d'_1-d'_a}=a \mid B^{T_0+M+d'_1-d'_a-1}, A^{T_0+M+d'_1-d'_a-1}, \bar{X}^{T_0+M+d'_1-d'_a-1})\right) \cdot \left(\prod_{a=1}^{K}(P_h^a)^{M+d'_1-d'_a}(i'_a|i_a)\right) \\
&\qquad \cdot \left(\prod_{a=1}^{K-1}\prod_{t=T_0+M+d'_1-d'_a+1}^{T_0+M+d'_1-d'_{a+1}} P(A_t \notin \{1,\ldots,a\} \mid B^{t-1}, A^{t-1}, \bar{X}^{t-1})\right) \\
&\overset{(a)}{\geq} \left(\frac{\eta}{K}\right)^{K-1} \cdot \left[\prod_{a=1}^{K}(P_h^a)^{M+d'_1-d'_a}(i'_a|i_a)\right] \cdot \left[\prod_{a=1}^{K-1}\prod_{t=T_0+M+d'_1-d'_a+1}^{T_0+M+d'_1-d'_{a+1}} \frac{\eta(K-a)}{K}\right] \\
&\overset{(b)}{\geq} \left(\frac{\eta}{K}\right)^{K-1} \cdot \left[\prod_{a=1}^{K}(P_h^a)^{M+d'_1-d'_a}(i'_a|i_a)\right] \cdot \left[\prod_{a=1}^{K-1}\prod_{t=T_0+M+d'_1-d'_a+1}^{T_0+M+d'_1-d'_{a+1}} \frac{\eta}{K}\right] \\
&> 0, \qquad (62)
\end{aligned}$$

where (a) above follows from the observation that the right-hand side of (60), for each t, is $\geq \eta/K$ and the fact that

$$P(A_t \notin \{1,\ldots,a\} \mid B^{t-1}, A^{t-1}\bar{X}^{t-1}) = \sum_{a'=a+1}^{K} P(A_t=a' \mid B^{t-1}, A^{t-1}, \bar{X}^{t-1}) \geq \frac{\eta(K-a)}{K},$$

and (b) follows by noting that $K-a\geq 1$ for $a\in\{1,\ldots,K-1\}$. Setting $N=M+d'_1-d'_K$, we see that the Markov process $\{(\underline{d}(t),\underline{i}(t):t\geq K)\}$ is in the state $(\underline{d}',\underline{i}')$ at time $t=T_0+N$. This establishes irreducibility. ■

Proof of Aperiodicity: Fix an arbitrary $(\underline{d},\underline{i})\in\mathbb{S}$. We shall now demonstrate that starting from the state $(\underline{d},\underline{i})$, there is a strictly positive probability for the Markov process $\{(\underline{d}(t),\underline{i}(t)):t\geq K\}$ to return back to the state $(\underline{d},\underline{i})$ in M' steps as well as in $(M'+1)$ steps, where M' is sufficiently large and such that (61) holds for all $m\geq M'$. This will then establish the desired aperiodicity property since the period of the state $(\underline{d},\underline{i})$ is equal to the gcd of M' and $M'+1$, which is 1.

Assume, without loss of generality, that $\underline{d}$ is such that $d_1>d_2>\cdots d_K=1$. Let M be such that (61) holds for all $m\geq M$. Using arguments similar to that in the proof of irreducibility presented above, the probability of starting from the state $(\underline{d},\underline{i})$ at some time $t=T_0$ and returning back to the state $(\underline{d},\underline{i})$ after $M+d_1$ time instants may be lower bounded, under hypothesis $\mathcal{H}_h$, by

$$\left(\frac{\eta}{K}\right)^{K-1} \cdot \left[\prod_{a=1}^{K}(P_h^a)^{M+d'_1-d'_a}(i'_a|i_a)\right] \cdot \left[\prod_{a=1}^{K-1}\prod_{t=T_0+M+d_1-d_a+1}^{T_0+M+d_1-d_{a+1}} \frac{\eta}{K}\right] > 0. \qquad (63)$$

Setting $M'=M+d_1-d_K$ yields the desired result. ■

Proof of positive recurrence: Let

$$p_\eta \coloneqq \frac{\eta}{K}\min\left\{\min\{P_1^M(j|i):i,j\in\mathcal{S}\}, \min\{P_2^M(j|i):i,j\in\mathcal{S}\}\right\}; \qquad (64)$$

here, once again, M is such that (61) holds for all $m\geq M$. Therefore, it follows that $p_\eta>0$. Let

$$r(\pi^\lambda) \coloneqq \min\{t>K : \underline{d}(t)=\underline{d}(K), \underline{i}(t)=\underline{i}(K)\} \qquad (65)$$

denote the first return time of the Markov process $\{(\underline{d}(t),\underline{i}(t)):t\geq K\}$ to its initial state (i.e., the state at time $t=K$) under the SRS policy π^λ. Note that

$$r(\pi^\lambda) \leq M\cdot K\cdot \tau_\eta \quad \text{almost surely}, \qquad (66)$$

where τ_η is a Geometric random variable with parameter p_η. In other words, $r(\pi^\lambda)$ may be almost surely upper bounded by the first return time of $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ to its initial state measured only at time instants that are integer multiples of $M \cdot K$. It then follows that

$$\begin{aligned} E\left[r(\pi^\lambda)\right] &\leq M \cdot K \cdot E[\tau_\eta] \\ &= M \cdot K \cdot \frac{1}{p_\eta} \\ &< \infty, \end{aligned} \tag{67}$$

thus implying that the Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is positive recurrent under π^λ. This completes the proof of positive recurrence, as also the proof of the Lemma. ■

APPENDIX B
PROOF OF PROPOSITION 1

This proof is organised as follows. Given $\epsilon > 0$, we first obtain in Lemma 5 a lower bound for $E_h[Z_{hh'}(\tau(\pi)]$ for all $\pi \in \Pi(\epsilon)$ using a change of measure argument of Kaufmann et al. [14]. Following this, we obtain an upper bound for $E_h[Z_{hh'}(\tau(\pi)]$ in terms of $E_h[\tau(\pi)]$. Combining the upper and lower bounds, and letting $\epsilon \downarrow 0$, we arrive at the desired result. The ergodicity property established in Lemma 1 for SRS policies plays a crucial role in deriving the final lower bound of (12).

A. A Lower Bound on $E_h[Z_{hh'}(\tau(\pi))]$ for $\pi \in \Pi(\epsilon)$

As a first step towards deriving the lower bound, we use a result of Kaufmann et al. [14] to obtain a lower bound for $E_h[Z_{hh'}(\tau(\pi))]$ in terms of the error probability parameter ϵ. However, this requires a generalisation of [14, Lemma 18], a change of measure argument for iid observations from the arms, to the setting of restless arms with Markov observations. We present this generalisation in the following Lemma.

Lemma 5. *Fix $\pi \in \Pi(\epsilon)$, and let $\tau(\pi)$ be the stopping time of policy π. Let $\mathcal{F}_{\tau(\pi)}$ be the σ-algebra*

$$\mathcal{F}_{\tau(\pi)} = \{E \in \mathcal{F} : E \cap \{\tau(\pi) = t\} \in \mathcal{F}_t \text{ for all } t \geq 0\}, \tag{68}$$

where $\mathcal{F}_0 = \sigma(\Omega, \emptyset)$ and $\mathcal{F}_t = \sigma(B^t, A^t, \bar{X}^t)$ for all $t \geq 1$. Then, for any $h, h' \in \mathcal{A}$ such that $h' \neq h$, the relation

$$P_{h'}(E) = E_h[1_E \, \exp(-Z_{hh'}(\tau(\pi))] \tag{69}$$

holds for all $E \in \mathcal{F}_{\tau(\pi)}$.

Proof of Lemma 5: We prove the Lemma by demonstrating, through mathematical induction, that the relation

$$E_{h'}[g(B^t, A^t, \bar{X}^t)] = E_h[g(B^t, A^t, \bar{X}^t) \, \exp(-Z_{hh'}(t))] \tag{70}$$

holds for all $t \geq 0$ and for all measurable functions $g : \mathcal{A}^{t+1} \times \mathcal{A}^{t+1} \times \mathcal{S}^{t+1} \to \mathbb{R}$. The proof for the case $t = 0$ may be obtained as follows. For any measurable $g : \mathcal{A} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, we have

$$\begin{aligned} E_{h'}[g(B_0, A_0, \bar{X}_0)] &= \sum_{b=1}^{K} \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} g(b, a, i) \, P_{h'}(B_0 = b, A_0 = a, \bar{X}_0 = i) \\ &= \sum_{b=1}^{K} \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} g(b, a, i) \, P_{h'}(B_0 = b) \, P_{h'}(A_0 = a | B_0 = b) \, P_{h'}(\bar{X}_0 = i | B_0 = b, A_0 = a) \end{aligned}$$

$$\stackrel{(a)}{=} \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}} g(b,a,i)\, P_h(B_0=b)\, P_h(A_0=a|B_0=b)\, P_{h'}(\bar{X}_0=i|A_0=a)$$
$$= \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}} g(b,a,i)\, P_h(B_0=b)\, P_h(A_0=a|B_0=b)\, P_{h'}(X_0^a=i), \tag{71}$$

where (a) follows using the facts that $P_h(B_0=a)=P_{h'}(B_0=b)$ and $P_h(A_0=a|B_0=b)=P_{h'}(A_0=a|B_0=b)$ (see Section V). Assuming that $X_0^a\sim\nu$, where ν is a probability distribution on $\mathcal{S}$, independent of the true hypothesis (which is not known to the policy π), we have

$$E_h[g(A_0,\bar{X}_0)|\mathcal{H}_{h'}] = \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}} g(b,a,i)\, P_h(B_0=b)\, P_h(A_0=a|B_0=b)\,\nu(i) \tag{72}$$
$$= \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}} g(b,a,i)\, P_h(B_0=b)\, P_h(A_0=a|B_0=b)\, P_h(X_0^a=i|A_0=a). \tag{73}$$

Also, we have (see Section V)

$$Z_{hh'}(0) = \log\frac{P_h(B_0,A_0,\bar{X}_0)}{P_{h'}(B_0,A_0,\bar{X}_0)} = 0. \tag{74}$$

Combining (73) and (74), we get $E_{h'}[g(B_0,A_0,\bar{X}_0)] = E_h[g(B_0,A_0,\bar{X}_0)\exp(-Z_{hh'}(0))]$, thus proving (70) for $t=0$.

We now assume that (70) is true for some $t>0$, and demonstrate that it also true for $t+1$. By law of iterated expectations,

$$E_{h'}[g(B^{t+1},A^{t+1},\bar{X}^{t+1})] = E_{h'}[E_{h'}[g(B^{t+1},A^{t+1},\bar{X}^{t+1})|\mathcal{F}_t]]. \tag{75}$$

Noting that $E_{h'}[g(B^{t+1},A^{t+1},\bar{X}^{t+1})|\mathcal{F}_t]$ is a measurable function of $(B^t,A^t,\bar{X}^t)$, by the induction hypothesis, we have

$$E_{h'}[g(B^{t+1},A^{t+1},\bar{X}^{t+1})|\mathcal{F}_t] = E_h[E_{h'}[g(B^{t+1},A^{t+1},\bar{X}^{t+1})|\mathcal{F}_t]\,\exp(-Z_{hh'}(t))]. \tag{76}$$

We now note that

$$\begin{aligned}
&E_{h'}[g(B^{t+1},A^{t+1},\bar{X}^{t+1})|\mathcal{F}_t]\,\exp(-Z_{hh'}(t))\\
&\stackrel{(a)}{=} E_{h'}[g(B^{t+1},A^{t+1},\bar{X}^{t+1})\,\exp(-Z_{hh'}(t))|\mathcal{F}_t]\\
&= \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}}\Big[g(B^t,A^t,\bar{X}^t,b,a,i)\cdot P_{h'}(B_{t+1}=b|B^t,A^t,\bar{X}^t)\\
&\qquad\cdot P_{h'}(A_{t+1}=a|B^{t+1}=b,B^t,A^t,\bar{X}^t)\cdot P_{h'}(\bar{X}_{t+1}=i|B^{t+1}=b,A_{t+1}=a,B^t,A^t,\bar{X}^t)\cdot\exp(-Z_{hh'}(t))\Big]\\
&\stackrel{(b)}{=} \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}}\Big[g(B^t,A^t,\bar{X}^t,b,a,i)\cdot P_h(B_{t+1}=b|B^t,A^t,\bar{X}^t)\\
&\qquad\cdot P_h(A_{t+1}=a|B^{t+1}=b,B^t,A^t,\bar{X}^t)\cdot P_{h'}(\bar{X}_{t+1}=i|A_{t+1}=a,A^t,\bar{X}^t)\cdot\exp(-Z_{hh'}(t))\Big],
\end{aligned} \tag{77}$$

where (a) above is due to the fact that $Z_{hh'}(t)$ is a measurable function of $(B^t,A^t,\bar{X}^t)$, and in writing (b), we use the following facts: for any t,

- $P_{h'}(B_{t+1}=b|B^t,A^t,\bar{X}^t)=P_h(B_{t+1}=b|B^t,A^t,\bar{X}^t)$,
- $P_{h'}(A_{t+1}=a|B_{t+1}=b,B^t,A^t,\bar{X}^t)=P_h(A_{t+1}=a|B_{t+1}=b,B^t,A^t,\bar{X}^t)$, and
- $P_{h'}(\bar{X}_{t+1}=i|B_{t+1}=b,A_{t+1}=a,B^t,A^t,\bar{X}^t)=P_{h'}(\bar{X}_{t+1}=i|A_{t+1}=a,A^t,\bar{X}^t)$.

See Section V for a justification of why the above facts are true. It then follows that

$$\sum_{i\in\mathcal{S}} P_{h'}(\bar{X}_{t+1}=i|A_{t+1}=a,A^t,\bar{X}^t)\,\exp(-Z_{hh'}(t))$$

$$= \sum_{i \in \mathcal{S}} \frac{P_{h'}(\bar{X}_{t+1} = i | A_{t+1} = a, A^t, \bar{X}^t)}{P_h(\bar{X}_{t+1} = i | A_{t+1} = a, A^t, \bar{X}^t)} \exp(-Z_{hh'}(t)) \, P_h(\bar{X}_{t+1} = i | A_{t+1} = a, A^t, \bar{X}^t)$$

$$= \sum_{i \in \mathcal{S}} \exp(-Z_{hh'}(t+1, a, i)) \, P_h(\bar{X}_{t+1} = i | A_{t+1} = a, A^t, \bar{X}^t). \tag{78}$$

where in (78), the quantity $Z_{hh'}(t+1, a, i)$ is defined as

$$Z_{hh'}(t+1, a, i) \coloneqq Z_{hh'}(t) + \log \frac{P_h(\bar{X}_{t+1} = i | A_{t+1} = a, A^t, \bar{X}^t)}{P_{h'}(\bar{X}_{t+1} = i | A_{t+1} = a, A^t, \bar{X}^t)}.$$

Substituting (78) in (77) and simplifying, we get

$$\begin{aligned}
&E_{h'}[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1}) | \mathcal{F}_t] \, \exp(-Z_{hh'}(t)) \\
&= \sum_{b=1}^{K} \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} \Big[g(B^t, A^t, \bar{X}^t, b, a, i) \cdot P_h(B_{t+1} = b | B^t, A^t, \bar{X}^t) \\
&\qquad \cdot P_h(A_{t+1} = a | B_{t+1} = b, B^t, A^t, \bar{X}^t) \cdot P_h(\bar{X}_{t+1} = i | B_{t+1} = b, A_{t+1} = a, B^t, A^t, \bar{X}^t) \cdot \exp(-Z_{hh'}(t+1, a, i)) \Big]
\end{aligned} \tag{79}$$

$$= E_h[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1}) | \mathcal{F}_t]. \tag{80}$$

Applying $E_h[\cdot]$ to both sides of (80) and using (76) along with the law of iterated expectations, we arrive at the desired relation. This proves (70) for all $t \geq 0$.

Finally, for any $E \in \mathcal{F}_{\tau(\pi)}$, we have

$$\begin{aligned}
P_{h'}(E) &= E_{h'}[1_E] \\
&= E_{h'} \left[\sum_{t \geq 0} 1_{E \cap \{\tau(\pi) = t\}} \right] \\
&\overset{(a)}{=} \sum_{t \geq 0} E_{h'} \left[1_{E \cap \{\tau(\pi) = t\}} \right] \\
&\overset{(b)}{=} \sum_{t \geq 0} E_h \left[1_{E \cap \{\tau(\pi) = t\}} \, \exp(-Z_{hh'}(t)) \right] \\
&= \sum_{t \geq 0} E_h \left[1_{E \cap \{\tau(\pi) = t\}} \, \exp(-Z_{hh'}(\tau(\pi))) \right] \\
&= E_h \left[1_E \, \exp(-Z_{hh'}(\tau(\pi))) \right],
\end{aligned} \tag{81}$$

where (a) is due to monotone convergence theorem, and (b) above follows from (70) and the fact that $E \cap \{\tau(\pi) = t\} \in \mathcal{F}_t$ for all $t \geq 0$ since $E \in \mathcal{F}_{\tau(\pi)}$. This completes the proof of the Lemma. ■

Lemma 5, in conjunction with [14, Lemma 19], yields the following inequality for all policies $\pi \in \Pi(\epsilon)$ and all $h' \neq h$:

$$E_h[Z_{hh'}(\tau(\pi))] \geq \sup_{E \in \mathcal{F}_{\tau(\pi)}} d(P_h(E), P_{h'}(E)), \tag{82}$$

where for any $x, y \in [0, 1]$,

$$d(x, y) \coloneqq x \log(x/y) + (1 - x) \log((1 - x)/(1 - y))$$

is the binary relative entropy function. As noted in [14], $x \mapsto d(x, y)$ is monotone increasing for $x < y$ and the $y \mapsto d(x, y)$ is monotone decreasing for any fixed x. Also, for any $\pi \in \Pi(\epsilon)$, we have

$$P_h(\theta(\pi) = h) \geq 1 - \epsilon, \quad P_{h'}(\theta(\pi) = h) \leq \epsilon$$

for all $h' \neq h$. Combining the aforementioned facts, we get

$$\min_{h' \neq h} E_h[Z_{hh'}(\tau(\pi))] \geq d(\epsilon, 1-\epsilon). \tag{83}$$

for all $\pi \in \Pi(\epsilon)$.

B. An Upper Bound for $E_h[Z_{hh'}(\tau(\pi))]$ in Terms of $E_h[\tau(\pi)]$

We now obtain an upper bound for the left-hand side of (83). Fix $\pi \in \Pi(\epsilon)$ and $h' \neq h$ arbitrarily. Then, from (31),

$$\begin{aligned} &E_h[Z_{hh'}(\tau(\pi))] \\ &= E_h\left[\sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)}\right] + E_h\left[\sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \sum_{j \in \mathcal{S}} N(\tau(\pi), \underline{d}, \underline{i}, a, j) \log \frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)}\right]. \end{aligned} \tag{84}$$

To simplify the second expectation term on the right-hand side of (84), we use the following Lemma.

Lemma 6. *Fix $h \in \mathcal{A}$. For any policy π, let E_h and $E_{h'}$ denote the expectations computed under hypothesis $\mathcal{H}_h$ and under policy π. Then, for all $(\underline{d}, \underline{i}) \in \mathbb{S}$, $a \in \mathcal{A}$ and $j \in \mathcal{S}$,*

$$E_h[E_h[N(\tau(\pi), \underline{d}, \underline{i}, a, j)|X_{a-1}^a]|\tau(\pi)] = E_h[E_h[N(\tau(\pi), \underline{d}, \underline{i}, a)|X_{a-1}^a]|\tau(\pi)]\,(P_h^a)^{d_a}(j|i_a). \tag{85}$$

Proof of Lemma 6: Substituting $n = \tau(\pi)$ in (29), we have

$$\begin{aligned} E_h[E_h[N(\tau(\pi), \underline{d}, \underline{i}, a, j)|X_{a-1}^a]|\tau(\pi)] &= E_h\left[E_h\left[\sum_{t=K}^{\tau(\pi)} 1_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a, X_t^a=j\}} \middle| X_{a-1}^a\right] \middle| \tau(\pi)\right] \\ &= E_h\left[\sum_{t=K}^{\tau(\pi)} P_h(\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a, X_t^a=j|X_{a-1}^a) \middle| \tau(\pi)\right]. \end{aligned} \tag{86}$$

For each t in the range of the summation in (86), the conditional probability term for t may be expressed as

$$\begin{aligned} &P_h(\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a, X_t^a=j|X_{a-1}^a) \\ &= P_h(\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a|X_{a-1}^a) \cdot P_h(X_t^a=j|A_t=a, \underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, X_{a-1}^a) \\ &= P_h(\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a|X_{a-1}^a) \cdot (P_h^a)^{d_a}(j|i_a). \end{aligned} \tag{87}$$

Plugging (87) back in (86) and simplifying, we arrive at the desired relation in (85). ■

Using Lemma 6, the second expectation term on the right-hand side of (84) can be simplified as follows.

$$\begin{aligned} &E_h\left[\sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \sum_{j \in \mathcal{S}} N(\tau(\pi), \underline{d}, \underline{i}, a, j) \log \frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)}\right] \\ &= E_h\left[\sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \sum_{j \in \mathcal{S}} N(\tau(\pi), \underline{d}, \underline{i}, a, j) \log \frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)}\right] \\ &= E_h\left[\sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \sum_{j \in \mathcal{S}} E_h[E_h[N(\tau(\pi), \underline{d}, \underline{i}, a, j)|X_{a-1}^a]|\tau(\pi)] \log \frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)}\right] \\ &\overset{(a)}{=} E_h\left[\sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \sum_{j \in \mathcal{S}} E_h[E_h[N(\tau(\pi), \underline{d}, \underline{i}, a)|X_{a-1}^a]|\tau(\pi)] \cdot (P_h^a)^{d_a}(j|i) \cdot \log \frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)}\right] \\ &= E_h\left[\sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} E_h[E_h[N(\tau(\pi), \underline{d}, \underline{i}, a)|X_{a-1}^a]|\tau(\pi)] \cdot D((P_h^a)^{d_a}(\cdot|i_a) \| (P_{h'}^a)^{d_a}(\cdot|i_a))\right] \end{aligned}$$

$$= \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} E_h[N(\tau(\pi),\underline{d},\underline{i},a)] \cdot D((P_h^a)^{d_a}(\cdot|i_a)\|(P_{h'}^a)^{d_a}(\cdot|i_a)), \tag{88}$$

where in the above set of equations, (a) follows from Lemma 6, and (88) is due to monotone convergence theorem and the fact that

$$E_h[E_h[E_h[N(\tau(\pi),\underline{d},\underline{i},a)|X_{a-1}^a]|\tau(\pi)]] = E_h[N(\tau(\pi),\underline{d},\underline{i},a)].$$

Plugging (88) back in (84), we get

$$\begin{aligned} &E_h[Z_{hh'}(\tau(\pi))] \\ &= E_h\left[\sum_{a=1}^{K} \log\frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)}\right] + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} E_h[N(\tau(\pi),\underline{d},\underline{i},a)] \cdot D((P_h^a)^{d_a}(\cdot|i_a)\|(P_{h'}^a)^{d_a}(\cdot|i_a)). \end{aligned} \tag{89}$$

Noting that

$$\begin{aligned} \sum_{a=1}^{K}\sum_{(\underline{d},\underline{i}\in\mathbb{S}} E_h[N(\tau(\pi),\underline{d},\underline{i},a)] &\overset{(a)}{=} E_h\Big[\sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} N(\tau(\pi),\underline{d},\underline{i},a)\Big] \\ &= E_h\Big[\sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K}\sum_{t=K}^{\tau(\pi)} 1_{\{\underline{d}^\pi(t)=\underline{d},\underline{i}^\pi(t)=\underline{i},A_t=a\}}\Big] \\ &= E_h\Big[\sum_{t=K}^{\tau(\pi)} 1\Big] \qquad (90) \\ &= E_h[\tau(\pi)-K+1], \qquad (91) \end{aligned}$$

where (a) above is due to monotone convergence theorem, we write (89) as

$$\begin{aligned} &E_h[Z_{hh'}(\tau(\pi))] \\ &= E_h\left[\sum_{a=1}^{K} \log\frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)}\right] + \Big(E_h[\tau(\pi)-K+1]\Big) \cdot \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \frac{E_h[N(\tau(\pi),(\underline{d},\underline{i}),a)]}{E_h[\tau(\pi)-K+1]} \cdot D((P_h^a)^{d_a}(\cdot|i_a)\|(P_{h'}^a)^{d_a}(\cdot|i_a)). \end{aligned} \tag{92}$$

Combining (83) and (92), and noting that (92) holds for all $h' \neq h$, we get

$$\begin{aligned} d(\epsilon,1-\epsilon) &\leq \min_{h'\neq h}\Bigg\{E_h\left[\sum_{a=1}^{K} \log\frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)}\right] \\ &\qquad + \Big(E_h[\tau(\pi)-K+1]\Big) \cdot \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \frac{E_h[N(\tau(\pi),\underline{d},\underline{i},a)]}{E_h[\tau(\pi)-K+1]} \cdot D((P_h^a)^{d_a}(\cdot|i_a)\|(P_{h'}^a)^{d_a}(\cdot|i_a))\Bigg\} \\ &\leq \sup_{\nu}\min_{h'\neq h}\Bigg\{E_h\left[\sum_{a=1}^{K} \log\frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)}\right] \\ &\qquad + \Big(E_h[\tau(\pi)-K+1]\Big) \cdot \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a)\, D((P_h^a)^{d_a}(\cdot|i_a)\|(P_{h'}^a)^{d_a}(\cdot|i_a))\Bigg\}, \end{aligned} \tag{93}$$

where the supremum in (93) is over all state-action occupancy measures satisfying

$$\sum_{a=1}^{K} \nu(\underline{d}',\underline{i}',a) = \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a)\, Q(\underline{d}',\underline{i}'|\underline{d},\underline{i},a) \quad \text{for all } (\underline{d}',\underline{i}')\in\mathbb{S}, \tag{94}$$

$$\sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a) = 1, \tag{95}$$

$$\nu(\underline{d},\underline{i},a) \geq 0 \quad \text{for all } (\underline{d},\underline{i},a)\in\mathbb{S}\times\mathcal{A}. \tag{96}$$

Recall that Q in (94) denotes the transition probability matrix given by (7). The left-hand side of (94) represents the long-term probability of leaving the state $(\underline{d}, \underline{i})$, while the right-hand side of (95) represents the long-term probability of entering into the state $(\underline{d}, \underline{i})$. Thus, (94) is the *global balance equation* for the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$. Equations (95) and (96) together imply that ν is a probability measure on $\mathbb{S} \times \mathcal{A}$.

As outlined in Section III, the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$, together with the sequence $\{B_t : t \geq 0\}$ of intended arm selections (or equivalently the sequence $\{A_t : t \geq 0\}$ of actual arm selections), defines a Markov decision problem (MDP) with state space $\mathbb{S}$ and action space $\mathcal{A}$. Note that $\mathbb{S}$ is a countable set. A simple extension of [16, Theorem 8.8.2] to countable state space MDPs, in conjunction with Lemma 1, implies a one-one correspondence between any feasible solution to (94)-(96) and policies in Π_{SRS}. In other words, [16, Theorem 8.8.2] implies that for any given ν satisfying (94)-(96), we can find an SRS policy $\pi^\lambda \in \Pi_{\mathsf{SRS}}$ such that $\nu^\lambda(\underline{d}, \underline{i}, a) = \nu(\underline{d}, \underline{i}, a)$ for all $(\underline{d}, \underline{i}, a) \in \mathbb{S} \times \mathcal{A}$. Recall that under the SRS policy π^λ, the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is, in fact, a Markov process whose transition probability matrix is ergodic (Lemma 1) and possesses μ^λ as its unique stationary distribution. The associated ergodic state occupancy measure, ν^λ, is then defined according to (15).

On account of [16, Theorem 8.8.2], we may replace the supremum in (93) by a supremum over all SRS policies. Doing so leads us to the relation

$$d(\epsilon, 1-\epsilon) \leq \sup_{\pi^\lambda \in \Pi_{\mathsf{SRS}}} \min_{h' \neq h} \Bigg\{ E_h \left[\sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)} \right] + \Big(E_h[\tau(\pi) - K + 1] \Big) \cdot \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu^\lambda(\underline{d}, \underline{i}, a)\, D((P_h^a)^{d_a}(\cdot|i_a) \| (P_{h'}^a)^{d_a}(\cdot|i_a)) \Bigg\}. \tag{97}$$

for all $\pi \in \Pi(\epsilon)$. Observe that the constant term multiplying $(E_h[\tau(\pi) - K + 1])$ in (97) is finite; further, it is not a function of either ϵ or of $\pi \in \Pi(\epsilon)$. The finiteness of this constant follows from the following observation: denote by μ_h^a the stationary distribution of the transition probability matrix P_h^a (i.e., $\mu_h^a = \mu_1$ for $a = h$ and $= \mu_2$ for all $a \neq h$). An application of the ergodic theorem to the Markov process of arm a yields

$$D((P_h^a)^{d_a}(\cdot|i_a) \| (P_{h'}^a)^{d_a}(\cdot|i_a)) \longrightarrow D(\mu_h^a \| \mu_{h'}^a) < \infty \quad \text{as } d_a \to \infty. \tag{98}$$

Since every convergent sequence is bounded, we may write $D((P_h^a)^{d_a}(\cdot|i_a) \| (P_{h'}^a)^{d_a}(\cdot|i_a)) \leq C$ for all $(\underline{d}, \underline{i}, a) \in \mathbb{S} \times \mathcal{A}$, where $0 < C < \infty$. Using (95), it follows that the constant term multiplying $(E_h[\tau(\pi) - K + 1])$ in (97) is bounded above by C.

Let us also note that the first term inside the braces in (97) does not depend on ϵ. Since $d(\epsilon, 1-\epsilon) \to d(0,1) = +\infty$ as $\epsilon \downarrow 0$, the boundedness of $R^*(P_1, P_2)$ shows that $\epsilon \downarrow 0$ is equivalent to $E_h[\tau(\pi)] \to \infty$ for all $\pi \in \Pi(\epsilon)$. Letting $\epsilon \downarrow 0$, and using $d(\epsilon, 1-\epsilon)/\log(1/\epsilon) \longrightarrow 1$ as $\epsilon \downarrow 0$, we arrive at the lower bound in (12). This completes the proof of the Proposition.

APPENDIX C
PROOF OF LEMMA 2

We begin the proof of this Lemma by recalling that the key ingredient in the proof of Lemma 1 is the fact that for each $t \geq K$, the probability term in (60) is $\geq \eta/K > 0$ whenever the trembling hand parameter $\eta > 0$. This property holds true even for the policy $\pi^\star(L, \delta)$. We leverage this to first show below that for all $(\underline{d}, \underline{i}) \in \mathbb{S}$,

$$\liminf_{n \to \infty} \frac{N(n, \underline{d}, \underline{i})}{n} > 0 \quad \text{almost surely} \tag{99}$$

under the policy $\pi^\star(L, \delta)$, where for each $n \geq K$,

$$N(n, \underline{d}, \underline{i}) := \sum_{t=K}^{n} \mathbb{I}_{\{\underline{d}(t) = \underline{d},\ \underline{i}(t) = \underline{i}\}} \tag{100}$$

denotes the number of times the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ visits the state $(\underline{d}, \underline{i})$. From [20, Proposition 1.7], we know that there exists an integer M such that for all $m \geq M$,

$$P_1^m(j|i) > 0 \text{ for all } i, j \in \mathcal{S}, \quad P_2^m(j|i) > 0 \text{ for all } i, j \in \mathcal{S}. \tag{101}$$

Fix an arbitrary $(\underline{d}, \underline{i}) \in \mathbb{S}$, and assume without loss of generality that $\underline{d}$ is such that $d_1 > d_2 > \cdots > d_K = 1$. Also assume, again without loss of generality, that the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ starts in the state $(\underline{d}, \underline{i})$, i.e., $\underline{d}(K) = \underline{d}$, $\underline{i}(K) = \underline{i}$. From Appendix A, we know that the probability of the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ starting in the state $(\underline{d}, \underline{i})$ and returning back to the state $(\underline{d}, \underline{i})$ after $M + d_1 - d_K$ time instants, call this $p(\underline{d}, \underline{i})$, is lower bounded by the quantity in (63). Since (63) is strictly positive, it follows that $p(\underline{d}, \underline{i}) > 0$.

Clearly, then, the term $N(n, \underline{d}, \underline{i})$ may be lower bounded almost surely by the number of visits to the state $(\underline{d}, \underline{i})$ measured only at times $t = K + M + d_1 - d_K, K + 2(M + d_1 - d_K), K + 3(M + d_1 - d_K)$ and so on until time $t = n$. Note that at each of these time instants, the probability that the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is in the state $(\underline{d}, \underline{i})$ is equal to $p(\underline{d}, \underline{i})$. Thus, we have

$$N(n, \underline{d}, \underline{i}) \geq \text{Bin}\left(\frac{n - K + 1}{M + d_1 - d_K}, \quad p(\underline{d}, \underline{i})\right) \quad \text{almost surely,} \tag{102}$$

where the notation $\text{Bin}(m, q)$ denotes a Binomial random variable with parameters m and q. It then follows that, almost surely,

$$\begin{aligned}
\liminf_{n\to\infty} \frac{N(n, \underline{d}, \underline{i})}{n} &\geq \liminf_{n\to\infty} \frac{\text{Bin}\left(\frac{n-K+1}{M+d_1-d_K}, \quad p(\underline{d}, \underline{i})\right)}{n} \\
&= \liminf_{n\to\infty} \frac{\text{Bin}\left(\frac{n-K+1}{M+d_1-d_K}, \quad p(\underline{d}, \underline{i})\right)}{\frac{n-K+1}{M+d_1-d_K}} \cdot \frac{n - K + 1}{n} \cdot \frac{1}{M + d_1 - d_K} \\
&\overset{(a)}{=} \frac{p(\underline{d}, \underline{i})}{M + d_1 - d_K} \\
&> 0,
\end{aligned} \tag{103}$$

where (a) above follows from the strong law of large numbers. This establishes (99).

We now show that for all $a \in \mathcal{A}$,

$$\liminf_{n\to\infty} \frac{N(n, \underline{d}, \underline{i}, a)}{n} > 0 \quad \text{almost surely.} \tag{104}$$

Subsequently, we use (104) to establish (34). Fix an arbitrary $a \in \mathcal{A}$, and define

$$S(n, \underline{d}, \underline{i}, a) \coloneqq \sum_{t=K}^{n} \left[\mathbb{I}_{\{A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}\}} - P(A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | B^{t-1}, A^{t-1}, \bar{X}^{t-1})\right]. \tag{105}$$

For each $t \geq K$, since $|\mathbb{I}_{\{A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}\}} - P(A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | B^{t-1}, A^{t-1}, \bar{X}^{t-1})| \leq 2$ almost surely, and

$$E[\mathbb{I}_{\{A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}\}} - P(A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | B^{t-1}, A^{t-1}, \bar{X}^{t-1}) | A^{t-1}, \bar{X}^{t-1}] = 0,$$

the collection $\{\mathbb{I}_{\{A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}\}} - P(A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | B^{t-1}, A^{t-1}, \bar{X}^{t-1})\}_{t \geq K}$ is a bounded martingale difference sequence. Using [21, Theorem 1.2A], we get that

$$\frac{S(n, \underline{d}, \underline{i}, a)}{n} \longrightarrow 0 \quad \text{almost surely} \tag{106}$$

as $n \to \infty$. This implies that for every choice of $\varepsilon > 0$, there exists N_ε sufficiently large such that

$$\frac{N(n, \underline{d}, \underline{i}, a)}{n} \geq \frac{1}{n} \sum_{t=K}^{n} P(A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | B^{t-1}, A^{t-1}, \bar{X}^{t-1}) - \varepsilon \quad \text{for all } n \geq N_\varepsilon \text{ almost surely.} \tag{107}$$

Now, for each $t \geq K$,

$$\begin{aligned}
&P(A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | B^{t-1}, A^{t-1}, \bar{X}^{t-1}) \\
&= P(A_t = a | \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, B^{t-1}, A^{t-1}, \bar{X}^{t-1}) \cdot P(\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | B^{t-1}, A^{t-1}, \bar{X}^{t-1}) \\
&= \left[\frac{\eta}{K} + (1-\eta)\, \lambda_{\theta(t),\delta}(a \mid \underline{d}, \underline{i}) \right] P(\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | B^{t-1}, A^{t-1}, \bar{X}^{t-1}) \\
&\geq \frac{\eta}{K} \cdot \mathbb{I}_{\{(\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}\}}, \qquad (108)
\end{aligned}$$

where (108) follows from the fact that $\underline{d}(t)$ and $\underline{i}(t)$ are measurable with respect to the history $(B^{t-1}, A^{t-1}, \bar{X}^{t-1})$. Plugging (108) in (107), we get

$$\frac{N(n, \underline{d}, \underline{i}, a)}{n} \geq \frac{\eta}{K} \cdot \frac{N(n, \underline{d}, \underline{i})}{n} - \varepsilon \quad \text{almost surely} \qquad (109)$$

for all $n \geq N_\varepsilon$. Using (103) in (109), we get

$$\frac{N(n, \underline{d}, \underline{i}, a)}{n - K + 1} \geq \frac{\eta}{K} \cdot \frac{p(\underline{d}, \underline{i})}{2(M + d_1 - d_K)} - \varepsilon \quad \text{almost surely} \qquad (110)$$

for all sufficiently large values of n. Setting $\varepsilon = \frac{\eta}{2K} \cdot \frac{p(\underline{d}, \underline{i})}{2(M + d_1 - d_K)}$ establishes (104).

Proof of Lemma 2: For any $h' \neq h$, we have

$$\frac{1}{n} Z_{hh'}(n) = \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, h, j)}{n} \log \frac{P_1^{d_h}(j | i_h)}{P_2^{d_h}(j | i_h)} + \frac{N(n, \underline{d}, \underline{i}, h', j)}{n} \log \frac{P_1^{d_{h'}}(j | i_{h'})}{P_2^{d_{h'}}(j | i_{h'})}. \qquad (111)$$

Since $N(n, \underline{d}, \underline{i}, a) \to \infty$ almost surely as $n \to \infty$ (this follows from the fact that $\liminf_{n \to \infty} N(n, \underline{d}, \underline{i}, a)/n > 0$ almost surely) for every $a \in \mathcal{A}$, we apply the Ergodic theorem to deduce that

$$\frac{N(n, \underline{d}, \underline{i}, a, j)}{N(n, \underline{d}, \underline{i}, a)} \longrightarrow (P_h^a)^{d_a}(j | i_a) \quad \text{as } n \to \infty \quad \text{almost surely.} \qquad (112)$$

Using (112) in (111), we get that for every choice of ε, there exists N_ε sufficiently large such that for all $n \geq N_\varepsilon$, almost surely,

$$\begin{aligned}
\frac{1}{n} Z_{hh'}(n) \geq & \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, h)}{n} (P_1^{d_h}(j | i_h) + \varepsilon) \log P_1^{d_h}(j | i_h) \\
& + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, h)}{n} (P_1^{d_h}(j | i_h) - \varepsilon) \log \frac{1}{P_2^{d_h}(j | i_h)} \\
& + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, h')}{n} (P_2^{d_{h'}}(j | i_{h'}) + \varepsilon) \log P_2^{d_{h'}}(j | i_{h'}) \\
& + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, h')}{n} (P_2^{d_{h'}}(j | i_{h'}) - \varepsilon) \log \frac{1}{P_1^{d_{h'}}(j | i_{h'})} \\
= & \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \frac{N(n, \underline{d}, \underline{i}, h)}{n} D(P_1^{d_h}(\cdot | i_h) \| P_2^{d_h}(\cdot | i_h)) + \frac{N(n, \underline{d}, \underline{i}, h')}{n} D(P_2^{d_{h'}}(\cdot | i_{h'}) \| P_1^{d_{h'}}(\cdot | i_{h'})) \\
& + \varepsilon \left[\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \frac{N(n, \underline{d}, \underline{i}, h)}{n} \left(\sum_{j \in \mathcal{S}} \log P_1^{d_h}(j | i_h)\, P_2^{d_h}(j | i_h) \right) + \frac{N(n, \underline{d}, \underline{i}, h')}{n} \left(\sum_{j \in \mathcal{S}} \log P_1^{d_{h'}}(j | i_{h'})\, P_2^{d_{h'}}(j | i_{h'}) \right) \right]. \qquad (113)
\end{aligned}$$

As a consequence of the convergence theorem for finite state Markov processes [20, Theorem 4.9], we have

$$\begin{aligned}
P_1^d(j | i) &\longrightarrow \mu_1(j) > 0 \quad \text{as } d \to \infty \\
P_2^d(j | i) &\longrightarrow \mu_2(j) > 0 \quad \text{as } d \to \infty \qquad (114)
\end{aligned}$$

for all $i, j \in \mathcal{S}$. This implies that the term inside the square brackets in (113) is bounded from below (say by a constant $C < 0$). We then have

$$\begin{aligned}\frac{1}{n} Z_{hh'}(n) &\geq \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \frac{N(n,\underline{d},\underline{i},h)}{n} D(P_1^{d_h}(\cdot|i_h) \| P_2^{d_h}(\cdot|i_h)) + \frac{N(n,\underline{d},\underline{i},h')}{n} D(P_2^{d_{h'}}(\cdot|i_{h'}) \| P_1^{d_{h'}}(\cdot|i_{h'})) + C\varepsilon \\ &\geq \frac{N(n,\underline{d},\underline{i},h)}{n} D(P_1^{d_h}(\cdot|i_h) \| P_2^{d_h}(\cdot|i_h)) + \frac{N(n,\underline{d},\underline{i},h')}{n} D(P_2^{d_{h'}}(\cdot|i_{h'}) \| P_1^{d_{h'}}(\cdot|i_{h'})) + C\varepsilon \end{aligned} \tag{115}$$

for all $(\underline{d},\underline{i}) \in \mathbb{S}$ and for all $n \geq N_\varepsilon$, almost surely. Now, fix an arbitrary $(\underline{d},\underline{i}) \in \mathbb{S}$ such that $d_1 > d_2 > \cdots > d_K = 1$. From (109), we know that there exist constants $N_h, N_{h'}$ sufficiently large such that

$$\frac{N(n,\underline{d},\underline{i},h)}{n} \geq \frac{\eta}{K} \cdot \frac{p(\underline{d},\underline{i})}{2(M + d_1 - d_K)} - \varepsilon, \quad \frac{N(n,\underline{d},\underline{i},h')}{n} \geq \frac{\eta}{K} \cdot \frac{p(\underline{d},\underline{i})}{2(M + d_1 - d_K)} - \varepsilon \tag{116}$$

for all $n \geq \max\{N_h, N_{h'}, N_\varepsilon\}$, almost surely. Combining (116) and (115), we may choose $\varepsilon > 0$ appropriately so that the right-hand side of (115) is strictly positive. This establishes the desired result. ■

APPENDIX D
PROOF OF LEMMA 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, thanks to Lemma 2. Thus, the probability of error of policy $\pi = \pi^\star(L,\delta)$ may be evaluated as follows: suppose $\mathcal{H}_h$ is the true hypothesis. Then,

$$P_h(\theta(\tau(\pi)) \neq h) = P_h\Big(\exists\ n \text{ and } \ h' \neq h \text{ such that } \theta(\tau(\pi)) = h' \text{ and } \tau(\pi) = n\Big). \tag{117}$$

We now let

$$\mathcal{R}_{h'}(n) \coloneqq \{\omega : \tau(\pi)(\omega) = n,\ \theta(\tau(\pi))(\omega) = h'\} \tag{118}$$

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

$$\begin{aligned} P_h(\theta(\tau(\pi)) \neq h) &= P_h\left(\bigcup_{h' \neq h} \bigcup_{n=0}^{\infty} \mathcal{R}_{h'}(n)\right) \\ &= \sum_{h' \neq h} \sum_{n=0}^{\infty} P_h(\tau(\pi) = n,\ \theta(\tau(\pi)) = h') \\ &= \sum_{h' \neq h} \sum_{n=0}^{\infty} \int\limits_{\mathcal{R}_{h'}(n)} dP_h(\omega) \\ &\overset{(a)}{=} \sum_{h' \neq h} \sum_{n=0}^{\infty} \int\limits_{\mathcal{R}_{h'}(n)} \exp(Z_h(n)(\omega)) \quad d(B^n(\omega), A^n(\omega), \bar{X}^n(\omega)) \\ &\overset{(b)}{=} \sum_{h' \neq h} \sum_{n=0}^{\infty} \int\limits_{\mathcal{R}_{h'}(n)} \exp(-Z_{h'h}(n)(\omega)) \quad \exp(Z_{h'}(n)(\omega)) \quad d(B^n(\omega), A^n(\omega), \bar{X}^n(\omega)) \\ &\overset{(c)}{\leq} \sum_{h' \neq h} \sum_{n=0}^{\infty} \left\{ \int\limits_{\mathcal{R}_{h'}(n)} \frac{1}{(K-1)L}\ dP_{h'}(\omega) \right\} \end{aligned}$$

$$= \sum_{h' \neq h} \frac{1}{(K-1)L} \; P_{h'} \left(\bigcup_{n=0}^{\infty} \mathcal{R}_{h'}(n) \right) \leq \frac{1}{L}, \tag{119}$$

where in (a) above,

$$Z_h(n) \coloneqq \log P_h(B^n, A^n, \bar{X}^n)$$

denotes the log-likelihood of all the intended arm pulls, the actual arm pulls and the observations up to time n under the hypothesis $\mathcal{H}_h$, (b) above follows by noting that for $h \neq h'$, $Z_{hh'}(n) = Z_h(n) - Z_{h'}(n) = -Z_{h'h}(n)$, and (c) follows from the fact that when $\mathcal{H}_{h'}$ is the true hypothesis, the condition $M_{h'}(n) \geq \log((K-1)L)$ is satisfied when the policy $\pi = \pi^\star(L, \delta)$ stops at time $\tau(\pi) = n$, which in particular implies that $Z_{h'h}(n) \geq \log((K-1)L)$. Finally, setting $L = 1/\epsilon$ yields the desired result. This completes the proof of the Lemma.

APPENDIX E
PROOF OF PROPOSITION 2

This section is organised as follows. First, we show in Proposition 3 that under the policy $\pi^\star(L, \delta)$, the test statistic $M_h(n)$ has the correct drift, one that comes from the ergodic occupancy measure corresponding to $\pi^{\lambda_{h,\delta}}$ when $\mathcal{H}_h$ is the true hypothesis. We then show in Lemma 7 that the stopping time of the policy $\pi^\star(L, \delta)$ grows with L (i.e., lower probability of error implies more time required to stop and declare the odd arm location correctly with high confidence). More specifically, we show in Lemma 8 that ratio $\tau(\pi)/\log L$ has, in the limit as $L \to \infty$, an almost sure upper bound that matches with the right-hand side of (35). Finally, we prove in Proposition 4 that the family $\{\tau(\pi)/\log L : L > 1\}$ is uniformly integrable. The almost sure upper bound of Lemma 8 combined with uniform integrability result of Proposition 4 yields the desired upper bound in (35).

Proposition 3. *Fix an arbitrary $L > 1$, $\delta > 0$ and $h \in \mathcal{A}$, and let $\mathcal{H}_h$ be the true hypothesis. For every $h' \neq h$, under the non-stopping version of policy $\pi^\star(L, \delta)$, we have, almost surely,*

$$\lim_{n\to\infty} \frac{Z_{hh'}(n)}{n} = \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \nu^{\lambda_{h,\delta}}(\underline{d}, \underline{i}, h)\, D(P_1^{d_h}(\cdot|i_h) \| P_2^{d_h}(\cdot|i_h)) + \nu^{\lambda_{h,\delta}}(\underline{d}, \underline{i}, h')\, D(P_2^{d_{h'}}(\cdot|i_{h'}) \| P_1^{d_{h'}}(\cdot|i_{h'})). \tag{120}$$

Consequently, it follows that almost surely,

$$\lim_{n\to\infty} \frac{M_h(n)}{n} = \min_{h' \neq h} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \nu^{\lambda_{h,\delta}}(\underline{d}, \underline{i}, h)\, D(P_1^{d_h}(\cdot|i_h) \| P_2^{d_h}(\cdot|i_h)) + \nu^{\lambda_{h,\delta}}(\underline{d}, \underline{i}, h')\, D(P_2^{d_{h'}}(\cdot|i_{h'}) \| P_1^{d_{h'}}(\cdot|i_{h'})). \tag{121}$$

Proof of Proposition 3: From Lemma 2, it follows that when $\mathcal{H}_h$ is the true hypothesis,

$$\liminf_{n\to\infty} \frac{M_h(n)}{n} = \liminf_{n\to\infty} \min_{h' \neq h} \frac{Z_{hh'}(n)}{n} > 0 \quad \text{almost surely.} \tag{122}$$

This in turn implies that $\liminf_{n\to\infty} M_h(n) > 0$ almost surely. An immediate consequence of this is that for any $h' \neq h$, almost surely,

$$\begin{aligned}
\limsup_{n\to\infty} M_{h'}(n) &= \limsup_{n\to\infty} \min_{a \neq h'} Z_{h'a}(n) \\
&\leq \limsup_{n\to\infty} Z_{h'h}(n) \\
&= \limsup_{n\to\infty} -Z_{hh'}(n) \\
&= -\liminf_{n\to\infty} Z_{hh'}(n) \\
&\leq -\liminf_{n\to\infty} M_h(n) \\
&< 0.
\end{aligned} \tag{123}$$

The above set of inequalities imply the following important result: suppose $\mathcal{H}_h$ is the true hypothesis. Then, for any $L > 1$ and $\delta > 0$, under the non-stopping version of policy $\pi^\star(L, \delta)$, we have

$$\theta(n) = h \quad \text{for all sufficiently large values of } n, \text{ almost surely.} \tag{124}$$

The condition in (124) implies that for all $(\underline{d}, \underline{i}, a) \in \mathbb{S} \times \mathcal{A}$,

$$\begin{aligned}\lim_{n\to\infty} P(A_n = a|\underline{d}(n) = \underline{d}, \underline{i}(n) = \underline{i}, \{(\underline{d}(t), \underline{i}(t)) : K \le t < n\}) &= \lim_{n\to\infty} \frac{\eta}{K} + (1-\eta)\,\lambda_{\theta(n),\delta}(a|\underline{d}, \underline{i}) \\ &= \frac{\eta}{K} + (1-\eta)\,\lambda_{h,\delta}(a|\underline{d}, \underline{i}) \end{aligned} \tag{125}$$

which in turn leads to the following almost sure convergences as $n \to \infty$:

$$\frac{N(n, \underline{d}, \underline{i}, a)}{N(n, \underline{d}, \underline{i})} \longrightarrow \frac{\eta}{K} + (1-\eta)\,\lambda_{h,\delta}(a|\underline{d}, \underline{i}), \tag{126}$$

$$\frac{N(n, \underline{d}, \underline{i})}{n} \longrightarrow \mu^{\lambda_{h,\delta}}(\underline{d}, \underline{i}). \tag{127}$$

It now follows that for any $h' \neq h$, almost surely,

$$\begin{aligned}&\lim_{n\to\infty} \frac{Z_{hh'}(n)}{n} \\ &= \lim_{n\to\infty} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, h, j)}{n} \log \frac{P_1^{d_h}(j|i_h)}{P_2^{d_h}(j|i_h)} + \frac{N(n, \underline{d}, \underline{i}, h', j)}{n} \log \frac{P_2^{d_{h'}}(j|i_{h'})}{P_1^{d_{h'}}(j|i_{h'})} \\ &= \lim_{n\to\infty} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \left(\frac{N(n, \underline{d}, \underline{i})}{n}\right) \left(\frac{N(n, \underline{d}, \underline{i}, h)}{N(n, \underline{d}, \underline{i})}\right) \left(\frac{N(n, \underline{d}, \underline{i}, h, j)}{N(n, \underline{d}, \underline{i}, h)}\right) \log \frac{P_1^{d_h}(j|i_h)}{P_2^{d_h}(j|i_h)} \\ &\qquad + \lim_{n\to\infty} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \left(\frac{N(n, \underline{d}, \underline{i})}{n}\right) \left(\frac{N(n, \underline{d}, \underline{i}, h')}{N(n, \underline{d}, \underline{i})}\right) \left(\frac{N(n, \underline{d}, \underline{i}, h', j)}{N(n, \underline{d}, \underline{i}, h')}\right) \log \frac{P_2^{d_{h'}}(j|i_{h'})}{P_1^{d_{h'}}(j|i_{h'})}. \end{aligned} \tag{128}$$

Note that in each of the logarithmic terms in (128), when either the numerator or the denominator is equal to 0, the corresponding coefficient term is also equal to 0. Thus, we may assume without loss of generality that each term inside the summations in (128) is nonzero for all values of the summation indices. Under this assumption, it follows from the convergences in (114) that the logarithmic terms in (128) are bounded. Using the dominated convergence theorem to pass the limit inside the summation in each of the terms, and using the results in (112), (126) and (127), we arrive at the desired result. ■

We now show that the stopping time of policy $\pi^\star(L, \delta)$ grows with L.

Lemma 7. *Fix $h \in \mathcal{A}$ and $\delta > 0$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Then, under policy $\pi = \pi^\star(L, \delta)$, we have*

$$\liminf_{L\to\infty} \tau(\pi) = \infty \text{ almost surely.} \tag{129}$$

Proof of Lemma 7: Assume without loss of generality that the policy $\pi = \pi^\star(L, \delta)$ pulls arm 1 at time $t = 0$, arm 2 at time $t = 1$ and so on until arm K at time $t = K - 1$. In order to prove the Lemma, we note that it suffices to prove the following statement:

$$\text{for each } m \ge K, \quad \lim_{L\to\infty} P_h(\tau(\pi) \le m) = 0. \tag{130}$$

Fix $m \ge K$, and note that

$$\begin{aligned}\limsup_{L\to\infty} P_h(\tau(\pi) \le m) &= \limsup_{L\to\infty} P_h\Big(\exists\ K \le n \le m \text{ and } \tilde{h} \in \mathcal{A} \text{ such that } M_{\tilde{h}}(n) > \log((K-1)L)\Big) \\ &\le \limsup_{L\to\infty} \sum_{\tilde{h}\in\mathcal{A}} \sum_{n=K}^{m} P_h(M_{\tilde{h}}(n) > \log((K-1)L)) \end{aligned}$$

$$\leq \limsup_{L\to\infty} \frac{1}{\log((K-1)L)} \sum_{\tilde{h}\in\mathcal{A}} \sum_{n=K}^{m} E_h[M_{\tilde{h}}(n)], \tag{131}$$

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

We now show that for each $n \in \{K, \ldots, m\}$, the expectation term inside the summation in (131) is finite. This will then imply that the limit supremum on the right-hand side of (131) is equal to 0, thus proving the desired result. Note that

$$M_{\tilde{h}}(n) = \min_{h'\neq\tilde{h}} Z_{\tilde{h}h'}(n) \leq Z_{\tilde{h}h'}(n) \text{ for all } h' \neq \tilde{h}. \tag{132}$$

Fix an arbitrary $h' \neq \tilde{h}$. Then, almost surely,

$$\begin{aligned} Z_{\tilde{h}h'}(n) &= \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} N(n,\underline{d},\underline{i},\tilde{h},j) \log \frac{P_1^{d_{\tilde{h}}}(j|i_{\tilde{h}})}{P_2^{d_{\tilde{h}}}(j|i_{\tilde{h}})} + N(n,\underline{d},\underline{i},h',j) \log \frac{P_2^{d_{h'}}(j|i_{h'})}{P_1^{d_{h'}}(j|i_{h'})} \\ &\leq n \max\left\{ \max\left\{ \log \frac{P_1^d(j|i)}{P_2^d(j|i)} : d\in\mathbb{N}, i,j\in\mathcal{S} \right\}, \max\left\{ \log \frac{P_2^d(j|i)}{P_1^d(j|i)} : d\in\mathbb{N}, i,j\in\mathcal{S} \right\} \right\}. \end{aligned} \tag{133}$$

From to the convergences in (114), we note that the coefficient of n in (133) is finite. Thus, it follows that $E[M_{\tilde{h}}(n)] \leq E[Z_{\tilde{h}h'}(n)] \leq nC$ for all $h' \neq \tilde{h}$, where $C < \infty$ represents the constant multiplying n in (133). ∎

Going further, let $R_{\lambda_{h,\delta}}$ denote the right-hand side of (121).

Lemma 8. *Fix $h \in \mathcal{A}$ and $\delta > 0$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Then, under policy $\pi = \pi^*(L,\delta)$, we have*

$$\limsup_{L\to\infty} \frac{\tau(\pi)}{\log L} \leq \frac{1}{R_{\lambda_{h,\delta}}} \quad \textit{almost surely}. \tag{134}$$

Proof of Lemma 8: Note that as a consequence of Proposition 3 and Lemma 7, we have

$$\lim_{L\to\infty} \frac{M_h(\tau(\pi))}{\tau(\pi)} = R_{\lambda_{h,\delta}} \quad \text{almost surely}. \tag{135}$$

We now show that for any $h' \neq h$ and $n \geq K$, the increment $Z_{hh'}(n) - Z_{hh'}(n-1)$ is bounded almost surely. Observe that, almost surely,

$$\begin{aligned} &Z_{hh'}(n) - Z_{hh'}(n-1) \\ &= \log \frac{P_h(A^n,\bar{X}^n)}{P_{h'}(A^n,\bar{X}^n)} - \log \frac{P_h(A^{n-1},\bar{X}^{n-1})}{P_{h'}(A^{n-1},\bar{X}^{n-1})} \\ &= \log \frac{P_h^{A_n}(\bar{X}_n|A^{n-1},\bar{X}^{n-1})}{P_{h'}^{A_n}(\bar{X}_n|A^{n-1},\bar{X}^{n-1})} \\ &= \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \sum_{j\in\mathcal{S}} \mathbb{I}_{\{\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},A_n=a,X_n^a=j\}} \log \frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)} \\ &= \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \left[\mathbb{I}_{\{\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},A_n=h,X_n^h=j\}} \log \frac{P_1^{d_h}(j|i_h)}{P_2^{d_h}(j|i_h)} + \mathbb{I}_{\{\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},A_n=h',X_n^{h'}=j\}} \log \frac{P_2^{d_{h'}}(j|i_{h'})}{P_1^{d_{h'}}(j|i_{h'})} \right]. \end{aligned} \tag{136}$$

We now note that whenever either the numerator or the denominator of the logarithmic terms in (136) is equal to 0, then the corresponding indicator function is also equal to 0. This, together with the convergences in (114), implies that the right-hand side of (136) is bounded. This, together with the collection $\{Z_{hh'}(n) - Z_{hh'}(n-1) : 1 \leq n \leq K-1\}$ of finitely many terms, each of which is finite almost surely, establishes the almost sure boundedness of the increments $Z_{hh'}(n) - Z_{hh'}(n-1)$ for all $n \geq 1$ and all $h' \neq h$.

When $\mathcal{H}_h$ is the true hypothesis, we note from the definition of stopping time $\tau(\pi)$ that $M_h(\tau(\pi)-1) < \log((K-1)L)$, which implies that there exists $h'' \neq h$ such that $Z_{hh''}(\tau(\pi)-1) < \log((K-1)L)$. Using this, we have

$$\limsup_{L\to\infty} \frac{M_h(\tau(\pi))}{\log L} = \limsup_{L\to\infty} \min_{h'\neq h} \frac{Z_{hh'}(\tau(\pi))}{\log L}$$

$$\begin{aligned}
&\leq \limsup_{L\to\infty} \frac{Z_{hh''}(\tau(\pi))}{\log L} \\
&\overset{(a)}{=} \limsup_{L\to\infty} \frac{Z_{hh''}(\tau(\pi)-1)}{\log L} \\
&\leq \limsup_{L\to\infty} \frac{\log((K-1)L)}{\log L} \\
&= 1 \quad \text{almost surely,}
\end{aligned} \tag{137}$$

where (a) above is due to the almost sure boundedness of the increments established above. Then, using (135) along with (137) yields

$$\begin{aligned}
\limsup_{L\to\infty} \frac{\tau(\pi)}{\log L} &= \limsup_{L\to\infty} \left\{ \left(\frac{\tau(\pi)}{M_h(\tau(\pi))} \right) \left(\frac{M_h(\tau(\pi))}{\log L} \right) \right\} \\
&= \left(\lim_{L\to\infty} \frac{\tau(\pi)}{M_h(\tau(\pi))} \right) \left(\limsup_{L\to\infty} \frac{M_h(\tau(\pi))}{\log L} \right) \\
&\leq \frac{1}{R_{\lambda_{h,\delta}}} \quad \text{almost surely,}
\end{aligned} \tag{138}$$

thus completing the proof of the Lemma. ■

Since, by definition, $R_{\lambda_{h,\delta}} > \frac{R^*(P_1,P_2)}{1+\delta}$, it follows that

$$\limsup_{L\to\infty} \frac{\tau(\pi)}{\log L} \leq \frac{1+\delta}{R^*(P_1,P_2)} \quad \text{almost surely.} \tag{139}$$

We now prove that the family $\{\tau(\pi^\star(L,\delta))/\log L : L > 1\}$ is uniformly integrable for all $\delta > 0$. This, along with the almost sure upper bound of (139) yields the desired upper bound of (35).

Proposition 4. *For any fixed $\delta > 0$, the family of random variables $\{\tau(\pi^\star(L,\delta))/\log L : L > 1\}$ is uniformly integrable.*

Proof of Proposition 4: Fix $h \in \mathcal{A}$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Then, in order to establish the desired uniform integrability, it suffices to show that

$$\limsup_{L\to\infty} E_h\left[\exp\left(\frac{\tau(\pi)}{\log L} \right) \right] < \infty. \tag{140}$$

Towards this, let us first define

$$D_{hh'} \coloneqq \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \nu^{\lambda_{h,\delta}}(\underline{d},\underline{i},a)\, D((P_h^a)^{d_a}(\cdot \mid i_a) \| (P_{h'}^a)^{d_a}(\cdot \mid i_a)). \tag{141}$$

Let

$$\tilde{n}(L) \coloneqq \frac{4\log((K-1)L)}{D_{hh'}} + K - 1, \tag{142}$$

and let

$$u(L) \coloneqq \exp\left(\frac{1+\tilde{n}(L)}{\log L} \right). \tag{143}$$

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

$$\begin{aligned}
\limsup_{L\to\infty} E_h\left[\exp\left(\frac{\tau(\pi)}{\log L} \right) \right] &= \limsup_{L\to\infty} \int_0^\infty P_h\left(\frac{\tau(\pi)}{\log L} > \log x \right) dx \\
&\leq \limsup_{L\to\infty} \int_0^\infty P_h\Big(\tau(\pi_h^\star) \geq \lceil (\log x)(\log L) \rceil \Big)\, dx
\end{aligned}$$

$$\overset{(a)}{\leq} \limsup_{L\to\infty} \left\{ u(L) + \int_{u(L)}^{\infty} P_h\Big(\tau(\pi_h^\star) \geq \lceil (\log x)(\log L) \rceil\Big)\, dx \right\}$$
$$= \exp\left(\frac{4}{D_{hh'}}\right) + \limsup_{L\to\infty} \sum_{n\geq \tilde{n}(L)} \exp\left(\frac{n+1}{\log L}\right) P_h(M_h(n) < \log((K-1)L)), \tag{144}$$

where (a) above follows by upper bounding the probability term by 1 for all $x \leq u(L)$. In Lemma 9, we show that the probability term in (144) has an exponential upper bound. It then follows that this exponential upper bound results in the finiteness of the right-hand side of (144), thus completing the proof of the Proposition. ■

APPENDIX F
AN EXPONENTIAL UPPER BOUND FOR $P_h(M_h(n) < \log((K-1)L))$

We now demonstrate the stated exponential upper bound used in (144).

Lemma 9. *Fix $\delta > 0$ and $h \in \mathcal{A}$, and suppose that $\mathcal{H}_h$ is the true hypothesis. There exist constants $B > 0$ and $0 < \theta < \infty$ independent of L such for all $n \geq \tilde{n}(L)$,*

$$P_h(M_h(n) < \log((K-1)L)) \leq Be^{-n\theta}. \tag{145}$$

Proof of Lemma 9: Since

$$P_h(M_h(n) < \log((K-1)L)) = P_h\left(\min_{h'\neq h} Z_{hh'}(n) < \log((K-1)L)\right)$$
$$\leq \sum_{h'\neq h} P_h\left(Z_{hh'}(n) < \log((K-1)L)\right); \tag{146}$$

the last line above follows from the union bound. In order to prove the Lemma, it suffices to show that each term inside the summation in (146) is exponentially bounded. Going further, we drop the superscript π in $P_h(\cdot)$ for ease of notation.

Fix $h' \neq h$. Recall that under the hypothesis $\mathcal{H}_h$, the transition probability matrix of arm h is P_1, while that of arm h' is P_2, where $P_2 \neq P_1$. The latter condition of $P_2 \neq P_1$ implies that there exists $i^* \in \mathcal{S}$ such that $P_1(\cdot|i^*) \neq P_2(\cdot|i^*)$. Equivalently, we have

$$D(P_1(\cdot|i^*)\|P_2(\cdot|i^*)) > 0, \quad D(P_2(\cdot|i^*)\|P_1(\cdot|i^*)) > 0.$$

Going further, let us fix an arbitrary $(\underline{d}^*, \underline{i}^*) \in \mathbb{S}$ such that $d_h^* = 1$ and $i_h^* = i^*$, where i^* is as defined above.

For $n \geq K$, let

$$\Delta Z_{hh'}(n) \coloneqq Z_{hh'}(n) - Z_{hh'}(n-1)$$
$$= \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \sum_{j\in\mathcal{S}} \mathbb{I}_{\{\underline{d}(n)=\underline{d}, \underline{i}(n)=\underline{i}, A_n=a, X_n^a=j\}} \log \frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)} \tag{147}$$

denote the increment of the log-likelihood process of all the intended arm pulls, actual arm pulls and observations under hypothesis $\mathcal{H}_h$ with respect to those under hypothesis $\mathcal{H}_{h'}$; note that $\Delta Z_{h'h}(n) = -\Delta Z_{hh'}(n)$. We then have the following key property satisfied by $\Delta Z_{h'h}(n)$.

Lemma 10. *For any $(\underline{d}, \underline{i}) \in \mathbb{S}$, $a \in \mathcal{A}$ and $0 < s < 1$, we have*

$$E_h\left[e^{s\Delta Z_{h'h}(n)} \middle| A_n = a, \underline{d}(n) = \underline{d}, \underline{i}(n) = \underline{i}\right] \leq 1 \quad \forall n, \tag{148}$$

with strict inequality in (148) *if $(\underline{d}, \underline{i}) = (\underline{d}^*, \underline{i}^*)$ and $a = h$.*

Proof of Lemma 10: Note that

$$\begin{aligned}
E_h\left[e^{s\Delta Z_{h'h}(n)}\middle| A_n=h,\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i}\right] &= \sum_{j\in\mathcal{S}}\left(\frac{(P^a_{h'})^{d_a}(j|i_a)}{(P^a_h)^{d_a}(j|i_a)}\right)^s P_h(X^h_n=j|A_n=a,\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i})\\
&= \sum_{j\in\mathcal{S}}\left(\frac{(P^a_{h'})^{d_a}(j|i_a)}{(P^a_h)^{d_a}(j|i_a)}\right)^s (P^a_h)^{d_a}(j|i_a)\\
&= \sum_{j\in\mathcal{S}}((P^a_h)^{d_a}(j|i_a))^{1-s}\ ((P^a_{h'})^{d_a}(j|i_a))^s\\
&\overset{(a)}{\leq} \left(\sum_{j\in\mathcal{S}}(P^a_h)^{d_a}(j|i_a)\right)^{1-s}\cdot\left(\sum_{j\in\mathcal{S}}(P^a_{h'})^{d_a}(j|i_a)\right)^s\\
&= 1, \qquad (149)
\end{aligned}$$

where (a) above is due to Hölder's inequality, and the last line follows from the fact that $(P^a_h)^{d_a}(\cdot|i_a)$ and $(P^a_{h'})^{d_a}(\cdot|i_a)$ are probability distributions on $\mathcal{S}$. When $(\underline{d},\underline{i})=(\underline{d}^*,\underline{i}^*)$ and $a=h$, the inequality in (a) is a strict inequality since $(P^a_h)^{d_a}(\cdot|i_a)=P_1(\cdot|i^*)$ and $(P^a_{h'})^{d_a}(\cdot|i_a)=P_2(\cdot|i^*)$, and since by the definition of i^*, $P_1(\cdot|i^*)\neq P_2(\cdot|i^*)$. ■

As an immediate consequence of Lemma 10, we have the following result.

Lemma 11. *For any $(\underline{d},\underline{i})\in\mathbb{S}$, $a\in\mathcal{A}$ and $0<s<1$, we have*

$$E_h\left[e^{s\Delta Z_{h'h}(n)}\mid\mathcal{F}_{n-1}\right]\mathbb{I}_{\{\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i}\}}\leq 1\quad\forall n\quad\textit{almost surely}, \qquad (150)$$

with strict inequality in (148) *if $(\underline{d},\underline{i})=(\underline{d}^*,\underline{i}^*)$ and $a=h$.*

Proof of Lemma 11: We have, almost surely,

$$\begin{aligned}
&E_h\left[e^{s\Delta Z_{h'h}(n)}\mid\mathcal{F}_{n-1}\right]\mathbb{I}_{\{\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i}\}} = E_h\left[e^{s\Delta Z_{h'h}(n)}\mid\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},\mathcal{F}_{n-1}\right]\\
&=\sum_{a=1}^K P(A_n=a\mid\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},\mathcal{F}_{n-1})\cdot E_h\left[e^{s\Delta Z_{h'h}(n)}\mid A_n=a,\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},\mathcal{F}_{n-1}\right]\\
&\overset{(a)}{=} P(A_n=h\mid\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},\mathcal{F}_{n-1})\cdot E_h\left[e^{s\Delta Z_{h'h}(n)}\mid A_n=h,\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i}\right]\\
&\qquad+\sum_{a\neq h}P(A_n=a\mid\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},\mathcal{F}_{n-1})\cdot E_h\left[e^{s\Delta Z_{h'h}(n)}\mid A_n=a,\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i}\right]\\
&\overset{(b)}{\leq} P(A_n=h\mid\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},\mathcal{F}_{n-1})\cdot E_h\left[e^{s\Delta Z_{h'h}(n)}\mid A_n=h,\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i}\right]\\
&\qquad+(1-P(A_n=h\mid\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},\mathcal{F}_{n-1}))\\
&= P(A_n=h\mid\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},\mathcal{F}_{n-1})\cdot\left(E_h\left[e^{s\Delta Z_{h'h}(n)}\mid A_n=h,\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i}\right]-1\right)+1\\
&\overset{(c)}{\leq}\frac{\eta}{K}\left(E_h\left[e^{s\Delta Z_{h'h}(n)}\mid A_n=h,\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i}\right]-1\right)+1 \qquad (151)\\
&\overset{(d)}{\leq} 1, \qquad (152)
\end{aligned}$$

where (a) above follows by noting that

$$E_h\left[e^{s\Delta Z_{h'h}(n)}\mid A_n=a,\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},\mathcal{F}_{n-1}\right]=E_h\left[e^{s\Delta Z_{h'h}(n)}\mid A_n=a,\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i}\right],$$

(b) uses the result of Lemma 10, (c) follows from the fact that for any $n\geq K$, under the policy $\pi^\star(L,\delta)$,

$$P(A_n=h\mid\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},\mathcal{F}_{n-1})=\frac{\eta}{K}+(1-\eta)\,\lambda_{\theta(n),\delta}(h|\underline{d},\underline{i})$$

$$\geq \frac{\eta}{K},$$

and (d) is straightforward. Clearly, the inequalities in (b), (c) and (d) above are strict when $(\underline{d},\underline{i}) = (\underline{d}^*,\underline{i}^*)$ and $a = h$. ■

Going further, let c denote the constant on the right-hand side of (151) when $(\underline{d},\underline{i}) = (\underline{d}^*,\underline{i}^*)$. From the arguments above, we have $c < 1$. Then,

$$\begin{aligned}
&E_h\left[e^{s\Delta Z_{h'h}(n)} \mid \mathcal{F}_{n-1}\right] \\
&= \sum_{(\underline{d},\underline{i})\in\mathbb{S}} E_h\left[e^{s\Delta Z_{h'h}(n)} \mid \mathcal{F}_{n-1}\right] \cdot \mathbb{I}_{\{\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i}\}} \\
&= c\,\mathbb{I}_{\{\underline{d}(n)=\underline{d}^*,\underline{i}(n)=\underline{i}^*\}} + \sum_{(\underline{d},\underline{i})\neq(\underline{d}^*,\underline{i}^*)} E_h\left[e^{s\Delta Z_{h'h}(n)} \mid \underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},\mathcal{F}_{n-1}\right] \cdot \mathbb{I}_{\{\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i}\}} \\
&= \begin{cases} c, & \underline{d}(n)=\underline{d}^*,\underline{i}(n)=\underline{i}^*, \\ \leq 1, & \text{otherwise.} \end{cases}
\end{aligned} \tag{153}$$

The above set of inequalities immediately lead us to the following important result.

Lemma 12. *For $0 < s < 1$,*

$$E_h\left[e^{sZ_{h'h}(n)}\right] \leq B_1\, e^{-\theta_1 n}, \tag{154}$$

where $B_1 > 0$ and $\theta_1 > 0$ are constants which depend on h, h' and s.

Proof of Lemma 12: We have

$$\begin{aligned}
E_h\left[e^{sZ_{h'h}(n)}\right] &= E_h[e^{sZ_{h'h}(n-1)}\, E_h[e^{s\Delta Z_{h'h}(n)} \mid \mathcal{F}_{n-1}]] \\
&\overset{(a)}{\leq} E_h\left[c^{N(n,\underline{d}^*,\underline{i}^*)}\right] \\
&\overset{(b)}{=} E_h\left[c^{N(n,\underline{d}^*,\underline{i}^*)}\ ;\ N(n,\underline{d}^*,\underline{i}^*) > \frac{n\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*)}{2}\right] + E_h\left[c^{N(n,\underline{d}^*,\underline{i}^*)}\ ;\ N(n,\underline{d}^*,\underline{i}^*) \leq \frac{n\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*)}{2}\right] \\
&\leq c^{n\frac{\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*)}{2}} + P_h\left(N(n,\underline{d}^*,\underline{i}^*) \leq \frac{n\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*)}{2}\right).
\end{aligned} \tag{155}$$

In the above set of equations, (a) follows from by repeatedly applying (153), the notation $E[X;A]$ in (b) stands for $E[X\,\mathbb{I}_A]$, and the last line follows by noting that $c^{n\frac{\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*)}{2}} \leq 1$ almost surely. We now note that $\{N(n,\underline{d}^*,\underline{i}^*) - N(K,\underline{d}^*,\underline{i}^*) : n \geq K\}$ is a bounded martingale. Using the Azuma-Hoeffding inequality, we then have

$$\begin{aligned}
P_h\left(N(n,\underline{d}^*,\underline{i}^*) \leq \frac{n\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*)}{2}\right) &= P_h\left(N(n,\underline{d}^*,\underline{i}^*) - N(K,\underline{d}^*,\underline{i}^*) \leq \frac{n\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*)}{2} - N(K,\underline{d}^*,\underline{i}^*)\right) \\
&\leq P_h\left(N(n,\underline{d}^*,\underline{i}^*) - N(K,\underline{d}^*,\underline{i}^*) \leq \frac{n\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*)}{2}\right) \\
&\leq \exp\left(-\frac{n(\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*))^2}{8}\right).
\end{aligned} \tag{156}$$

Plugging (156) back in (155), and noting that c is a function of s, we arrive at (154). ■

As a consequence of Lemma 12, we have the following result.

Lemma 13. *Fix an arbitrary $h \in \mathcal{A}$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Consider the non-stopping version of the policy $\pi = \pi^\star(L,\delta)$. There exist constants C_R and $\gamma > 0$ such that*

$$P_h\left(\min_{h'\neq h} Z_{hh'}(n) < R\right) \leq C_R\, e^{-\gamma n}. \tag{157}$$

In (157)*, C_R is independent of h but γ depends on h.*

Proof of Lemma 13: Observe that

$$\begin{aligned}
P_h\left(\min_{h'\neq h} Z_{hh'}(n) < R\right) &= P_h\left(\max_{h'\neq h} Z_{h'h}(n) > -R\right) \\
&\leq \sum_{h'\neq h} P_h\left(Z_{h'h}(n) > -R\right) \\
&= \sum_{h'\neq h} P_h\left(sZ_{h'h}(n) > -sR\right) \quad \forall\ 0 < s < 1 \\
&\overset{(a)}{\leq} \sum_{h'\neq h} e^{sR}\ E_h\left[e^{sZ_{h'h}(n)}\right] \\
&\overset{(b)}{\leq} e^{sR} \sum_{h'\neq h} B_1\ e^{-\theta n} \\
&\leq e^{sR}\cdot(K-1)\cdot \max_{h'\neq h} B_1\, e^{-\theta n} \\
&\leq C_R\ e^{-\gamma n}, \qquad (158)
\end{aligned}$$

where $\max\limits_{h'\neq h} B_1\, e^{-\theta n} = e^{-\gamma}$ and $C_R = Ke^{sR}$. In the above set of equations, (a) is due to Chernoff's bound for $0 < s < 1$, and (b) is due to Lemma 12. ■

From (124), we know that under the non-stopping version of the policy $\pi^\star(L,\delta)$, the guess of the odd arm $\theta(n)$ eventually settles at h with probability 1 under the hypothesis $\mathcal{H}_h$. Indeed, we now show using Lemma 13 that something stronger holds. Towards this, fix $h \in \mathcal{A}$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Let

$$T_h \coloneqq \inf\{n : \theta(n') = h \text{ for all } n' \geq n\}. \qquad (159)$$

We have the following result for T_h.

Lemma 14. *Fix an arbitrary $h \in \mathcal{A}$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Consider the non-stopping version of the policy $\pi^\star(L,\delta)$. There exist constants $C > 0$ and $b > 0$, both finite and possibly depending on h, such that*

$$P_h\left(T_h > n\right) \leq C\, e^{-bn}. \qquad (160)$$

Proof of Lemma 14: We have

$$\begin{aligned}
P_h\left(T_h > n\right) &\leq P_h\left(\exists\ n' \geq n \text{ such that } \theta(n') \neq h\right) \\
&\leq \sum_{n'\geq n} P_h(\theta(n') \neq h) \\
&= \sum_{n'\geq n} P_h\left(\exists\ h' \neq h \text{ such that } \theta(n') = h'\right) \\
&\leq \sum_{n'\geq n} P_h\left(M_{h'}(n') > \max_{h''\neq h'} M_{h''}(n)\right) \\
&\leq \sum_{n'\geq n} P_h\left(M_h(n') - M_{h'}(n') < 0\right). \qquad (161)
\end{aligned}$$

We now note that, almost surely,

$$\begin{aligned}
M_h(n') - M_{h'}(n') &= M_h(n') - \min_{h''\neq h'} Z_{h'h''}(n') \\
&\geq M_h(n') - Z_{h'h}(n') \\
&= M_h(n') + Z_{hh'}(n')
\end{aligned}$$

$$\geq 2 \min_{h' \neq h} Z_{hh'}(n'). \tag{162}$$

Using (162) in (161), we get

$$P_h\left(T_h > n\right) \leq \sum_{n' \geq n} P_h\left(\min_{h' \neq h} Z_{hh'}(n) < 0\right). \tag{163}$$

The result now follows from Lemma 13. ■

We now use the results presented above to derive the desired exponential upper bound for each term of the summation in (146) to finish the proof of Lemma 9. Note that for any $\epsilon' > 0$, we have

$$\begin{aligned}
&P_h\left(Z_{hh'}(n) < \log((K-1)L)\right) \\
&= P_h\left(\sum_{k=K}^{n} \Delta Z_{hh'}(k) < \log((K-1)L)\right) \\
&= P_h\Bigg(\sum_{k=K}^{n} (\Delta Z_{hh'}(k) - E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] + \epsilon') \\
&\qquad\qquad + \sum_{k=K}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \epsilon') \\
&\qquad\qquad\qquad + (n-K+1)\left(D_{hh'} - 2\epsilon'\right) < \log((K-1)L)\Bigg) \\
&\leq P_h\Bigg(\sum_{k=K}^{n} (\Delta Z_{hh'}(k) - E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] + \epsilon') < 0\Bigg) + P_h\Bigg(\sum_{k=K}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \epsilon') < 0\Bigg) \\
&\qquad\qquad + P_h\Big((n-K+1)\left(D_{hh'} - 2\epsilon'\right) < \log((K-1)L)\Big).
\end{aligned} \tag{164}$$

We first choose ϵ' such that

$$(n-K+1)\left(D_{hh'} - 2\epsilon'\right) \geq \log((K-1)L) \quad \forall n \geq \tilde{n}(L).$$

In particular, $\epsilon' = D_{hh'}/4$ works. Let us fix this ϵ' for the rest of the proof, and note that this choice of ϵ' ensures that the third probability term in (164) is equal to 0. We now focus on the first probability term in (164), and note that each term inside the summation has strictly positive mean. Thus, from Chernoff's bounding technique [22, Lemma 2], we get that there exists $b(\epsilon')$ such that

$$P_h\Bigg(\sum_{k=K}^{n} (\Delta Z_{hh'}(k) - E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] + \epsilon') < 0\Bigg) \leq e^{-(n-K+1)\, b(\epsilon')}. \tag{165}$$

It thus remains to show that the second probability term in (164) is bounded above exponentially. To do so, we use the proof technique of Vaidhiyan et al. [1, pp. 4793-4794] and adapt it to our setting of restless arms.

Let

$$\begin{aligned}
\tilde{C} &\coloneqq \inf_{(\underline{d},\underline{i}) \in \mathbb{S},\ a \in \mathcal{A}} E_h[\Delta Z_{hh'}(n) \mid A_n = a, \underline{d}(n) = \underline{d}, \underline{i}(n) = \underline{i}] - D_{hh'} \\
&= \inf_{(\underline{d},\underline{i}) \in \mathbb{S},\ a \in \mathcal{A}} D((P_h^a)^{d_a}(\cdot \mid i_a) \| (P_{h'}^a)^{d_a}(\cdot \mid i_a)) - D_{hh'}.
\end{aligned} \tag{166}$$

Note that $\tilde{C} \leq 0$ by the definition of $D_{hh'}$. Choose ϵ'' such that

$$\tilde{\epsilon} \coloneqq \epsilon' + \epsilon'' \tilde{C} > 0;$$

here, $\epsilon' = D_{hh'}/4$ as chosen earlier. We may then write the second probability in (164) as follows:

$$P_h\Bigg(\sum_{k=K}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \epsilon') < 0\Bigg)$$

$$= P_h\Bigg(\sum_{k=K}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \epsilon') < 0,\ T_h \leq n\epsilon'' \Bigg) + P_h\Bigg(\sum_{k=K}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \epsilon') < 0,\ T_h > n\epsilon'' \Bigg)$$

$$\leq P_h\Bigg(\sum_{k=K}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \epsilon') < 0,\ T_h \leq n\epsilon'' \Bigg) + P_h\Bigg(T_h > n\epsilon'' \Bigg). \tag{167}$$

From Lemma 14, the second probability term in (167) is bounded above exponentially. The first probability term in (167) may be upper bounded as

$$\begin{aligned}
&P_h\Bigg(\sum_{k=K}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \epsilon') < 0,\ T_h \leq n\epsilon'' \Bigg) \\
&= P_h\Bigg(\sum_{k=K}^{\lfloor n\epsilon'' \rfloor} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \epsilon') + \sum_{k=\lfloor n\epsilon'' \rfloor+1}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \epsilon') < 0,\ T_h \leq n\epsilon'' \Bigg) \\
&\overset{(a)}{\leq} P_h\Bigg((\lfloor n\epsilon'' \rfloor - K + 1)(\tilde{C} + \epsilon') + \sum_{k=\lfloor n\epsilon'' \rfloor+1}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \epsilon') < 0,\ T_h \leq n\epsilon'' \Bigg) \\
&= P_h\Bigg((\lfloor n\epsilon'' \rfloor - K + 1)(\tilde{C} + \epsilon') + (n - \lfloor n\epsilon'' \rfloor)(\epsilon' - \tilde{\epsilon}) + \sum_{k=\lfloor n\epsilon'' \rfloor+1}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \tilde{\epsilon}) < 0,\ T_h \leq n\epsilon'' \Bigg) \\
&\overset{(b)}{\leq} P_h\Bigg(\lfloor n\epsilon'' \rfloor(\epsilon''\tilde{C} + \epsilon') - (K-1)\epsilon' + \sum_{k=\lfloor n\epsilon'' \rfloor+1}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \tilde{\epsilon}) < 0,\ T_h \leq n\epsilon'' \Bigg) \\
&\overset{(c)}{\leq} P_h\Bigg(\sum_{k=\lfloor n\epsilon'' \rfloor+1}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \tilde{\epsilon}) < 0,\ T_h \leq n\epsilon'' \Bigg) \\
&= \tilde{P}_h\Bigg(\sum_{k=\lfloor n\epsilon'' \rfloor+1}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \tilde{\epsilon}) < 0 \Bigg),
\end{aligned} \tag{168}$$

where in writing (a), we use the fact that for each $k \geq K$, we have $E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] \geq \tilde{C}$, (b) follows by noting that

$$\begin{aligned}
&(\lfloor n\epsilon'' \rfloor - K + 1)(\tilde{C} + \epsilon') + (n - \lfloor n\epsilon'' \rfloor)(\epsilon' - \tilde{\epsilon}) \\
&= (\lfloor n\epsilon'' \rfloor - K + 1)(\tilde{C} + \epsilon') - (n - \lfloor n\epsilon'' \rfloor)\,\epsilon''\tilde{C} \\
&= \lfloor n\epsilon'' \rfloor(\epsilon' + \epsilon''\tilde{C}) + \tilde{C}(\lfloor n\epsilon'' \rfloor - n\epsilon'') - (K-1)(\tilde{C} + \epsilon') \\
&\geq \lfloor n\epsilon'' \rfloor(\epsilon''\tilde{C} + \epsilon') - (K-1)\epsilon'
\end{aligned} \tag{169}$$

since $\tilde{C} \leq 0$, (c) and the equality in (168) hold for all n such that $\lfloor n\epsilon'' \rfloor(\epsilon''\tilde{C} + \epsilon') - (K-1)\epsilon' \geq 0$, and in (168), $\tilde{P}_h$ is a new probability measure under which at each time instant, an arm is selected according to the policy $\pi^\star(L, \delta)$ but assuming that the guess of the odd arm $\theta(k) = h$ for all k.

We now note that under the measure $\tilde{P}_h$,

$$\tilde{E}_h[E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}]] = \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \tilde{P}_h(\underline{d}(k) = \underline{d}, \underline{i}(k) = \underline{i}) \left(\frac{\eta}{K} + (1-\eta)\lambda_{h,\delta}(a|\underline{d},\underline{i}) \right) D((P_h^a)^{d_a}(\cdot \mid i_a) \| (P_{h'}^a)^{d_a}(\cdot \mid i_a)), \tag{170}$$

where $\tilde{E}_h$ in (170) denotes expectation under the measure $\tilde{P}_h$. We claim that under the measure $\tilde{P}_h$, the collection $\{(\underline{d}(k), \underline{i}(k)) : k \geq \lfloor n\epsilon'' \rfloor + 1\}$ is a Markov process. Indeed, for all $k \geq \lfloor n\epsilon'' \rfloor + 1$,

$$\tilde{P}_h(\underline{d}(k+1) = \underline{d}', \underline{i}(k+1) = \underline{i}' \mid (\underline{d}(t), \underline{i}(t)),\ \lfloor n\epsilon'' \rfloor + 1 \leq t \leq k)$$

$$= \begin{cases} \left(\frac{\eta}{K} + (1-\eta)\,\lambda_{h,\delta}(a|\underline{d}(k),\underline{i}(k))\right)\,(P_h^a)^{d_a(k)}(i'_a|i_a(k)), & \text{if } d'_a = 1 \text{ and } d'_b = d_b(k)+1 \text{ for all } b \neq a, \\ & i'_b = i_b(k) \text{ for all } b \neq a, \\ 0, & \text{otherwise.} \end{cases} \tag{171}$$

Let us fix $\underline{d}' = (K, K-1, \ldots, 1)$ and $\underline{i}' = (1, \ldots, 1)$, where we assume without loss of generality that $1 \in \mathcal{S}$. From [20, Proposition 1.7], we know that there exists an integer M sufficiently large such that each entry of the transition probability matrices P_1^M and P_2^M is strictly positive. We now use this fact to demonstrate that for any $(\underline{d},\underline{i}) \in \mathbb{S}$, if the process $\{(\underline{d}(t),\underline{i}(t)) : t \geq K\}$ starts in the state $(\underline{d},\underline{i})$ at some time $T_0 \geq \lfloor n\epsilon'' \rfloor + 1$, then it has a strictly positive probability of being in the state $(\underline{d}',\underline{i}')$ at time $t = T_0 + M + d'_1 - d'_K$. Indeed, following the arguments in the proof of irreducibility in Appendix A, this probability may be lower bounded as

$$\begin{aligned} &\tilde{P}_h(\underline{d}(T_0+M+d'_1-d'_K) = \underline{d}', \underline{i}(T_0+M+d'_1-d'_K) = \underline{i}' \mid \underline{d}(K) = \underline{d}, \underline{i}(K) = \underline{i}) \\ &\geq \left(\frac{\eta}{K}\right)^{K-1} \cdot \left[\prod_{a=1}^{K} (P_h^a)^{M+d'_1-d'_a}(1|i_a)\right] \cdot \left[\prod_{a=1}^{K-1} \prod_{t=T_0+M+d'_1-d'_{a+1}+1}^{T_0+M+d'_1-d'_a} \frac{\eta}{K}\right]. \end{aligned} \tag{172}$$

Denoting the right-hand side of (172) by α, and noting that $\alpha > 0$, we have that

$$(\tilde{P}_h)^{M+d'_1-d'_K}((\underline{d}'',\underline{i}'') \mid \underline{d},\underline{i}) \geq \alpha\, \mathbb{I}_{\{(\underline{d}'',\underline{i}'')=(\underline{d}',\underline{i}')\}} \quad \text{for all } (\underline{d},\underline{i}),\ (\underline{d}'',\underline{i}'') \in \mathbb{S}. \tag{173}$$

The condition in (173) is referred to as the "Doeblin's minorisation condition" (see [23, Eq. (5)]). Noting that (a) the Markov process $\{(\underline{d}(k),\underline{i}(k)) : k \geq \lfloor n\epsilon'' \rfloor + 1\}$ is ergodic under the measure $\tilde{P}_h$ with $\mu^{\lambda_{h,\delta}}$ as the unique stationary distribution, (b) (173) holds, and (c) the increment $\Delta Z_{hh'}(k)$ is almost surely bounded for each k as demonstrated in (136), we apply [23, Theorem 1] to deduce that the second probability term in (164) is bounded above exponentially. This completes the proof of the Lemma. ■

REFERENCES

[1] N. K. Vaidhiyan, S. Arun, and R. Sundaresan, "Neural dissimilarity indices that predict oddball detection in behaviour," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 4778–4796, 2017.

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

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

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

[5] A. P. Sripati and C. R. Olson, "Global image dissimilarity in macaque inferotemporal cortex predicts human visual search efficiency," *Journal of Neuroscience*, vol. 30, no. 4, pp. 1258–1269, 2010.

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

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

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

[9] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1902–1916, 2012.

[10] R. Ortner, D. Ryabko, P. Auer, and R. Munos, "Regret bounds for restless markov bandits," in *International Conference on Algorithmic Learning Theory*. Springer, 2012, pp. 214–228.

[11] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[12] S. Grünewälder and A. Khaleghi, "Approximations of the restless bandit problem," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 514–550, 2019.

[13] S. Bubeck, R. Munos, and G. Stoltz, "Pure Exploration in Finitely-armed and Continuous-armed Bandits," *Theor. Comput. Sci.*, vol. 412, no. 19, pp. 1832–1852, Apr. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.tcs.2010.12.059

[14] 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.

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

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

[17] K. Avrachenkov and V. S. Borkar, "Whittle index based q-learning for restless bandits with average reward," 2020. [Online]. Available: https://arxiv.org/abs/2004.14427

[18] P. Milgrom and I. Segal, "Envelope theorems for arbitrary choice sets," *Econometrica*, vol. 70, no. 2, pp. 583–601, 2002.

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

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

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

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

[23] I. Kontoyiannis, L. A. Lastras-Montaño, and S. P. Meyn, "Relative entropy and exponential deviation bounds for general markov chains," in *Proceedings. International Symposium on Information Theory, 2005. ISIT 2005.* IEEE, 2005, pp. 1563–1567.