

Learning to Detect an Odd Restless Markov Arm with a Trembling Hand

P. N. Karthik and Rajesh Sundaresan

Abstract

This paper studies the problem of finding an anomalous arm in a multi-armed bandit when (a) each arm is a finite-state Markov process, and (b) the arms are restless. Here, anomaly means that the transition probability matrix (TPM) of one of the arms (the odd arm) is different from the common TPM of each of the non-odd arms. The TPMs are unknown to a decision entity that wishes to find the index of the odd arm as quickly as possible, subject to an upper bound on the error probability. We derive a problem instance specific asymptotic lower bound on the expected time required to find the odd arm index, where the asymptotics is as the error probability vanishes. Further, we devise a policy based on the principle of certainty equivalence, and demonstrate that under a continuous selection assumption and a certain regularity assumption on the TPMs, the policy achieves the lower bound arbitrarily closely. Thus, while the lower bound is shown for all problem instances, the upper bound is shown only for those problem instances satisfying the regularity assumption. Our achievability analysis is based on resolving the identifiability problem in the context of a certain countable-state controlled Markov process.

Index Terms

Odd arm identification, restless multi-armed bandits, controlled Markov process, certainty equivalence, identifiability.

I. INTRODUCTION

Consider a multi-armed bandit in which each arm is a time-homogeneous and ergodic discrete-time Markov process taking values in a common finite state space. Assume that the Markov process of each arm is independent of those of the other arms. Suppose that the state evolution on one of the arms (the *odd* arm) is governed by a transition probability matrix (TPM) P_1 and that on each of the non-odd arms is governed by $P_2 \neq P_1$. A decision entity that knows neither P_1 nor P_2 wishes to find the index of the odd arm as quickly as possible subject to an upper bound on the error probability. Clearly, the smaller the error probability, the longer the decision entity will have to wait before declaring the odd arm index. Our goal is to characterise the asymptotic growth and growth rate of the expected time required to find the odd arm index as the error probability vanishes.

To arrive at a decision of the odd arm index, the decision entity samples the arms sequentially, one at each discrete time instant $t \in \{0, 1, 2, \dots\}$. Following [1], we assume that the decision entity has a *trembling hand*. This means that for some fixed $\eta > 0$, the decision entity samples the intended arm with probability $1 - \eta$, but with probability η , the decision entity samples a uniformly randomly chosen arm. Here, η is known as the trembling hand parameter. See [1] for an example where human subjects exhibit a trembling hand in a visual search experiment that involves searching for an oddball image in a sea of

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.

distracter images. It is likely that in such visual search experiments, the human subject scans multiple images at once before narrowing down the search to the oddball image [2], in a way different from that modelled in [1]. Nevertheless, the model in [1] captures the search dynamics in a way that makes the problem amenable to analysis. We therefore follow the model in [1].

At each time t , the decision entity observes the (noiseless) state of the pulled arm post-trembling. The Markov processes of the unobserved arms continue to evolve, thus making the arms *restless* [3]. The decision entity continues to sample the arms sequentially until it is sufficiently confident¹ of its estimate of the odd arm index, at which time it stops further sampling and declares its estimate of the odd arm index.

For the above described problem of identifying the odd arm in a restless multi-armed bandit, the paper [1] studies the case when both P_1 and P_2 are known to the decision entity beforehand. In this paper, we study the more difficult case when neither P_1 nor P_2 is known beforehand.

A. Observation ‘Delays’ and a Markov Decision Problem

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

B. Certainty Equivalence and Identifiability

When neither P_1 nor P_2 is known beforehand, as is the case in this paper, the transition probabilities of the MDP may be regarded as being parameterised by a triplet of unknowns consisting of (a) the odd arm index, (b) the TPM P_1 of the odd arm, and (c) the common TPM P_2 of each non-odd arm. Call this triplet an *arms configuration*. Because the true (underlying) arms configuration is not known beforehand, it must at least partially be learnt along the way. A commonly used approach to learn the true parameter governing the transition probabilities of an MDP is *certainty equivalence*. The idea behind this approach is to (a) maintain an estimate of the parameter at each time t , and (b) take an action at time t supposing that the estimated value is indeed the true parameter value. The key challenge in this approach is to show that the parameter estimates converge to the true parameter value as $t \rightarrow \infty$, i.e., the system is *identifiable*.

Sufficient conditions that lead to system identifiability have been proposed in the literature. In an important paper [4], Mandl demonstrated that for a finite-state MDP, when the parameter estimate at each time t is chosen so as to minimise a “contrast” function computed using all the observations and actions up to time t , the parameter estimates converge to the true parameter value as $t \rightarrow \infty$ [4, Theorem 6]. In particular, when the contrast function is the negative log-likelihood, the resulting parameter estimates are the maximum likelihood (ML) estimates. Mandl demonstrated the convergence of the ML estimates to the true parameter value under an additional condition (known as *Mandl’s identifiability condition* [4, Eq. (35)]) on the MDP transition

¹Based on the given error probability.

probabilities. However, it is not clear if Mandl’s identifiability condition is sufficient for identifiability in countable-state MDPs. In [5], the authors consider the same problem as Mandl’s, but for countable-state MDPs when Mandl’s identifiability condition is relaxed. The authors of [5] show that under some regularity assumptions on the MDP transition probabilities, certainty equivalence based on ML estimation renders the system identifiable.

In this paper, we use certainty equivalence with ML estimation as in [5] to learn the true arms configuration. Due to the presence of arm delays in the likelihood function, closed-form expressions for the ML estimates of the TPMs are not available. Nevertheless, we show that the system is identifiable under a mild regularity assumption on the TPMs.

C. Prior Works on Odd Arm Identification, Restless Arms, and Other Related Works

As mentioned earlier, the paper [1] studies the problem of odd arm identification in a restless multi-armed bandit when the TPMs of the odd arm and the non-odd arms are known beforehand. In addition to [1], prior works on odd arm identification have studied the cases when either each arm yields iid observations or each arm yields Markov observations and the arms are rested. The specifics of such works are as follows. Paper [6] studies the case when each arm yields iid Poisson observations and the Poisson rates of the odd arm and the non-odd arms are known beforehand. A study of the case of unknown Poisson rates appears in [7]. Extension of the results in [7] to the case when each arm yields iid observations coming from a generic exponential family appears in [8]. The setting when each arm yields Markov observations and the arms are rested has been studied in [9]. The recent papers [10]–[12] study more general sequential hypothesis testing problems in multi-armed bandits of which the problem of odd arm identification is a special case. For a related problem of best arm identification instead of odd arm identification, see [13], [14].

Restless multi-armed bandits have been studied in the works of [3], [15], [16]. In contrast to these works which are based on the theme of minimising regret, our work is based on the theme of optimal stopping. In [17], a policy based on repeated sampling of the arms has been proposed to achieve a logarithmic order regret for a restless multi-armed bandit with unknown dynamics. The setting of our paper is similar to that of [17] in that the TPMs of the odd arm and the non-odd arms are not known beforehand. However, it is not clear if the policy of [17], or more generally, a policy that learns the TPMs by repeatedly sampling the arms, performs well for the optimal stopping problem studied in this paper.

We note here that the problem of identifying the odd (anomalous) arm is closely related to the problem of *anomaly detection* that has been explored sufficiently well in the literature. In works on anomaly detection, the time at which the anomaly occurs is a random variable, and the goal is to detect the anomaly in the shortest possible time after it occurs. However, in our work, it is known to the decision entity beforehand that one of the arms is anomalous.

D. Prior Works on Certainty Equivalence, Identification, and Adaptive Control of Markov Processes

The principle of certainty equivalence seems to have been rigorously explored first in the context of linear systems by Åström and Wittenmark [18] where it has been referred to as a *self-tuning regulator*. The paper by Mandl [4] applies the principle of certainty equivalence to the problem of adaptively controlling a finite-state controlled Markov process (equivalently, a finite-state MDP). Mandl demonstrated that under a regularity condition (known as Mandl’s identifiability condition [4, Eq. (35)]) on the MDP transition probabilities, certainty equivalence based on ML estimation renders the system identifiable. Doshi and Shreve [19] consider a problem similar to Mandl’s and show that under a slightly weaker condition than Mandl’s, identifiability holds for a scheme based on certainty equivalence with modified maximum likelihood estimates (estimates that nearly maximise the log-likelihood).

It is not clear if either Mandl's identifiability condition or its weaker version in [19] is sufficient for identifiability in infinite-state MDPs. Also, as remarked in [20], Mandl's condition may be too restrictive in practical applications. The authors of [20] provide an example of a linear, real-valued Markovian system for which Mandl's identifiability condition fails to hold. For this example, the authors of [20] demonstrate that the parameter estimates may not converge, and even if they do, the convergence is not necessarily to the true parameter value. This suggests that in order to show system identifiability in infinite-state MDPs, it may be necessary to impose additional regularity conditions on the MDP transition probabilities.

One such set of such regularity conditions that ensures identifiability in countable-state MDPs when Mandl's condition is relaxed may be found in [5]. We show that the regularity conditions of [5] are satisfied for the problem studied in this paper under a mild assumption on the unknown TPMs of the odd arm and the non-odd arm Markov processes. For a detailed survey of results in stochastic adaptive control of Markov chains, see [21].

E. Contributions and Key Challenges to Overcome

In this section, we bring out the contributions of this paper and highlight the key challenges that we must overcome in our work.

- 1) We derive an asymptotic lower bound on the expected time required to find the odd arm index subject to an upper bound on the error probability, where the asymptotics is as the error probability vanishes. Specifically, given an arms configuration $C = (h, P_1, P_2)$ in which h is the odd arm index, P_1 is the TPM of arm h , and $P_2 \neq P_1$ is the common TPM of each non-odd arm, we show that the lower bound is $1/R^*(h, P_1, P_2)$, where $R^*(h, P_1, P_2)$ is an arms configuration dependent (or problem instance dependent) constant and is the value of a max-min optimisation problem. See Section III, equation (12) for the exact mathematical expression for $R^*(h, P_1, P_2)$. The max-min expression in (12) consists of an outer supremum over all conditional probability distributions on the arms conditioned on the arm delays and the last observed states, and an inner infimum over all arms configurations in which the odd arm index is any $h' \neq h$.
- 2) Given an arms configuration $C = (h, P_1, P_2)$, the question of whether there exists an optimal conditional distribution that attains the outer supremum in the expression for $R^*(h, P_1, P_2)$ is still under study. Notwithstanding this, we look at conditional distributions that attain the supremum to within a factor of $1/(1 + \delta)$ for any $\delta > 0$. We refer to such conditional distributions as δ -optimal solutions for the arms configuration $C = (h, P_1, P_2)$, and note that there may be multiple such solutions in general for each arms configuration.

It is worth noting here that in the prior works [7]–[9], the outer supremum in the expression for the constant appearing in the lower bound is over all *unconditional* probability distributions on the arms which are simpler objects to deal with than conditional probability distributions. Further, in these works, this outer supremum is attained by a unique (unconditional) probability distribution on the arms.

- 3) An important property of the optimal solutions to the lower bounds in [7]–[9] is that they are continuous functions of the arms configuration. It is this continuity property that is used to show that the certainty equivalence-based arm sampling policies of these works (using ML estimation) render the system identifiable. This, in turn, is used to show that these policies achieve the respective lower bounds asymptotically. Furthermore, a close look at [7]–[9] reveals that the ML estimates in these works have closed-form expressions. This readily helps in showing the convergence of the ML estimates to their true values by using either the law of large numbers (for the case of iid observations from the arms in [7], [8]) or the ergodic theorem (for the case of Markov observations from the arms in [9]).

In the context of this paper, we note that neither the existence of unique optimal solutions to $R^*(h, P_1, P_2)$ nor the above mentioned continuity property is available. Furthermore, the presence of arm delays in the likelihood function poses difficulty in obtaining closed-form expressions for the ML estimates of the TPMs of the odd arm and the non-odd arm Markov processes. As a result, showing that the ML estimates converge to their true values is a challenging task.

- 4) Notwithstanding the difficulties highlighted in the previous point, we demonstrate that under two key assumptions, certainty equivalence based on ML estimation renders the system identifiable. The first of these assumptions is on the existence of a continuous selection of δ -optimal solutions (the analogue of continuity of optimal solutions in the prior works). The second assumption, analogous to that appearing in [20], is a mild regularity on the TPMs of the odd arm and the non-odd arm Markov processes.
- 5) Given an error probability $\epsilon > 0$ and $\delta > 0$, we construct a policy with parameters $L = 1/\epsilon$ and δ based on the principle of certainty equivalence with ML estimation and using the δ -optimal solutions. Under the assumptions mentioned in the previous point, we demonstrate that the policy stops in finite time almost surely and that the error probability of the policy at stoppage is upper bounded by ϵ . Further, we show that as $\epsilon \downarrow 0$, the policy's expected time to find the odd arm index satisfies an upper bound that is away from the lower bound only by a multiplicative factor of $(1 + \delta)^2$. Our achievability analysis relies crucially on (a) resolving the identifiability problem for the countable-state MDP arising from the arm delays and the last observed states, and (b) showing that the policy eventually samples the arms according to the δ -optimal solution for the underlying arms configuration. Finally, by letting $\delta \downarrow 0$, we show that our policy meets the lower bound.

In summary, we prove the lower bound for all arms configurations, and prove the upper bound only for those arms configurations satisfying two regularity assumptions: (a) the existence of continuous selection of δ -optimal solutions, and (b) the regularity of the TPMs.

- 6) As highlighted in point 3 above, because of the presence of arm delays in the expression for the likelihood function, it is difficult to obtain closed-form expressions for the ML estimates of the TPMs (and hence for the maximum likelihood) at any given time. This difficulty can be circumvented by repeatedly sampling the Markov process of each arm and using the successive observations from each of the arms to estimate the TPMs. Because the Markov process of each arm is ergodic, these TPM estimates will converge to their true values asymptotically.

However, unlike the policy we propose in Section IV, repeated sampling of the arms may not ensure that given $\delta > 0$ and an underlying arms configuration C , the arms are eventually sampled according to the δ -optimal solution for C , a condition that is crucial in order to meet the constant in the lower bound. Moreover, it is not clear if the desired error probability criterion can be met. While policies that sample the arms repeatedly have been proposed and shown to perform well for the problem of minimising regret [17], it is not clear if such policies perform well for optimal stopping problems such as that studied in this paper.

F. Organisation of the Paper

The rest of this paper is organised as follows. In Section II, we set up the notations and provide some preliminaries on MDPs that will be used in the remainder of the paper. In Section III, we present the asymptotic lower bound on the expected time to identify the odd arm index. In Section IV, we state the two assumptions, present a policy based on certainty equivalence with ML estimation, and demonstrate that it achieves the lower bound asymptotically. We provide some concluding remarks in Section VI. The proofs of all the results are contained in Appendices A-G.

II. NOTATIONS AND PRELIMINARIES

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

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

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

for some fixed $\eta > 0$; here, η is called the *trembling hand parameter*. Notice that A_t is chosen uniformly at random with probability η , and $A_t = B_t$ with probability $1 - \eta$. The decision entity observes A_t and therefore knows whether its hand trembled at time t . Further, the decision entity observes the (noiseless) state of the sampled arm A_t , which we denote by \bar{X}_t . Therefore, at any given time t , the decision entity has knowledge of the history $(B_0, A_0, \bar{X}_0, \dots, B_t, A_t, \bar{X}_t)$ of all the intended arm samples, the actual arm samples and the states of the actually sampled arms up to time t . We write (B^t, A^t, \bar{X}^t) to compactly represent the history $(B_0, A_0, \bar{X}_0, \dots, B_t, A_t, \bar{X}_t)$. While the decision entity observes the state of only one arm at each time instant, the Markov processes of the other arms continue to evolve (*restless* arms).

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

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

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

$$\Pi(\epsilon) := \left\{ \pi : P^\pi \left(\theta(\tau(\pi)) \neq h \mid C = (h, P_1, P_2) \right) \leq \epsilon \quad \forall C = (h, P_1, P_2), h \in \mathcal{A}, P_2 \neq P_1 \right\}. \quad (2)$$

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

Note the requirement of policies in (2) to work for all arms configurations. A careful reader may raise the question that the error probability criterion may not hold for arms configurations $C = (h, P_1, P_2)$ in which the entries of P_1 and P_2 are arbitrarily close to each other, with some entries of P_1 possibly matching with those of P_2 . However, the key here is to note that, for such arms configurations, a policy $\pi \in \Pi(\epsilon)$ might have to observe longer before it stops and declares the odd arm index with an error probability $\leq \epsilon$. Therefore, the “closer” the TPMs P_1 and P_2 are, the harder they are to distinguish, which results in larger stopping times of the policies.

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

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

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

A. Arm Delays and Last Observed States

Due to the restless nature of the arms, the decision entity has to keep a record of (a) the time elapsed since an arm was previously selected (the arm's *delay*), and (b) the state of the arm at its previous selection time instant (the arm's *last observed state*). As noted in [1], arm delays and the last observed states are striking features of the setting of restless arms. When each arm yields independent and identically distributed (iid) observations, or when each arm yields Markov observations and the arms are rested, the notion of arm delays is superfluous.

Following the notations in [1], let $d_a(t)$ and $i_a(t)$ respectively denote the delay and the last observed state of arm $a \in \mathcal{A}$ at time t . Let $\underline{d}(t) = (d_a(t) : a \in \mathcal{A})$ and $\underline{i}(t) = (i_a(t) : a \in \mathcal{A})$ denote the vectors of arm delays and last observed states at time t . The arm delays and the last observed states make sense when each arm is sampled at least once, and shall therefore be defined for $t \geq K$ keeping in mind that $A_0 = 1, \dots, A_{K-1} = K$ under every policy. For $t = K$, we set $\underline{d}(K) = (K, K-1, \dots, 1)$. This means that with reference to time $t = K$, arm 1 was sampled K time instants earlier (i.e., at $t = 0$), arm 2 was sampled $K-1$ time instants earlier (i.e., at $t = 1$) and so on. The rule for updating $(\underline{d}(t), \underline{i}(t))$, based on the value of A_t , is straightforward and is as follows: if $A_t = a$, then

$$d_{\tilde{a}}(t+1) = \begin{cases} d_{\tilde{a}}(t) + 1, & \tilde{a} \neq a, \\ 1, & \tilde{a} = a', \end{cases} \quad i_{\tilde{a}}(t+1) = \begin{cases} i_{\tilde{a}}(t), & \tilde{a} \neq a, \\ \bar{X}_t, & \tilde{a} = a', \end{cases} \quad (4)$$

where, to recall, \bar{X}_t is the state of the arm $A_t = a$ at time t . Therefore, the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ takes values in a subset, say \mathbb{S} , of $\{1, 2, \dots\}^K \times \mathcal{S}^K$. The set \mathbb{S} is countably infinite and includes among many others the constraint that, for each $t \geq K$, exactly one component of the vector $\underline{d}(t)$ is equal to 1 and all other components are strictly greater than 1.

Therefore, it follows that under any policy π , the decision entity first samples each of the K arms in such a way that $A_0 = 1$, $A_1 = 2$ and so on until $A_{K-1} = K$. Then, for all $K \leq t < \tau(\pi)$, based on (a) the history $\{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}$ of arm delays and last observed states, and (b) the history $\{B_s : K \leq s < t\}$ of all the intended arm samples, the decision maker chooses to sample the arm B_t . Notice that this is equivalent to choosing B_t based the history $(B^{t-1}, A^{t-1}, \bar{X}^{t-1})$. Due to the trembling hand, the decision entity observes that arm A_t is instead pulled at time t . Subsequently, the decision entity observes the state \bar{X}_t of arm A_t , and updates $(\underline{d}(t), \underline{i}(t))$ to $(\underline{d}(t+1), \underline{i}(t+1))$ based on the update rule in (4). At time $t = \tau(\pi)$, the decision entity announces its estimate $\theta(\tau(\pi))$ of the odd arm index.

B. Controlled Markov Process and the Associated Markov Decision Problem

From [1, Section II.C], we know that $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is a *controlled Markov process* with $(\underline{d}(t), \underline{i}(t))$ regarded as the state at time t and B_t regarded as the control at time t . That is, we are in the setting of a Markov decision problem (MDP)

whose state space is \mathbb{S} , action space is \mathcal{A} , and the transition probabilities under an arms configuration C are as follows: for all $(\underline{d}', \underline{i}'), (\underline{d}, \underline{i}) \in \mathbb{S}$ and $b \in \mathcal{A}$,

$$P^\pi(\underline{d}(t+1) = \underline{d}', \underline{i}(t+1) = \underline{i}' \mid \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, B_t = b, C) = \begin{cases} \left(\frac{\eta}{K} + (1-\eta) \mathbb{I}_{\{a=b\}}\right) (P_C^a)^{d_a}(i'_a|i_a), & \text{if } d'_a = 1 \text{ and } d'_a = d_a + 1 \text{ for all } \tilde{a} \neq a, \\ i'_a = i_a \text{ for all } \tilde{a} \neq a, & \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

In (5), P_C^a is the TPM of arm a under the arms configuration C . For instance, if $C = (h, P_1, P_2)$, then

$$P_C^a = \begin{cases} P_1, & a = h, \\ P_2, & a \neq h. \end{cases} \quad (6)$$

Also, d'_a and i'_a denote the component corresponding to arm a in the vectors \underline{d}' and \underline{i}' respectively. Similarly, $d_{\tilde{a}}$ and $i_{\tilde{a}}$ denote the component corresponding to arm \tilde{a} in the vectors \underline{d} and \underline{i} respectively. Notice that the transition probabilities in (5) are stationary across time and are parameterised by the arms configuration. We write $Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b)$ as a short hand representation of the transition probabilities in (5) to be in line with the standard notation for controlled state transitions in MDPs.

Our objective, however, is nonstandard in the context of MDPs and more in line with what information theorists study. Given an arms configuration $C = (h, P_1, P_2)$, we are interested in determining the following:

$$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^\pi[\tau(\pi) \mid C]}{\log(1/\epsilon)}. \quad (7)$$

C. SRS Policies and State-Action Occupancy Measures

Call a policy π a *stationary randomised strategy (SRS)* if there exists a Cartesian product λ of the form²

$$\lambda = \bigotimes_{(\underline{d}, \underline{i}) \in \mathbb{S}} \lambda(\cdot \mid \underline{d}, \underline{i}), \quad (8)$$

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

$$\nu^\lambda(\underline{d}, \underline{i}, a) := \mu^\lambda(\underline{d}, \underline{i}) \left(\frac{\eta}{K} + (1-\eta) \lambda(a \mid \underline{d}, \underline{i}) \right) \quad (9)$$

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

III. CONVERSE: LOWER BOUND

In this section, we present a lower bound for (7). Given two probability distributions μ and ν on \mathcal{S} , the Kullback-Leibler (KL) divergence (also called the *relative entropy*) between μ and ν is defined as

$$D(\mu \parallel \nu) := \sum_{i \in \mathcal{S}} \mu(i) \log \frac{\mu(i)}{\nu(i)}, \quad (10)$$

²Writing $\mathcal{P}(\mathcal{A})$ to denote the space of all probability distributions on \mathcal{A} , it follows that (8) is an element of the product space $\bigotimes_{(\underline{d}, \underline{i}) \in \mathbb{S}} \mathcal{P}(\mathcal{A})$.

where, by convention, $0 \log \frac{0}{0} = 0$. Also, given a TPM P on \mathcal{S} , an integer $d \geq 1$, and $i, j \in \mathcal{S}$, let $P^d(j|i)$ denote the (i, j) th entry of the matrix P^d .

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

$$\liminf_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^\pi[\tau(\pi)|C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h, P_1, P_2)}, \quad (11)$$

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

$$R^*(h, P_1, P_2) := \sup_{\pi^\lambda \in \Pi_{\text{SRS}}} \inf_{\substack{C' = (h', P'_1, P'_2): \\ h' \neq h, \\ P'_1 \neq P'_2}} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K \nu^\lambda(\underline{d}, \underline{i}, a) k_{CC'}(\underline{d}, \underline{i}, a), \quad (12)$$

with

$$\begin{aligned} k_{CC'}(\underline{d}, \underline{i}, a) &:= D((P_C^a)^{d_a}(\cdot|i_a) \parallel (P_{C'}^a)^{d_a}(\cdot|i_a)) \\ &= \begin{cases} D(P_1^{d_a}(\cdot|i_a) \parallel (P'_2)^{d_a}(\cdot|i_a)), & a = h, \\ D(P_2^{d_a}(\cdot|i_a) \parallel (P'_1)^{d_a}(\cdot|i_a)), & a = h', \\ D(P_2^{d_a}(\cdot|i_a) \parallel (P'_2)^{d_a}(\cdot|i_a)), & a \neq h, h'. \end{cases} \end{aligned} \quad (13)$$

The infimum in (12) is over all alternative odd arm configurations $C' = (h', P'_1, P'_2)$ satisfying (a) $h' \neq h$, and (b) $P'_1 \neq P'_2$.

Proof: See Appendix A. ■

Notice that closer the TPMs P_1 and P_2 are (in terms of relative entropies of the corresponding rows), the smaller is the value of $R^*(h, P_1, P_2)$, resulting in a larger value of the lower bound (11). The key ingredients in the proof of the lower bound are (a) a data processing inequality for the setting of restless arms based on a change of measure argument presented in [13], (b) a Wald-type lemma for Markov processes, (c) a recognition of the fact that for any $(\underline{d}, \underline{i}) \in \mathbb{S}$, the long-term fraction of exits from $(\underline{d}, \underline{i})$ matches the long-term fraction of entries to $(\underline{d}, \underline{i})$, and (d) the restriction of the supremum in (12) to the class of SRS policies, which is possible thanks to an analogue of [22, Theorem 8.8.2] for countable-state controlled Markov processes. A formal statement of this theorem as applicable to this paper may be found in [1, Appendix H].

A. Simplifying $R^*(h, P_1, P_2)$

Note that

$$\begin{aligned} R^*(h, P_1, P_2) &= \sup_{\pi^\lambda \in \Pi_{\text{SRS}}} \inf_{\substack{C' = (h', P'_1, P'_2): \\ h' \neq h, \\ P'_1 \neq P'_2}} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \left[\nu^\lambda(\underline{d}, \underline{i}, h) D(P_1^{d_h}(\cdot|i_h) \parallel (P'_2)^{d_h}(\cdot|i_h)) + \nu^\lambda(\underline{d}, \underline{i}, h') D(P_2^{d_{h'}}(\cdot|i_{h'}) \parallel (P'_1)^{d_{h'}}(\cdot|i_{h'})) \right. \\ &\quad \left. + \sum_{a \neq h, h'} \nu^\lambda(\underline{d}, \underline{i}, a) D(P_2^{d_a}(\cdot|i_a) \parallel (P'_2)^{d_a}(\cdot|i_a)) \right]. \end{aligned} \quad (14)$$

Because P'_1 appears only in the second term within the square brackets, it follows that the infimum over all P'_1 of this term is equal to zero and may be achieved by setting $P'_1 = P_2$. Therefore,

$$\begin{aligned} R^*(h, P_1, P_2) &= \sup_{\pi^\lambda \in \Pi_{\text{SRS}}} \inf_{\substack{h', P'_2: \\ h' \neq h, P'_2 \neq P_2}} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \left[\nu^\lambda(\underline{d}, \underline{i}, h) D(P_1^{d_h}(\cdot|i_h) \parallel (P'_2)^{d_h}(\cdot|i_h)) + \sum_{a \neq h, h'} \nu^\lambda(\underline{d}, \underline{i}, a) D(P_2^{d_a}(\cdot|i_a) \parallel (P'_2)^{d_a}(\cdot|i_a)) \right]. \end{aligned} \quad (15)$$

It is worth comparing (15) with [9, Eq. (19), pp. 4328], the final expression for the lower bound when the arms are rested. The absence of arm delays in [9, Eq. (19), pp. 4328] makes it possible to simplify this term further. In fact, a close examination of the steps presented in [9, pp. 4337-4338] shows that when the arms are rested, the inner infimum over all $h' \neq h$ and $P'_2 \neq P_2$ may be simplified by first computing the infimum over all $h' \neq h$ to arrive at the expression [9, Eq. (91), pp. 4338], following which the optimal value of P'_2 (the one that attains the infimum over all $P'_2 \neq P_2$) may be obtained using the method of Lagrange multipliers. The final expression for the optimal value of P'_2 is given by [9, Eq. (20), pp. 4328].

The presence of ergodic state-action occupancy measures inside the summation in (15) does not allow the simplification of infimum over $P'_2 \neq P_2$, as was possible in the setting of rested arms. Further, because of the presence of arm delays in the relative entropy terms in (15), it may not be possible to obtain a closed-form expression for the choice of P'_2 that attains the inner infimum in (15).

B. Near-Optimal Solutions to the Supremum in (15)

It is not clear if there exists an optimal SRS policy π^λ that attains the supremum in (12). However, for each $\delta > 0$, there exists $\lambda = \lambda_{h,P_1,P_2,\delta}(\cdot|\cdot) \in \bigotimes_{(d,i) \in \mathbb{S}} \mathcal{P}(\mathcal{A})$ such that

$$\inf_{\substack{h', P'_2: \\ h' \neq h, P'_2 \neq P_2}} \sum_{(d,i) \in \mathbb{S}} \sum_{a=1}^K \nu^\lambda(d, i, a) k_{CC'}(d, i, a) \geq \frac{R^*(h, P_1, P_2)}{1 + \delta}. \quad (16)$$

Call $\lambda_{h,P_1,P_2,\delta}$ a δ -optimal solution for the arms configuration $C = (h, P_1, P_2)$. More generally, let $\lambda_{h,P,Q,\delta}$ denote a δ -optimal solution for $C = (h, P, Q)$. Notice that more than one λ may satisfy (16).

In the next section, we show that under some regularity on the choice of the δ -optimal solutions for the various possible arms configurations, a time-varying policy based on certainty equivalence and the choice of the δ -optimal solutions achieves the lower bound asymptotically.

IV. ACHIEVABILITY

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

A. Two Key Assumptions

Given $\delta > 0$, (16) suggests that in order to approach the constant $R^*(h, P_1, P_2)$ in the lower bound to within a factor of $1/(1 + \delta)$, the arms must eventually be sampled according to $\lambda_{h,P_1,P_2,\delta}$ or one of the δ -optimal solutions for $C = (h, P_1, P_2)$. Because the underlying arms configuration is unknown and may be any one among the uncountably infinite collection $\{C = (h, P, Q) : h \in \mathcal{A}, P \neq Q\}$ of all possible arms configurations, at best one can hope for a (P, Q) that approaches (P_1, P_2) . A feasible option is to sample the arms according to $\lambda_{h,P,Q,\delta}$ and hope that it is eventually close to $\lambda_{h,P_1,P_2,\delta}$ when (P, Q) is close to (P_1, P_2) . For this to work, we need a regularity condition.

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

The paper [12] considers a similar assumption as above for a more general sequential hypothesis testing problem in multi-armed bandits, but with independent observations (see [12, Assumption A]). Also, the analogue of Assumption 1 for the

maximisers, instead of δ -optimal solutions, holds in the settings of the prior works [7]–[9] as a consequence of Berge’s maximum theorem [23]. Henceforth, for each $\delta > 0$, fix a selection $\{\lambda_{h,P,Q,\delta} : h \in \mathcal{A}, P \neq Q\}$ of δ -optimal solutions satisfying Assumption 1.

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

$$\mathcal{P}(\bar{\varepsilon}^*) := \{P \in \mathcal{P}(\mathcal{S}) : P \text{ is ergodic, } \forall d \geq 1, i, j \in \mathcal{S}, P^d(j|i) > 0 \implies P^d(j|i) \geq \bar{\varepsilon}^*\}. \quad (17)$$

Eq. (17) defines the class of all ergodic $P \in \mathcal{P}(\mathcal{S})$ such that each non-zero entry of P^d is lower bounded by $\bar{\varepsilon}^*$ uniformly in d . Clearly, every ergodic P belongs to $\mathcal{P}(\bar{\varepsilon}^*)$ for some P -dependent $\bar{\varepsilon}^*$. To see this, fix an arbitrary ergodic P , and let $\mu = (\mu(j) : j \in \mathcal{S})$ be the unique stationary distribution for P . From [24, Theorem 4.9], we have $P^d(j|i) \rightarrow \mu(j) > 0$ as $d \rightarrow \infty$ for all $i, j \in \mathcal{S}$. Let $\mu_{\min} = \min_j \mu(j)$. Then, there exists D such that $\forall d \geq D$, each non-zero entry of P^d is lower bounded by $\mu_{\min}/2$. Further, let

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

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

$$\mathcal{C}(\bar{\varepsilon}^*) := \{(P, Q) : P(\cdot|i) \text{ is mutually absolutely continuous with } Q(\cdot|i) \text{ for all } i \in \mathcal{S}, P \in \mathcal{P}(\bar{\varepsilon}^*), Q \in \mathcal{P}(\bar{\varepsilon}^*)\}, \quad (18)$$

Assumption 2. *There exists $\bar{\varepsilon}^* \in (0, 1)$ such that for every arms configuration $C = (h, P, Q)$, $(P, Q) \in \mathcal{C}(\bar{\varepsilon}^*)$.*

Some remarks are in order. An arms configuration $C = (h, P, Q)$ satisfying Assumption 2 only increases the difficulty of identifying the odd arm index h . If $P, Q \in \mathcal{P}(\bar{\varepsilon}^*)$ for some $\bar{\varepsilon}^* > 0$, then they are harder to “distinguish” from one another. To see this, note that for any ergodic P, Q , we know from [24, Proposition 2.4] that there exists $M = M(P, Q)$ such that all the entries of P^d and Q^d are strictly positive $\forall d \geq M$, thus implying that $\forall d \geq M$ and $i \in \mathcal{S}$,

$$D(P^d(\cdot|i) \| Q^d(\cdot|i)) < \infty, \quad D(Q^d(\cdot|i) \| P^d(\cdot|i)) < \infty. \quad (19)$$

For $d < M$, it may be the case that one or both of the relative entropy terms in (19) equals $+\infty$ and discrimination becomes easier. However, when $P, Q \in \mathcal{P}(\bar{\varepsilon}^*)$, it follows that $\forall d \geq 1$, each row of P^d is mutually absolutely continuous with the corresponding row of Q^d . Furthermore, $\forall d \geq 1$ and $i, j \in \mathcal{S}$ such that $P^d(j|i) > 0$, $Q^d(j|i) > 0$, the relation

$$\bar{\varepsilon}^* \leq \frac{P^d(j|i)}{Q^d(j|i)} \leq \frac{1}{\bar{\varepsilon}^*} \quad (20)$$

holds, thus implying that $\forall d \geq 1$, each of the relative entropy terms in (19) is at most $\log(1/\bar{\varepsilon}^*)$. Therefore, $P, Q \in \mathcal{P}(\bar{\varepsilon}^*)$ cannot have an arbitrarily large separation (in terms of relative entropy) and are harder to distinguish from one another.

It is worth noting here that Assumption 2 is equivalent to [20, Assumption I] which, in the context of this paper, states that there exists $\bar{\varepsilon} > 0$ such that for all $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$, $b \in \mathcal{A}$ and $C = (h, P_1, P_2)$,

$$Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b) = 0 \quad \text{or} \quad Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b) > \bar{\varepsilon}.$$

Indeed, it is easy to see from (5) that

$$Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b) = \sum_{a=1}^K \left(\frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{a=b\}} \right) (P_C^a)^{d_a} (i'_a | i_a) \mathbb{I}_{\{d'_a=1 \text{ and } d'_a=d_{\bar{a}}+1 \text{ for all } \bar{a} \neq a\}} \mathbb{I}_{\{i'_a=i_{\bar{a}} \text{ for all } \bar{a} \neq a\}}, \quad (21)$$

where P_C^a in (21) is as defined in (6). It follows that if $Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b) > 0$, then it must be true that $d'_a = 1$, $d'_{\bar{a}} = d_{\bar{a}} + 1$ for all $\bar{a} \neq a$, and $i'_a = i_{\bar{a}}$ for all $\bar{a} \neq a$ for some $a \in \mathcal{A}$. In this case,

$$Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b) = \left(\frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{a=b\}} \right) (P_C^a)^{d_a} (i'_a | i_a). \quad (22)$$

Because the term within brackets in (22) is $\geq \frac{\eta}{K} > 0$, it follows that any lower bound on $Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b)$ implies a lower bound on $(P_C^a)^{d_a}(i'_a \mid i_a)$ and vice-versa. This establishes the equivalence between Assumption 2 and [20, Assumption I].

B. Test Statistic

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

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

$$\bar{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_h) = \int_{\mathcal{P}(\mathcal{S}) \times \mathcal{P}(\mathcal{S})} f(B^n, A^n, \bar{X}^n \mid C = (h, P, Q)) D(P) D(Q) dP dQ. \quad (23)$$

Further, let $\hat{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_h)$ denote the maximum likelihood of all the arm samples and observations up to time n under the hypothesis \mathcal{H}_h , i.e.,

$$\hat{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_h) = \sup_{P, Q} f(B^n, A^n, \bar{X}^n \mid C = (h, P, Q)). \quad (24)$$

The supremum in (24) is over the compact set $\mathcal{P}(\mathcal{S}) \times \mathcal{P}(\mathcal{S})$, and is attained because the likelihood is a continuous function of the TPMs. Let $(\hat{P}_{h,1}(n), \hat{P}_{h,2}(n))$ attain the supremum in (24). These are the ML estimates of the TPMs under the hypothesis \mathcal{H}_h (whose closed-form expressions are not available in the general case). Notice that we do not impose the constraint $P \neq Q$ while computing the supremum in (24) for the following reason. We show later that under an arms configuration $C = (h, P_1, P_2)$, the convergences $\hat{P}_{h,1}(n) \rightarrow P_1, \hat{P}_{h,2}(n) \rightarrow P_2$ hold as $n \rightarrow \infty$ almost surely. Because $P_1 \neq P_2$ by the definition of arms configuration, the constraint $\hat{P}_{h,1}(n) \neq \hat{P}_{h,2}(n)$ is almost surely automatically satisfied for all n sufficiently large.

For $h, h' \in \mathcal{A}$ such that $h \neq h'$, our test statistic, which we denote by $M_{hh'}(n)$ at time n , is defined as

$$M_{hh'}(n) := \log \frac{\bar{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_h)}{\hat{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_{h'})}. \quad (25)$$

Before presenting the exact mathematical expression for (25), we introduce a few notations. Given a policy π and an arms configuration C , let $Z_C^\pi(n)$ denote the log-likelihood of all the arm samples and the observations up to time n under the policy π and the arms configuration C . That is,

$$\begin{aligned} Z_C^\pi(n) &= \log f(B^n, A^n, \bar{X}^n \mid C) \\ &= \sum_{a=1}^K \log P^\pi(X_{a-1}^a \mid C) + \sum_{t=K}^n \log P^\pi(B_t, A_t, \bar{X}_t \mid B^{t-1}, A^{t-1}, \bar{X}^{t-1}, C) \\ &= \sum_{a=1}^K \log P^\pi(X_{a-1}^a \mid C) + \sum_{t=K}^n \log P^\pi(B_t, A_t \mid B^{t-1}, A^{t-1}, \bar{X}^{t-1}, C) + \sum_{t=K}^n \log P^\pi(\bar{X}_t \mid B^t, A^t, \bar{X}^{t-1}, C) \\ &= \sum_{a=1}^K \log P^\pi(X_{a-1}^a \mid C) + \sum_{t=K}^n \log P^\pi(B_t, A_t \mid B^{t-1}, A^{t-1}, \bar{X}^{t-1}, C) + \sum_{t=K}^n \log P^\pi(\bar{X}_t \mid A^t, \bar{X}^{t-1}, C). \end{aligned} \quad (26)$$

In writing (26), we use the fact that \bar{X}_t is conditionally independent of B^t , conditioned on the actually sampled arm A_t , A^{t-1} and \bar{X}^{t-1} . The last summation term in (26) may be written as

$$\begin{aligned} \sum_{t=K}^n \log P^\pi(\bar{X}_t | A^t, \bar{X}^{t-1}, C) &= \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K \sum_{j \in \mathcal{S}} \sum_{t=K}^n \mathbb{I}_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a, \bar{X}_t^a=j\}} \log P^\pi(\bar{X}_t | A^t, \bar{X}^{t-1}, C) \\ &\stackrel{(a)}{=} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K \sum_{j \in \mathcal{S}} \sum_{t=K}^n \mathbb{I}_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a, \bar{X}_t^a=j\}} \log (P_C^a)^{d_a}(j|i_a) \\ &= \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, a, j) \log (P_C^a)^{d_a}(j|i_a), \end{aligned} \quad (27)$$

where (a) above follows by noting that when $\underline{d}(t) = \underline{d}$, $\underline{i}(t) = \underline{i}$, $A_t = a$ and $\bar{X}_t^a = j$,

$$P^\pi(\bar{X}_t | A^t, \bar{X}^{t-1}, C) = P^\pi(X_t^a = j | X_{t-d_a}^a = i_a, C) = (P_C^a)^{d_a}(j|i_a). \quad (28)$$

Also, the term $N(n, \underline{d}, \underline{i}, a, j)$ in (27) is defined as

$$N(n, \underline{d}, \underline{i}, a, j) := \sum_{t=K}^n \mathbb{I}_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a, \bar{X}_t^a=j\}}, \quad (29)$$

and denotes the number of times up to time n the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is observed to be in the state $(\underline{d}, \underline{i})$, arm a is sampled subsequently, and the state $j \in \mathcal{S}$ is observed on arm a . Plugging (27) into (26), we get

$$\begin{aligned} Z_C^\pi(n) &= \sum_{a=1}^K \log P^\pi(X_{a-1}^a | C) + \sum_{t=K}^n \log P^\pi(B_t, A_t | B^{t-1}, A^{t-1}, \bar{X}^{t-1}, C) + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, a, j) \log (P_C^a)^{d_a}(j|i_a). \end{aligned} \quad (30)$$

Under an arms configuration $C = (h, P, Q)$, (30) may be written more explicitly as

$$Z_C^\pi(n) = \sum_{t=K}^n \log P^\pi(B_t, A_t | B^{t-1}, A^{t-1}, \bar{X}^{t-1}, C) \quad (31)$$

$$+ \log P^\pi(X_{h-1}^h | C) + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h, j) \log P^{d_h}(j|i_h) \quad (32)$$

$$+ \sum_{a \neq h} \log P^\pi(X_{a-1}^a | C) + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a \neq h} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, a, j) \log Q^{d_a}(j|i_a). \quad (33)$$

Suppose that $X_0^a \sim \phi \forall a \in \mathcal{A}$ for some probability distribution ϕ on \mathcal{S} . Then, we have

$$Z_C^\pi(n) = \sum_{t=K}^n \log P^\pi(B_t, A_t | B^{t-1}, A^{t-1}, \bar{X}^{t-1}, C) \quad (34)$$

$$+ \log \left(\sum_{i \in \mathcal{S}} \phi(i) P^{h-1}(X_{h-1}^h | i) \right) + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h, j) \log P^{d_h}(j|i_h) \quad (35)$$

$$+ \sum_{a \neq h} \log \left(\sum_{i \in \mathcal{S}} \phi(i) Q^{a-1}(X_{a-1}^a | i) \right) + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a \neq h} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, a, j) \log Q^{d_a}(j|i_a). \quad (36)$$

With the above notations in place, (25) may be written as

$$M_{hh'}(n) = T_1(n) + T_2(n) + T_3(n) + T_4(n), \quad (37)$$

where the terms $T_1(n)$, $T_2(n)$, $T_3(n)$, and $T_4(n)$ are as given below.

1) The term $T_1(n)$ is given by

$$T_1(n) = \log \mathbb{E} \left[\exp \left(\log \left(\sum_{i \in \mathcal{S}} \phi(i) P^{h-1}(X_{h-1}^h | i) \right) + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h, j) \log P^{d_h}(j | i_h) \right) \right], \quad (38)$$

where the expectation is with respect to $P \sim D$.

2) The term $T_2(n)$ is given by

$$T_2(n) = \log \mathbb{E} \left[\exp \left(\sum_{a \neq h} \log \left(\sum_{i \in \mathcal{S}} \phi(i) Q^{a-1}(X_{a-1}^a | i) \right) + \sum_{a \neq h} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, a, j) \log Q^{d_a}(j | i_a) \right) \right], \quad (39)$$

where the expectation is with respect to $Q \sim D$.

3) The term $T_3(n)$ is given by

$$T_3(n) = \log \frac{1}{\sum_{i \in \mathcal{S}} \phi(i) \hat{P}_{h',1}^n(n)(X_{h'-1}^{h'} | i)} + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h', j) \log \frac{1}{(\hat{P}_{h',1}^n)^{d_{h'}}(j | i_{h'})}. \quad (40)$$

4) The term $T_4(n)$ is given by

$$T_4(n) = \sum_{a \neq h'} \log \frac{1}{\sum_{i \in \mathcal{S}} \phi(i) \hat{P}_{h',2}^n(X_{a-1}^a | i)} + \sum_{a \neq h'} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, a, j) \log \frac{1}{(\hat{P}_{h',2}^n)^{d_a}(j | i_a)}. \quad (41)$$

Notice that the term (34) does not appear in the expression for $M_{hh'}(n)$. This is because π is oblivious to the knowledge of the underlying arms configuration C , as a result of which at any given time t , the probability of selecting arm B_t based on the history $(B^{t-1}, A^{t-1}, \bar{X}^{t-1})$ is independent of C . Therefore, (34), which appears in both the numerator and the denominator of (25), cancels out.

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

C. Policy Based on Certainty Equivalence

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

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

Without loss of generality, let $A_0 = 1$, $A_1 = 2$, and so on until $A_{K-1} = K$. Follow the below mentioned steps $\forall n \geq K$.

- (1) Compute $\theta(n) \in \arg \max_{h \in \mathcal{A}} \min_{h' \neq h} M_{hh'}(n)$. Resolve ties, if any, uniformly at random.
 - (2) If $M_{\theta(n)}(n) \geq \log((K-1)L)$, stop further sampling and declare $\theta(n)$ as the odd arm.
 - (3) If $M_{\theta(n)}(n) < \log((K-1)L)$, sample arm B_{n+1} according to $\lambda_{\theta(n), \hat{P}_{\theta(n),1}(n), \hat{P}_{\theta(n),2}(n), \delta}(\cdot | \underline{d}(n), \underline{i}(n))$.
 - (4) Update $n \leftarrow n + 1$ and go back to (1).
-

In item (1) above, $\theta(n)$ denotes the guess of the odd arm at time n . In item (2), we check if $M_{\theta(n)}(n)$ has exceeded a certain fixed threshold ($\geq \log((K-1)L)$). If this is the case, then the policy is confident that the true odd arm index is $\theta(n)$, in which case the policy terminates at time n and outputs $\theta(n)$ as the odd arm index. If the condition in item (2) fails, the policy supposes that $C_n = (\theta(n), \hat{P}_{\theta(n),1}(n), \hat{P}_{\theta(n),2}(n))$ is the true arms configuration, and samples arm B_{n+1} according to the δ -optimal solution for C_n , thus following the principle of certainty equivalence. Observe that the policy does not rely on the knowledge of the constant $\bar{\varepsilon}^*$ from Assumption 2.

Remark 1. 1) The presence of arm delays in (34)-(36), the expression for the log-likelihood function, makes obtaining the closed-form expressions for $(\hat{P}_{h,1}(n), \hat{P}_{h,2}(n))$ and the exact computation of $M_{hh'}(n)$ difficult. This difficulty can

be circumvented by repeatedly sampling the Markov process of each arm and using the successive observations from each of the arms to estimate the TPMs. Because the Markov process of each arm is ergodic, these TPM estimates will converge to their true values asymptotically.

However, repeated sampling of the arms may not ensure that given $\delta > 0$ and an arms configuration $C = (h, P_1, P_2)$, the arms are sampled eventually according to $\lambda_{h,P_1,P_2,\delta}$. This, as we shall see shortly, is crucial in order to achieve the lower bound (11). Whereas, in spite of the computational intractability of $M_{hh'}(n)$, we shall see that $\pi^*(L, \delta)$ samples the arms eventually according to $\lambda_{h,P_1,P_2,\delta}$ and therefore meets the lower bound asymptotically as $L \rightarrow \infty$.

- 2) Moreover, when sampling the arms repeatedly, it is not clear if the desired error probability criterion can be met. Whereas, as we shall see, the policy $\pi^*(L, \delta)$ meets the desired error probability criterion. A key step in the proof of this result is to upper bound the likelihood $f(B^n, A^n, \bar{X}^n | C)$ under an arms configuration $C = (h, P_1, P_2)$ by the maximum likelihood $\hat{f}(B^n, A^n, \bar{X}^n | \mathcal{H}_h)$. Clearly, this is not possible for a policy that does not use the maximum likelihood (for e.g., the policy that uses TPM estimates resulting from repeated sampling of the arms).

D. Results on the Performance of the Policy

We now present the results on the performance of the above described policy. The proofs are relegated to the Appendices. Let Assumptions 1 and 2 hold.

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

$$\hat{P}_{h,1}(n) \longrightarrow P_1, \quad \hat{P}_{h,2}(n) \longrightarrow P_2 \quad \text{as } n \rightarrow \infty \quad \text{almost surely.} \quad (42)$$

Proof: See Appendix B. ■

The proof of Proposition 2 is based on verifying that the assumptions of [5] hold in the context of this paper. The result then simply follows from [5, Theorem 4.3]. It is instructive to note here that the proof of [5, Theorem 4.3] is based on a notion of “ $\{\varepsilon_i\}$ -randomisation” which, for controlled Markov processes, ensures a strictly positive probability of choosing each control at each time instant. The trembling hand model (1) of this paper ensures that the probability of sampling an arm at any given time is $\geq \frac{\eta}{K} > 0$, thus alleviating the need to consider $\{\varepsilon_i\}$ -randomisations.

An immediate consequence of Proposition 2 is the following: suppose $C = (h, P_1, P_2)$ is the underlying arms configuration. Then, for any $h' \in \mathcal{A}$ such that $h' \neq h$,

$$\hat{P}_{h',1}(n) \longrightarrow P_2, \quad \hat{P}_{h',2}(n) \longrightarrow P \quad \text{as } n \rightarrow \infty \quad \text{almost surely,} \quad (43)$$

where P in (43) is a transition probability matrix that is a function of P_1 and P_2 , but whose closed-form expression is not available. It is worth noting here that when the arms are rested, the (i, j) th entry of P is a convex combination of the (i, j) th entries of P_1 and P_2 whose closed-form expression given by [9, Eq. (20)].

The next result shows that the test statistic has a strictly positive drift under the non-stopping version of the policy (a policy, say $\pi_{ns}^*(L, \delta)$, that never stops and picks an arm at time n according to the rule in item (3) above).

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

$$\liminf_{n \rightarrow \infty} \frac{M_{hh'}(n)}{n} > 0 \quad \text{almost surely.} \quad (44)$$

Proof: The proof uses the convergences in (42). See Appendix C for the details. ■

An immediate consequence of Proposition 3 is that, almost surely, $\liminf_{n \rightarrow \infty} M_h(n) > 0$. For each $h' \in \mathcal{A}$, let $\pi_{h'}^*(L, \delta)$ denote a version of the policy $\pi^*(L, \delta)$ that waits until the event $M_{h'}(n) \geq \log((K-1)L)$ occurs, at which point it stops and always declares h' as the index of the odd arm. Under the arms configuration $C = (h, P_1, P_2)$, it then follows that the stopping time of the policy $\pi^*(L, \delta)$ may be upper bounded by that of $\pi_{h'}^*(L, \delta)$, i.e., $\tau(\pi^*(L, \delta)) \leq \tau(\pi_{h'}^*(L, \delta))$ almost surely under $C = (h, P_1, P_2)$. As a consequence, under $C = (h, P_1, P_2)$, the following set of inequalities hold almost surely:

$$\begin{aligned} \tau(\pi^*(L, \delta)) &\leq \tau(\pi_h^*(L, \delta)) \\ &= \inf\{n \geq 1 : M_h(n) \geq \log((K-1)L)\} \\ &\leq \inf\left\{n \geq 1 : M_{hh'}(n') \geq \log((K-1)L) \text{ for all } n' \geq n \text{ and for all } h' \neq h\right\} \\ &< \infty, \end{aligned} \tag{45}$$

where the last line above is due to Proposition 3. This establishes that the policy $\pi^*(L, \delta)$ stops in finite time almost surely.

For $h' \in \mathcal{A}$ such that $h' \neq h$, let

$$\text{GLR}_{hh'}(n) := \log \frac{\hat{f}(B^n, A^n, \bar{X}^n | \mathcal{H}_h)}{\hat{f}(B^n, A^n, \bar{X}^n | \mathcal{H}_{h'})} \tag{46}$$

denote the GLR test statistic of hypothesis \mathcal{H}_h with respect to the hypothesis $\mathcal{H}_{h'}$ at time n . Clearly, $\text{GLR}_{hh'}(n) \geq M_{hh'}(n)$ almost surely for all n . Also, $\text{GLR}_{h'h}(n) = -\text{GLR}_{hh'}(n)$. Then, almost surely,

$$\begin{aligned} \limsup_{n \rightarrow \infty} M_{h'}(n) &= \limsup_{n \rightarrow \infty} \min_{a \neq h'} M_{h'a}(n) \\ &\leq \limsup_{n \rightarrow \infty} M_{h'h}(n) \\ &\leq \limsup_{n \rightarrow \infty} \text{GLR}_{h'h}(n) \\ &= \limsup_{n \rightarrow \infty} -\text{GLR}_{hh'}(n) \\ &= -\liminf_{n \rightarrow \infty} \text{GLR}_{hh'}(n) \\ &\leq -\liminf_{n \rightarrow \infty} M_{hh'}(n) \\ &\leq -\liminf_{n \rightarrow \infty} \min_{a \neq h} M_{h,a}(n) \\ &= -\liminf_{n \rightarrow \infty} M_h(n) \\ &< 0, \end{aligned} \tag{47}$$

where (47) follows from Proposition 3. Under the arms configuration $C = (h, P_1, P_2)$ and under the policy $\pi_{\text{ns}}^*(L, \delta)$, almost surely,

$$\theta(n) = \arg \max_{h \in \mathcal{A}} M_h(n) = h \quad \forall n \text{ sufficiently large}, \tag{48}$$

which together with (42) implies that

$$\lambda_{\theta(n), \hat{P}_{\theta(n),1}(n), \hat{P}_{\theta(n),2}(n), \delta} \longrightarrow \lambda_{h, P_1, P_2, \delta} \quad \text{as } n \rightarrow \infty, \tag{49}$$

where the convergence in (49) is with respect to the product topology on the space $\bigotimes_{(\underline{d}, \underline{i}) \in \mathbb{S}} \mathcal{P}(\mathcal{A})$, i.e.,

$$\lambda_{\theta(n), \hat{P}_{\theta(n),1}(n), \hat{P}_{\theta(n),2}(n), \delta}(\cdot | \underline{d}, \underline{i}) \rightarrow \lambda_{h, P_1, P_2, \delta}(\cdot | \underline{d}, \underline{i}) \quad \text{as } n \rightarrow \infty \quad \forall (\underline{d}, \underline{i}) \in \mathbb{S}. \tag{50}$$

The property in (49) ensures that under the non-stopping policy $\pi_{\text{ns}}^*(L, \delta)$, as more observations are collected from the arms and as the ML estimates of the TPMs approach their true values defined by the underlying arms configuration, the arms are

eventually sampled according to the δ -optimal selection for the underlying arms configuration. As we shall see shortly, it is this property of $\pi_{\text{ns}}^*(L, \delta)$ that plays a key role in showing the asymptotic optimality of the policy $\pi^*(L, \delta)$.

Let us return to the policy $\pi^*(L, \delta)$. The next result shows that any error probability $\epsilon > 0$ can be met by setting the parameter $L = 1/\epsilon$.

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

Proof: The proof uses the fact that the policy stops in finite time almost surely. The details are in Appendix D. ■

An important step in the proof of Proposition 4 is to upper bound the likelihood function under the arms configuration $C = (h, P_1, P_2)$ by the maximum likelihood under the hypothesis \mathcal{H}_h . Such an exercise of upper bounding the likelihood function is not possible if, for instance, the ML estimates in the expression for the maximum likelihood are replaced by the estimates of the TPMs computed by repeatedly sampling each of the arms.

The next result improves upon Proposition 3 and shows that the modified GLR test statistic has the correct drift under the non-stopping version of the policy $\pi^*(L, \delta)$.

Proposition 5. *Suppose that $C = (h, P_1, P_2)$ is the underlying arms configuration. Fix $L > 1$ and $\delta > 0$. For all $h' \in \mathcal{A}$ such that $h' \neq h$, under the non-stopping version of the policy $\pi^*(L, \delta)$,*

$$\liminf_{n \rightarrow \infty} \frac{M_{hh'}(n)}{n} \geq \frac{R^*(h, P_1, P_2)}{(1 + \delta)^2} \quad \text{almost surely.} \quad (51)$$

Proof: It is in proving this proposition that we make use of Assumption 1. The proof uses the fact that almost surely, $\theta(n) = h$ for all sufficiently large n under the arms configuration $C = (h, P_1, P_2)$. As a result, $\hat{P}_{\theta(n),1}(n) = \hat{P}_{h,1}(n)$, $\hat{P}_{\theta(n),2}(n) = \hat{P}_{h,2}(n)$ for all sufficiently large n . The convergences in (42) imply that the pair $(\hat{P}_{h,1}(n), \hat{P}_{h,2}(n))$ lies inside a neighbourhood (chosen based on the value of δ) of (P_1, P_2) for all n sufficiently large. These together with Assumption 1 and (49) imply that $\lambda_{\theta(n), \hat{P}_{\theta(n),1}(n), \hat{P}_{\theta(n),2}(n), \delta}$ is close to $\lambda_{h, P_1, P_2, \delta}$ for all n sufficiently large. Using this, we arrive at (51). For the details, see Appendix E. ■

Next, we show that the stopping time of the policy $\pi^*(L, \delta)$ blows up as $L \rightarrow \infty$.

Proposition 6. *For each $\delta > 0$,*

$$\liminf_{L \rightarrow \infty} \tau(\pi^*(L, \delta)) = \infty \quad \text{almost surely.} \quad (52)$$

Proof: See Appendix F. ■

Combining the results of Proposition 5 and Proposition 6, it follows that for all $\delta > 0$,

$$\liminf_{L \rightarrow \infty} \frac{M_{hh'}(\tau(\pi^*(L, \delta)))}{\tau(\pi^*(L, \delta))} \geq \frac{R^*(h, P_1, P_2)}{(1 + \delta)^2} \quad \text{almost surely.} \quad (53)$$

The following result shows that the stopping time of the policy $\pi^*(L, \delta)$ satisfies an almost sure upper bound that is arbitrarily close to the lower bound in (11).

Proposition 7. *Suppose that $C = (h, P_1, P_2)$ is the underlying arms configuration. For all $\delta > 0$, the stopping time of policy $\pi^*(L, \delta)$ satisfies*

$$\limsup_{L \rightarrow \infty} \frac{\tau(\pi^*(L, \delta))}{\log L} \leq \frac{(1 + \delta)^2}{R^*(h, P_1, P_2)} \quad \text{almost surely.} \quad (54)$$

Proof of Proposition 7: By the definition of $\tau(\pi^*(L, \delta))$, we know that under the arms configuration $C = (h, P_1, P_2)$,

$$M_h(\tau(\pi^*(L, \delta))) \geq \log((K - 1)L), \quad M_h(\tau(\pi^*(L, \delta)) - 1) < \log((K - 1)L).$$

We then have, almost surely,

$$\begin{aligned}
1 &= \limsup_{L \rightarrow \infty} \frac{\log((K-1)L)}{\log L} \\
&\geq \limsup_{L \rightarrow \infty} \frac{M_h(\tau(\pi^*(L, \delta)) - 1)}{\log L} \\
&= \limsup_{L \rightarrow \infty} \frac{M_h(\tau(\pi^*(L, \delta)) - 1)}{\tau(\pi^*(L, \delta)) - 1} \cdot \frac{\tau(\pi^*(L, \delta)) - 1}{\log L} \\
&\geq \frac{R^*(h, P_1, P_2)}{(1 + \delta)^2} \cdot \limsup_{L \rightarrow \infty} \frac{\tau(\pi^*(L, \delta))}{\log L},
\end{aligned} \tag{55}$$

where (55) follows from (53) and Assumption 2; here, Assumption 2 guarantees that the increment $M_{hh'}(n) - M_{hh'}(n-1)$ is bounded almost surely. The desired result follows by rearranging (55). ■

The main result of this section on the expected stopping time of the policy is stated next.

Proposition 8. Fix $\delta > 0$. Under the arms configuration $C = (h, P_1, P_2)$, the expected stopping time of the policy $\pi = \pi^*(L, \delta)$ satisfies

$$\limsup_{L \rightarrow \infty} \frac{E^\pi[\tau(\pi)|C]}{\log L} \leq \frac{(1 + \delta)^2}{R^*(h, P_1, P_2)}. \tag{56}$$

Proof: In the proof, which we present in Appendix G, we show that for each $\delta > 0$, the family $\{\tau(\pi^*(L, \delta))/\log L : L > 1\}$ is uniformly integrable. Combining the almost sure upper bound in (54) with uniform integrability yields the desired upper bound in (56) for the expected stopping time of the policy $\pi^*(L, \delta)$. ■

V. MAIN RESULT

With the above ingredients in place, the main result of this paper is as below.

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 an arms configuration $C = (h, P_1, P_2)$; here, h is the odd arm, P_1 is the TPM of the odd arm h , and $P_2 \neq P_1$ is the TPM of each non-odd arm. Fix $\eta \in (0, 1]$, and suppose that a decision entity that wishes to identify the odd arm has a trembling hand with parameter η . When neither P_1 nor P_2 is known beforehand to the decision entity, the expected time required to identify the odd arm satisfies the asymptotic lower bound

$$\liminf_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^\pi[\tau(\pi)|C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h, P_1, P_2)}. \tag{57}$$

Further, under a continuous selection assumption and a regularity condition on the TPMs P_1 and P_2 , the policy $\pi^*(L, \delta)$ satisfies the asymptotic upper bound

$$\limsup_{\delta \downarrow 0} \limsup_{L \rightarrow \infty} \frac{E^{\pi^*(L, \delta)}[\tau(\pi^*(L, \delta))|C]}{\log L} \leq \frac{1}{R^*(h, P_1, P_2)}. \tag{58}$$

Combining the lower and the upper bounds, we have

$$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^\pi[\tau(\pi)|C]}{\log(1/\epsilon)} = \lim_{\delta \downarrow 0} \lim_{L \rightarrow \infty} \frac{E^{\pi^*(L, \delta)}[\tau(\pi^*(L, \delta))|C]}{\log L} = \frac{1}{R^*(h, P_1, P_2)}. \tag{59}$$

Proof: The asymptotic lower bound in (57) follows from Proposition 1. Taking limits as $\delta \downarrow 0$ in (56), we arrive at (58). From Proposition 4, we know that given any error tolerance parameter $\epsilon > 0$, by setting $L = 1/\epsilon$, we have $\pi^*(L, \delta) \in \Pi(\epsilon)$ for all $\delta > 0$. Therefore, it follows that for all $\epsilon, \delta > 0$,

$$\inf_{\pi \in \Pi(\epsilon)} \frac{E^\pi[\tau(\pi)|C]}{\log(1/\epsilon)} \leq \frac{E^{\pi^*(L, \delta)}[\tau(\pi^*(L, \delta))|C]}{\log L}. \tag{60}$$

Fixing $\delta > 0$ and letting $\epsilon \downarrow 0$ (which is identical to letting $L \rightarrow \infty$) in (60), and using the upper bound in (56), we get

$$\limsup_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^\pi[\tau(\pi)|C]}{\log(1/\epsilon)} \leq \limsup_{L \rightarrow \infty} \frac{E^{\pi^*(L, \delta)}[\tau(\pi^*(L, \delta))|C]}{\log L} \leq \frac{1 + \delta}{R^*(h, P_1, P_2)}. \quad (61)$$

Letting $\delta \downarrow 0$ in (61) and noting that the leftmost term in (61) does not depend on δ , we get

$$\limsup_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^\pi[\tau(\pi)|C]}{\log(1/\epsilon)} \leq \limsup_{\delta \downarrow 0} \limsup_{L \rightarrow \infty} \frac{E^{\pi^*(L, \delta)}[\tau(\pi^*(L, \delta))|C]}{\log L} \leq \frac{1}{R^*(h, P_1, P_2)}. \quad (62)$$

Combining the result in (62) with the lower bound in (11), we get

$$\begin{aligned} \frac{1}{R^*(h, P_1, P_2)} &\leq \liminf_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^\pi[\tau(\pi)|C]}{\log(1/\epsilon)} \\ &\leq \limsup_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^\pi[\tau(\pi)|C]}{\log(1/\epsilon)} \\ &\leq \limsup_{\delta \downarrow 0} \limsup_{L \rightarrow \infty} \frac{E^{\pi^*(L, \delta)}[\tau(\pi^*(L, \delta))|C]}{\log L} \\ &\leq \frac{1}{R^*(h, P_1, P_2)}. \end{aligned} \quad (63)$$

Thus, it follows that the limit infimum and the limit suprema in the chain of inequalities leading to (63) are indeed limits, thereby yielding (59). This completes the proof of the theorem. \blacksquare

Thus, under the arms configuration $C = (h, P_1, P_2)$, the asymptotic growth rate for the expected time required to identify the odd arm is $1/R^*(h, P_1, P_2)$. While the asymptotic lower bound holds for all arms configurations, the asymptotic upper bound is only established under a continuous selection assumption for those arms configurations whose TPMs satisfy a regularity condition. The continuous selection assumption and the regularity condition are used to prove identifiability. The trembling hand model (1) ensures that at any given time, each arm is selected with a strictly positive probability under every policy, and plays a key role in deriving the lower and the upper bounds.

VI. CONCLUDING REMARKS

- 1) In this paper, we showed that under the arms configuration $C = (h, P_1, P_2)$, when neither P_1 nor P_2 is known beforehand to the decision entity and the when arms are restless, the answer to (7) is $1/R^*(h, P_1, P_2)$. This is the growth rate of the expected time required to find the odd arm index. The closer the transition probability matrices (TPMs) P_1 and P_2 are in terms of the relative entropies of the corresponding rows, the smaller is the value of $R^*(h, P_1, P_2)$, and therefore the larger the growth rate.
- 2) The main results of this paper are the following. (a) We gave an asymptotic lower bound on the expected time required to find the odd arm, subject to an upper bound on the error probability (PAC setting). The asymptotics is as the error probability vanishes. (b) We gave a policy (called $\pi^*(L, \delta)$) based on the principle of certainty equivalence based on maximum likelihood (ML) estimates that achieves the lower bound asymptotically. The achievability analysis relied crucially on showing that (a) the ML estimates of the TPMs converge to their true values asymptotically, and (b) given $\delta > 0$ and an arms configuration C , the policy $\pi^*(L, \delta)$ eventually samples the arms according to the δ -optimal solution for C as given by the lower bound. These conditions were established under (a) the existence of a continuous selection of δ -optimal solutions, and (b) a certain regularity assumption on the TPMs.
- 3) The existence of a continuous selection of δ -optimal solutions was crucial in establishing that the ML estimates of the TPMs converge to their true values. In [7]–[9], the continuous selection property was established for the maximisers

instead of δ -optimal solutions. However, for more general settings such as the setting of this paper or the setting considered in [12], a continuous selection assumption seems inevitable and difficult to do away with.

- 4) The computability of $R^*(h, P_1, P_2)$ is an issue because it contains an outer supremum over all SRS policies which is difficult to evaluate. An algorithm such as Q -learning may need to be employed to compute $R^*(h, P_1, P_2)$.
- 5) The policy $\pi^*(L, \delta)$ is based a modified GLR test statistic that consists of an average average likelihood evaluated with respect to an artificial prior in the numerator, and the maximum likelihood in the denominator. The computability of the maximum likelihood is an issue because closed-form expressions for the ML estimates of the TPMs are not available. At best, the ML estimates can be evaluated numerically.
- 6) Repeatedly sampling each arm and using the consecutive observations from the arms to estimate the TPMs may lead to simple closed-form expressions for the TPM estimates. By virtue of the ergodic theorem, these estimates converge to their true values. However, in this case, it is not clear if the arms will be sampled eventually according to the δ -optimal solution for the underlying arms configuration. Also, it is not clear if the desired error probability can be met. Policies that sample the arms repeatedly are known to perform well for the problem of minimising regret [17]. However, it is not clear if such policies perform well for optimal stopping problems such as that studied in this paper.
- 7) It would be interesting to study the case when the trembling hand parameter $\eta = 0$. Preliminary analysis in [1, Section VII] for the case when the TPMs P_1 and P_2 are known beforehand reveals that for each $\eta > 0$, the upper and the lower bounds match (as in this paper). However, the limiting values of these bounds as $\eta \downarrow 0$ may exhibit a gap. We anticipate that a similar behaviour holds when neither P_1 nor P_2 is known beforehand. We leave the details open for exploration.

APPENDIX A

PROOF OF PROPOSITION 1

Given a policy π and two arm configurations $C = (h, P_1, P_2)$ and $C' = (h', P'_1, P'_2)$ such that $h \neq h'$, let $Z_{CC'}^\pi(n) := Z_C^\pi(n) - Z_{C'}^\pi(n)$ denote the log-likelihood ratio, under the policy π , of all the arm samples and observations up to time n under the arms configuration C with respect to that under C' . Note that (34), being independent of the arms configuration and common to the expressions for $Z_C^\pi(n)$ and $Z_{C'}^\pi(n)$, cancels out when writing the expression for $Z_C^\pi(n) - Z_{C'}^\pi(n)$. It then follows from (30) that

$$Z_{CC'}^\pi(n) = \sum_{a=1}^K \log \frac{P_C^\pi(X_{a-1}^a)}{P_{C'}^\pi(X_{a-1}^a)} + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, a, j) \log \frac{(P_C^a)^{d_a}(j|i_a)}{(P_{C'}^a)^{d_a}(j|i_a)}. \quad (64)$$

We organise the proof of Proposition 1 as follows. Given an error probability $\epsilon > 0$, we first obtain a lower bound for $E^\pi[Z_{CC'}^\pi(\tau(\pi))|C]$ for all $\pi \in \Pi(\epsilon)$ using a change of measure argument of Kaufmann et al. [13]. Following this, we obtain an upper bound for $E^\pi[Z_{CC'}^\pi(\tau(\pi))|C]$ in terms of $E^\pi[\tau(\pi)|C]$ using a Wald-type lemma for the setting of restless Markov arms. Combining the upper and the lower bounds, and letting $\epsilon \downarrow 0$, we arrive at the desired result. The ergodicity property established in [1, Lemma 1] for SRS policies plays a crucial role in deriving the final lower bound (11).

A. A Lower Bound on $E^\pi[Z_{CC'}^\pi(\tau(\pi))|C]$ for $\pi \in \Pi(\epsilon)$

As a first step towards deriving the lower bound, we use a result of Kaufmann et al. [13] to obtain a lower bound for $E^\pi[Z_{CC'}^\pi(\tau(\pi))|C]$ in terms of the error probability parameter ϵ . This is based on a generalisation of [13, 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 1. Let (Ω, \mathcal{F}) be a measurable space. Let the filtration $\{\mathcal{F}_t : t \geq 0\}$ be defined as follows: $\mathcal{F}_0 = \sigma(\Omega, \emptyset)$ and $\mathcal{F}_t = \sigma(B^t, A^t, \bar{X}^t)$ for all $t \geq 1$. 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

$$\mathcal{F}_{\tau(\pi)} = \{E \in \mathcal{F} : E \cap \{\tau(\pi) = t\} \in \mathcal{F}_t \text{ for all } t \geq 0\}. \quad (65)$$

Then, for all $C = (h, P_1, P_2)$ and $C' = (h', P'_1, P'_2)$ such that $h' \neq h$,

$$P^\pi(E|C') = E^\pi[1_E \exp(-Z_{C'}^\pi(\tau(\pi)))]|C] \quad (66)$$

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

Proof of Lemma 1: We prove the lemma by demonstrating, through mathematical induction, that the relation

$$E^\pi[g(B^t, A^t, \bar{X}^t)|C'] = E^\pi[g(B^t, A^t, \bar{X}^t) \exp(-Z_{C'}^\pi(t))|C] \quad (67)$$

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

$$\begin{aligned} E^\pi[g(B_0, A_0, \bar{X}_0)|C'] &= \sum_{b=1}^K \sum_{a=1}^K \sum_{i \in \mathcal{S}} g(b, a, i) P^\pi(B_0 = b, A_0 = a, \bar{X}_0 = i|C') \\ &= \sum_{b=1}^K \sum_{a=1}^K \sum_{i \in \mathcal{S}} g(b, a, i) P^\pi(B_0 = b|C') P^\pi(A_0 = a|B_0 = b, C') P^\pi(\bar{X}_0 = i|B_0 = b, A_0 = a, C') \\ &\stackrel{(a)}{=} \sum_{b=1}^K \sum_{a=1}^K \sum_{i \in \mathcal{S}} g(b, a, i) P^\pi(B_0 = b|C) P^\pi(A_0 = a|B_0 = b, C) P^\pi(\bar{X}_0 = i|A_0 = a, C') \\ &= \sum_{b=1}^K \sum_{a=1}^K \sum_{i \in \mathcal{S}} g(b, a, i) P^\pi(B_0 = b|C) P^\pi(A_0 = a|B_0 = b, C) P^\pi(X_0^a = i|C'), \end{aligned} \quad (68)$$

where (a) follows by noting that

- $P^\pi(B_0 = b|C') = P^\pi(B_0 = b|C)$ because π is oblivious to the underlying arms configuration, and
- $P^\pi(A_0 = a|B_0 = b, C') = P^\pi(A_0 = a|B_0 = b, C) = \frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{a=b\}}$.

Assuming that $X_0^a \sim \phi$ under π , where ϕ is a probability distribution on \mathcal{S} that is independent of the underlying arms configuration (which is not known to π), we have

$$E^\pi[g(B_0, A_0, \bar{X}_0)|C'] = \sum_{b=1}^K \sum_{a=1}^K \sum_{i \in \mathcal{S}} g(b, a, i) P^\pi(B_0 = b|C) P^\pi(A_0 = a|B_0 = b, C) \phi(i) \quad (69)$$

$$= \sum_{b=1}^K \sum_{a=1}^K \sum_{i \in \mathcal{S}} g(b, a, i) P^\pi(B_0 = b|C) P^\pi(A_0 = a|B_0 = b, C) P^\pi(X_0^a = i|A_0 = a, C) \quad (70)$$

$$= \sum_{b=1}^K \sum_{a=1}^K \sum_{i \in \mathcal{S}} g(b, a, i) P^\pi(B_0 = b|C) P^\pi(A_0 = a|B_0 = b, C) P^\pi(X_0^a = i|A_0 = a, B_0 = b, C). \quad (71)$$

Noting that

$$Z_{C'}^\pi(0) = \log \frac{P^\pi(B_0, A_0, \bar{X}_0|C)}{P^\pi(B_0, A_0, \bar{X}_0|C')} = 0, \quad (72)$$

and combining (71) with (72), we get

$$E^\pi[g(B_0, A_0, \bar{X}_0)|C'] = E^\pi[g(B_0, A_0, \bar{X}_0|C) \exp(-Z_{C'}^\pi(0))].$$

This proves (67) for $t = 0$.

Assume now that (67) is true for some $t > 0$. We shall demonstrate that (67) holds for $t + 1$. By the law of iterated expectations,

$$E^\pi[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})|C'] = E^\pi[E^\pi[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})|\mathcal{F}_t, C']|C']. \quad (73)$$

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

$$E^\pi[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})|\mathcal{F}_t, C'] = E^\pi[E^\pi[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})|\mathcal{F}_t, C'] \exp(-Z_{CC'}^\pi(t))|C]. \quad (74)$$

We now note that

$$\begin{aligned} & E^\pi[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})|\mathcal{F}_t, C'] \exp(-Z_{CC'}^\pi(t)) \\ & \stackrel{(a)}{=} E^\pi[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1}) \exp(-Z_{CC'}^\pi(t))|\mathcal{F}_t, C'] \\ & = \sum_{b=1}^K \sum_{a=1}^K \sum_{i \in \mathcal{S}} \left[g(B^t, A^t, \bar{X}^t, b, a, i) \cdot P^\pi(B_{t+1} = b|B^t, A^t, \bar{X}^t, C') \cdot P^\pi(A_{t+1} = a|B^{t+1} = b, B^t, A^t, \bar{X}^t, C') \right. \\ & \quad \left. \cdot P^\pi(\bar{X}_{t+1} = i|B^{t+1} = b, A_{t+1} = a, B^t, A^t, \bar{X}^t, C') \cdot \exp(-Z_{CC'}^\pi(t)) \right] \\ & \stackrel{(b)}{=} \sum_{b=1}^K \sum_{a=1}^K \sum_{i \in \mathcal{S}} \left[g(B^t, A^t, \bar{X}^t, b, a, i) \cdot P^\pi(B_{t+1} = b|B^t, A^t, \bar{X}^t, C) \cdot P^\pi(A_{t+1} = a|B^{t+1} = b, B^t, A^t, \bar{X}^t, C) \right. \\ & \quad \left. \cdot P^\pi(\bar{X}_{t+1} = i|A_{t+1} = a, A^t, \bar{X}^t, C') \cdot \exp(-Z_{CC'}^\pi(t)) \right], \end{aligned} \quad (75)$$

where (a) follows by observing that $Z_{CC'}^\pi(t)$ is a measurable function of (B^t, A^t, \bar{X}^t) , and in writing (b), we use the following facts: for any t ,

- $P^\pi(B_{t+1} = b|B^t, A^t, \bar{X}^t, C') = P^\pi(B_{t+1} = b|B^t, A^t, \bar{X}^t, C)$ because π is oblivious to the underlying arms configuration,
- $P^\pi(A_{t+1} = a|B_{t+1} = b, B^t, A^t, \bar{X}^t, C') = P^\pi(A_{t+1} = a|B_{t+1} = b, B^t, A^t, \bar{X}^t, C) = \frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{a=b\}}$, and
- $P^\pi(\bar{X}_{t+1} = i|B_{t+1} = b, A_{t+1} = a, B^t, A^t, \bar{X}^t, C') = P^\pi(\bar{X}_{t+1} = i|A_{t+1} = a, A^t, \bar{X}^t, C')$ because the observation obtained from arm a at time t is a function only of the delay and the last observed state of arm a as measured at time t , both of which may be deduced from (A^t, \bar{X}^t) .

Also, we have

$$\begin{aligned} & \sum_{i \in \mathcal{S}} P^\pi(\bar{X}_{t+1} = i|A_{t+1} = a, A^t, \bar{X}^t, C') \exp(-Z_{CC'}^\pi(t)) \\ & = \sum_{i \in \mathcal{S}} \frac{P^\pi(\bar{X}_{t+1} = i|A_{t+1} = a, A^t, \bar{X}^t, C')}{P^\pi(\bar{X}_{t+1} = i|A_{t+1} = a, A^t, \bar{X}^t, C)} \exp(-Z_{CC'}^\pi(t)) P^\pi(\bar{X}_{t+1} = i|A_{t+1} = a, A^t, \bar{X}^t, C) \\ & = \sum_{i \in \mathcal{S}} \exp(-Z_{CC'}^\pi(t+1, a, i)) P^\pi(\bar{X}_{t+1} = i|A_{t+1} = a, A^t, \bar{X}^t, C), \end{aligned} \quad (76)$$

since

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

Substituting (76) in (75) and simplifying, we get

$$\begin{aligned} & E^\pi[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})|\mathcal{F}_t, C'] \exp(-Z_{CC'}^\pi(t)) \\ & = \sum_{b=1}^K \sum_{a=1}^K \sum_{i \in \mathcal{S}} \left[g(B^t, A^t, \bar{X}^t, b, a, i) \cdot P^\pi(B_{t+1} = b|B^t, A^t, \bar{X}^t, C) \cdot P^\pi(A_{t+1} = a|B_{t+1} = b, B^t, A^t, \bar{X}^t, C) \right. \\ & \quad \left. \cdot P^\pi(\bar{X}_{t+1} = i|B_{t+1} = b, A_{t+1} = a, B^t, A^t, \bar{X}^t, C) \cdot \exp(-Z_{CC'}^\pi(t+1, a, i)) \right] \end{aligned} \quad (77)$$

$$= E^\pi[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1}) \exp(-Z_{CC'}^\pi(t+1)) | \mathcal{F}_t, C]. \quad (78)$$

Thus, we have

$$E^\pi[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1}) | \mathcal{F}_t, C'] \exp(-Z_{CC'}^\pi(t)) = E^\pi[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1}) \exp(-Z_{CC'}^\pi(t+1)) | \mathcal{F}_t, C]. \quad (79)$$

Applying $E^\pi[\cdot | C]$ to both sides of (79), plugging it into the right hand side of (74), and using the law of iterated expectations, we arrive at the desired relation for $t+1$. This proves (67) for all $t \geq 0$.

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

$$\begin{aligned} P^\pi(E|C') &= E^\pi[1_E | C'] \\ &= E^\pi \left[\sum_{t \geq 0} 1_{E \cap \{\tau(\pi)=t\}} \middle| C' \right] \\ &\stackrel{(a)}{=} \sum_{t \geq 0} E^\pi [1_{E \cap \{\tau(\pi)=t\}} | C'] \\ &\stackrel{(b)}{=} \sum_{t \geq 0} E^\pi [1_{E \cap \{\tau(\pi)=t\}} \exp(-Z_{CC'}^\pi(t)) | C] \\ &= \sum_{t \geq 0} E^\pi [1_{E \cap \{\tau(\pi)=t\}} \exp(-Z_{CC'}^\pi(\tau(\pi))) | C] \\ &= E^\pi [1_E \exp(-Z_{CC'}^\pi(\tau(\pi))) | C], \end{aligned} \quad (80)$$

where (a) is due to monotone convergence theorem, and (b) above follows from (67) 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. \blacksquare

Lemma 1 in conjunction with [13, Lemma 19] implies that when $C = (h, P_1, P_2)$ is the underlying arms configuration, the following inequality holds for all $\pi \in \Pi(\epsilon)$ and $C' = (h', P'_1, P'_2)$ such that $h' \neq h$:

$$E^\pi[Z_{CC'}^\pi(\tau(\pi)) | C] \geq \sup_{E \in \mathcal{F}_{\tau(\pi)}} d(P^\pi(E|C), P^\pi(E|C')), \quad (81)$$

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

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

is the binary relative entropy function. We now note that for any $\pi \in \Pi(\epsilon)$, when C is the actual (underlying) arms configuration,

$$P^\pi(\theta(\tau(\pi)) = h | C) \geq 1 - \epsilon, \quad P^\pi(\theta(\tau(\pi)) = h | C') \leq \epsilon.$$

As noted in [13], the mapping $x \mapsto d(x, y)$ is monotone increasing for $x < y$, and the mapping $y \mapsto d(x, y)$ is monotone decreasing for any fixed x . Using these facts in (81), we get

$$\begin{aligned} E^\pi[Z_{CC'}^\pi(\tau(\pi)) | C] &\geq d(P^\pi(\theta(\tau(\pi)) = h | C), P^\pi(\theta(\tau(\pi)) = h | C')) \\ &\geq d(\epsilon, 1 - \epsilon) \end{aligned} \quad (82)$$

for all $\pi \in \Pi(\epsilon)$ and for all $C' = (h', P'_1, P'_2)$ such that $h' \neq h$, from which it follows that

$$\inf_{\substack{C'=(h', P'_1, P'_2): \\ h' \neq h, P'_1 \neq P'_2}} E^\pi[Z_{CC'}^\pi(\tau(\pi)) | C] \geq d(\epsilon, 1 - \epsilon). \quad (83)$$

B. An Upper Bound for $E^\pi[Z_{CC'}^\pi(\tau(\pi))|C]$ in Terms of $E^\pi[\tau(\pi)|C]$

We now obtain an upper bound for the left-hand side of (83). From (64), we have

$$E^\pi[Z_{CC'}^\pi(\tau(\pi))|C] = E^\pi \left[\sum_{a=1}^K \log \frac{P_C^\pi(X_{a-1}^a)}{P_{C'}^\pi(X_{a-1}^a)} \middle| C \right] + E^\pi \left[\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K \sum_{j \in \mathcal{S}} N(\tau(\pi), \underline{d}, \underline{i}, a, j) \log \frac{(P_C^a)^{d_a}(j|i_a)}{(P_{C'}^a)^{d_a}(j|i_a)} \middle| C \right]. \quad (84)$$

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

Lemma 2. For every $(\underline{d}, \underline{i}) \in \mathbb{S}$, $a \in \mathcal{A}$ and $j \in \mathcal{S}$,

$$E^\pi[E^\pi[N(\tau(\pi), \underline{d}, \underline{i}, a, j)|X_{a-1}^a, C]|\tau(\pi), C] = E^\pi[E^\pi[N(\tau(\pi), \underline{d}, \underline{i}, a)|X_{a-1}^a, C]|\tau(\pi), C] (P_C^a)^{d_a}(j|i_a). \quad (85)$$

Proof of Lemma 2: Substituting $n = \tau(\pi)$ in (29), we have

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

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

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

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

Using Lemma 2, the second expectation term on the right-hand side of (84) can be simplified as follows.

$$\begin{aligned} &E^\pi \left[\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K \sum_{j \in \mathcal{S}} N(\tau(\pi), \underline{d}, \underline{i}, a, j) \log \frac{(P_C^a)^{d_a}(j|i_a)}{(P_{C'}^a)^{d_a}(j|i_a)} \middle| C \right] \\ &= E^\pi \left[\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K \sum_{j \in \mathcal{S}} E^\pi[E^\pi[N(\tau(\pi), \underline{d}, \underline{i}, a, j)|X_{a-1}^a, C]|\tau(\pi), C] \log \frac{(P_C^a)^{d_a}(j|i_a)}{(P_{C'}^a)^{d_a}(j|i_a)} \middle| C \right] \\ &\stackrel{(a)}{=} E^\pi \left[\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K \sum_{j \in \mathcal{S}} E^\pi[E^\pi[N(\tau(\pi), \underline{d}, \underline{i}, a)|X_{a-1}^a, C]|\tau(\pi), C] \cdot (P_C^a)^{d_a}(j|i) \cdot \log \frac{(P_C^a)^{d_a}(j|i_a)}{(P_{C'}^a)^{d_a}(j|i_a)} \right] \\ &= E^\pi \left[\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K E^\pi[E^\pi[N(\tau(\pi), \underline{d}, \underline{i}, a)|X_{a-1}^a, C]|\tau(\pi), C] \cdot D((P_C^a)^{d_a}(\cdot|i_a) \| (P_{C'}^a)^{d_a}(\cdot|i_a)) \right] \\ &= \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K E^\pi[N(\tau(\pi), \underline{d}, \underline{i}, a)|C] \cdot D((P_C^a)^{d_a}(\cdot|i_a) \| (P_{C'}^a)^{d_a}(\cdot|i_a)), \end{aligned} \quad (88)$$

where in the above set of equations, (a) follows from Lemma 2,

$$N(n, \underline{d}, \underline{i}, a) := \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, a, j)$$

for all $n \geq K$, $(\underline{d}, \underline{i}) \in \mathbb{S}$ and $a \in \mathcal{A}$, and (88) is due to monotone convergence theorem and the fact that

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

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

$$E^\pi[Z_{CC'}^\pi(\tau(\pi))|C]$$

$$= E^\pi \left[\sum_{a=1}^K \log \frac{P_C^\pi(X_{a-1}^a)}{P_{C'}^\pi(X_{a-1}^a)} \right] + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K E^\pi [N(\tau(\pi), \underline{d}, \underline{i}, a) | C] \cdot D((P_C^a)^{d_a}(\cdot | i_a) \| (P_{C'}^a)^{d_a}(\cdot | i_a)). \quad (89)$$

Noting that

$$\begin{aligned} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K E^\pi [N(\tau(\pi), \underline{d}, \underline{i}, a) | C] &\stackrel{(a)}{=} E^\pi \left[\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K N(\tau(\pi), \underline{d}, \underline{i}, a) \middle| C \right] \\ &= E^\pi \left[\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K \sum_{t=K}^{\tau(\pi)} 1_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a\}} \middle| C \right] \\ &= E^\pi \left[\sum_{t=K}^{\tau(\pi)} 1 \middle| C \right] \end{aligned} \quad (90)$$

$$= E^\pi [\tau(\pi) - K + 1 | C], \quad (91)$$

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

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

Combining (83) and (92), and noting that (92) holds for all $C' = (h', P'_1, P'_2)$ such that $h' \neq h$, we get

$$\begin{aligned} d(\epsilon, 1 - \epsilon) &\leq \inf_{\substack{C'=(h', P'_1, P'_2): \\ h' \neq h, P'_1 \neq P'_2}} \left\{ E^\pi \left[\sum_{a=1}^K \log \frac{P_C^\pi(X_{a-1}^a)}{P_{C'}^\pi(X_{a-1}^a)} \right] \right. \\ &\quad \left. + \left(E^\pi [\tau(\pi) - K + 1 | C] \right) \cdot \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K \frac{E^\pi [N(\tau(\pi), \underline{d}, \underline{i}, a) | C]}{E^\pi [\tau(\pi) - K + 1 | C]} \cdot D((P_C^a)^{d_a}(\cdot | i_a) \| (P_{C'}^a)^{d_a}(\cdot | i_a)) \right\} \\ &\leq \sup_{\nu} \inf_{\substack{C'=(h', P'_1, P'_2): \\ h' \neq h, P'_1 \neq P'_2}} \left\{ E^\pi \left[\sum_{a=1}^K \log \frac{P_C^\pi(X_{a-1}^a)}{P_{C'}^\pi(X_{a-1}^a)} \right] + \left(E^\pi [\tau(\pi) - K + 1 | C] \right) \cdot \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K \nu(\underline{d}, \underline{i}, a) k_{CC'}(\underline{d}, \underline{i}, a) \right\}, \end{aligned} \quad (93)$$

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

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

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

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

Recall that Q in (94) denotes the transition probability matrix given by (5). 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 II, 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} . From [1, Lemma 1], we know that $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is an ergodic

Markov process under every SRS policy. This suffices to apply [1, Theorem 2] to deduce a one-one correspondence between feasible solutions to (94)-(96) and policies in Π_{SRS} . In other words, [1, Theorem 2] implies that for any given ν satisfying (94)-(96), we can find an SRS policy $\pi^\lambda \in \Pi_{\text{SRS}}$ 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 stationary distribution of the Markov process $\{(d(t), i(t)) : t \geq K\}$ is μ^λ . The associated ergodic state occupancy measure, ν^λ , is then defined according to (9).)

Using [1, Theorem 2], we may replace the supremum in (93) by a supremum over all SRS policies. Doing so leads us to the relation

$$\begin{aligned} & d(\epsilon, 1 - \epsilon) \\ & \leq \sup_{\pi^\lambda \in \Pi_{\text{SRS}}} \inf_{\substack{C' = (h', P'_1, P'_2): \\ h' \neq h, P'_1 \neq P'_2}} \left\{ E^\pi \left[\sum_{a=1}^K \log \frac{P_C^\pi(X_{a-1}^a)}{P_{C'}^\pi(X_{a-1}^a)} \right] + \left(E^\pi[\tau(\pi) - K + 1|C] \right) \cdot \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^K \nu^\lambda(\underline{d}, \underline{i}, a) k_{CC'}(\underline{d}, \underline{i}, a) \right\}. \end{aligned} \quad (97)$$

for all $\pi \in \Pi(\epsilon)$. Observe that the constant term multiplying $E^\pi[\tau(\pi) - K + 1|C]$ 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_C^a the stationary distribution of the transition probability matrix P_C^a (i.e., $\mu_C^a = \mu_1$ for $a = h$ and $\mu_C^a = \mu_2$ for all $a \neq h$). An application of the ergodic theorem to the Markov process of arm a yields

$$D((P_C^a)^{d_a}(\cdot|i_a) \parallel (P_{C'}^a)^{d_a}(\cdot|i_a)) \longrightarrow D(\mu_C^a \parallel \mu_{C'}^a) < \infty \quad \text{as } d_a \rightarrow \infty. \quad (98)$$

Since every convergent sequence is bounded, we may write $D((P_C^a)^{d_a}(\cdot|i_a) \parallel (P_{C'}^a)^{d_a}(\cdot|i_a)) \leq M$ for all $(\underline{d}, \underline{i}, a) \in \mathbb{S} \times \mathcal{A}$, where $0 < M < \infty$. Using (95), it follows that the constant term multiplying $E^\pi[\tau(\pi) - K + 1|C]$ in (97) is bounded above by M . We also note that the first term inside the braces in (97) does not depend on ϵ . Since $d(\epsilon, 1 - \epsilon) \rightarrow d(0, 1) = +\infty$ as $\epsilon \downarrow 0$, the boundedness of the constant multiplying $E^\pi[\tau(\pi) - K + 1|C]$ implies that $\epsilon \downarrow 0$ is equivalent to $E^\pi[\tau(\pi)|C] \rightarrow \infty$ for all $\pi \in \Pi(\epsilon)$. Letting $\epsilon \downarrow 0$, and using $d(\epsilon, 1 - \epsilon)/\log(1/\epsilon) \rightarrow 1$ as $\epsilon \downarrow 0$, we arrive at the lower bound in (11). This completes the proof of the proposition.

APPENDIX B

PROOF OF PROPOSITION 2

We first state the analogues of [5, Assumptions A1-A5 and A6.1] as applicable to the context of this paper and then verify each of the assumptions. Throughout this section, we let $C_0 = (h, P_1, P_2)$ denote the underlying arms configuration.

- The analogue of [5, Assumption A1] is that for all arms configuration C , $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$, and $b \in \mathcal{A}$, the transition probabilities $Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b)$ (defined in (5)) are continuous in b and C .
- The analogue of [5, Assumption A2] simply states that the actual transition probabilities are those corresponding to the underlying arms configuration C_0 .
- The analogue of [5, Assumption A3] is that if $Q_{C_0}(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b) = 0$ for some $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$ and $b \in \mathcal{A}$, then $Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b') = 0$ for all C and $b' \in \mathcal{A}$. Further, if $Q_{C_0}(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b) > 0$ for some $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$ and $b \in \mathcal{A}$, then there exists $\bar{\epsilon} > 0$ independent of $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$ and $b \in \mathcal{A}$ such that for all C , the relation

$$\bar{\epsilon} \leq \frac{Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b)}{Q_{C_0}(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b)} \leq \frac{1}{\bar{\epsilon}}$$

holds.

- The analogue of [5, Assumption A4] is that for all $b \in \mathcal{A}$ and arms configuration C , the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ with the control process $\{B_t : t \geq 0\}$ given by $B_t = b$ for all t , and transition probabilities given by $Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b)$, is a positive recurrent Markov process whose state space \mathbb{S} is a single communicating class.
- Instead of stating the analogue of [5, Assumption A5], we state the analogue of its equivalent form [25, Condition C2]; for a proof of this equivalence, we refer the reader to [25] and the references therein. The analogue of [25, Condition C2] as applicable to the context of this paper is that there is a finite set K , an integer $d \geq 1$, and a number $\rho > 0$ such that under every SRS policy π^λ , the probability of the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ starting from any state $(\underline{d}, \underline{i}) \in \mathbb{S}$ and hitting the set K after d time steps is lower bounded by ρ .
- The analogue of [5, Assumption A6.1] is that for any two arm configurations $C \neq C'$, there exists $(\underline{d}, \underline{i}) \in \mathbb{S}$ (possibly depending on C and C') such that for every $b \in \mathcal{A}$,

$$Q_C(\cdot | \underline{d}, \underline{i}, b) \neq Q_{C'}(\cdot | \underline{d}, \underline{i}, b).$$

We now proceed to verify each of the above stated assumptions. From (5), it is clear that the transition probabilities $Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b)$ are continuous in C and b . This verifies the analogue of [5, Assumption A1].

Next, we note that if $Q_{C_0}(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b) = 0$ for some $(\underline{d}', \underline{i}'), (\underline{d}, \underline{i}) \in \mathbb{S}$ and $b \in \mathcal{A}$, then it follows from (21) that one of the following conditions must hold:

- For all $a \in \mathcal{A}$,

$$\mathbb{I}_{\{d'_a=1 \text{ and } d'_{\bar{a}}=d_{\bar{a}}+1 \text{ for all } \bar{a} \neq a\}} \cdot \mathbb{I}_{\{i'_{\bar{a}}=i_{\bar{a}} \text{ for all } \bar{a} \neq a\}} = 0.$$

That is, $(\underline{d}', \underline{i}')$ is not a valid state that can be reached in one step from the state $(\underline{d}, \underline{i})$. Clearly, then, we have $Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b) = 0$ for all C and $b \in \mathcal{A}$.

- $\mathbb{I}_{\{d'_a=1 \text{ and } d'_{\bar{a}}=d_{\bar{a}}+1 \text{ for all } \bar{a} \neq a\}} \cdot \mathbb{I}_{\{i'_{\bar{a}}=i_{\bar{a}} \text{ for all } \bar{a} \neq a\}} = 1$ for some $a \in \mathcal{A}$. In this case, $(\underline{d}', \underline{i}')$ is a valid state that can be reached in one step from the state $(\underline{d}, \underline{i})$. From (5), it follows that

$$Q_{C_0}(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b) = 0 \iff (P_{C_0}^a)^{d_a}(i'_a | i_a) = 0.$$

Using Assumption 2, we have $(P_{C_0}^a)^{d_a}(i'_a | i_a) = 0 \iff (P_C^a)^{d_a}(i'_a | i_a) = 0$ for all C . Therefore, it follows that if $Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b) = 0$, then $Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b') = 0$ for all C and $b' \in \mathcal{A}$.

Now, suppose that $Q_{C_0}(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b) > 0$ for some $(\underline{d}', \underline{i}'), (\underline{d}, \underline{i}) \in \mathbb{S}$ and $b \in \mathcal{A}$. From (22), this implies that there exists $a \in \mathcal{A}$ such that $P_{C_0}^{d_a}(i'_a | i_a) > 0$. From Assumption 2, we then have $P_{C_0}^{d_a}(i'_a | i_a) > \bar{\varepsilon}^*$, which together with (22) gives us that $Q_{C_0}(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b) \geq \frac{\eta}{K} \cdot \bar{\varepsilon}^*$. Setting $\bar{\varepsilon} = \frac{\eta}{K} \cdot \bar{\varepsilon}^*$, it follows that

$$\bar{\varepsilon} \leq \frac{Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b)}{Q_{C_0}(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b)} \leq \frac{1}{\bar{\varepsilon}}.$$

This verifies the analogue of [5, Assumption A3].

Next, from [1, Lemma 1], we know that under the trembling hand model and under every SRS policy π^λ , the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is ergodic, i.e., irreducible, aperiodic and positive recurrent. This verifies the analogue of [5, Assumption A4].

To verify the analogue of [25, Condition C2], let $\underline{d}' = (K, \dots, 1)$ and $\underline{i}' = (i, \dots, i)$, where $i \in \mathcal{S}$ is a fixed state. We now show that given any SRS policy π^λ , the probability of reaching the state $(\underline{d}', \underline{i}')$ starting from any state $(\underline{d}, \underline{i})$ is lower bounded by a strictly positive constant, say $\rho > 0$, that is independent of $(\underline{d}, \underline{i})$ and the underlying arms configuration C_0 . Note that the

Markov process of each arm evolves on the common, finite state space \mathcal{S} . Therefore, from [26, Proposition 1.7], we have that there exists M sufficiently large such that

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

Fix an arbitrary $(\underline{d}, \underline{i})$, and suppose that $\underline{d}(t) = \underline{d}$ and $\underline{i}(t) = \underline{i}$ at some time $t = T_0$. Let the following sequence of arm selections and observations be obtained under π^λ : arm 1 is pulled at time $t = T_0 + 1$, arm 2 is pulled at time $t = T_0 + 2$ and so on until arm K is pulled at time $t = T_0 + K$. Thereafter, arm 1 is pulled at time $t = T_0 + M + 1$ and the state i is observed on arm 1, arm 2 is pulled at time $t = T_0 + M + 2$ and the state i is observed on arm 2, and so on until arm K is pulled at time $t = T_0 + M + K$ and the state i is observed on arm K .

Clearly, then, we have $\underline{d}(T_0 + M + K + 1) = \underline{d}'$, $\underline{i}(T_0 + M + K + 1) = \underline{i}'$. Suppose that the observations obtained from the arms $1, \dots, K$ at times $T_0 + 1, \dots, T_0 + K$ are $s_1, \dots, s_K \in \mathcal{S}$ respectively. Let $\underline{s} = (s_1, \dots, s_K)$. Then,

$$\begin{aligned} & P^{\pi^\lambda}(\underline{d}(T_0 + M + K + 1) = \underline{d}', \underline{i}(T_0 + M + K + 1) = \underline{i}' \mid \underline{d}(T_0) = \underline{d}, \underline{i}(T_0) = \underline{i}, C_0) \\ & \geq \prod_{t=T_0+1}^{T_0+K} P^{\pi^\lambda}(A_t|\underline{d}(t), \underline{i}(t), C_0) \cdot \prod_{t=T_0+M+1}^{T_0+M+K+1} P^{\pi^\lambda}(A_t|\underline{d}(t), \underline{i}(t), C_0) \cdot \prod_{a=1}^K (P_{C_0}^a)^M(i|s_a) \\ & \stackrel{(a)}{\geq} \prod_{t=T_0+1}^{T_0+K} \frac{\eta}{K} \cdot \prod_{t=T_0+M+1}^{T_0+M+K+1} \frac{\eta}{K} \cdot \prod_{a=1}^K (P_{C_0}^a)^M(i|s_a) \\ & = \left(\frac{\eta}{K}\right)^{2K} \prod_{a=1}^K (P_{C_0}^a)^M(i|s_a), \end{aligned} \tag{99}$$

where (a) above follows by observing that for any t ,

$$\begin{aligned} P^{\pi^\lambda}(A_t|\underline{d}(t), \underline{i}(t), C_0) &= \frac{\eta}{K} + (1 - \eta) \lambda(A_t|\underline{d}(t), \underline{i}(t)) \\ &\geq \frac{\eta}{K}. \end{aligned}$$

From Assumption 2 we know that $P_1, P_2 \in \mathcal{P}(\bar{\varepsilon}^*)$ for some $\bar{\varepsilon}^* > 0$. This implies that $(P_{C_0}^a)^M(i|s_a) > \bar{\varepsilon}^*$ for all $a \in \mathcal{A}$. Using this in (99), and setting $\rho = \left(\frac{\eta}{K}\right)^{2K} (\bar{\varepsilon}^*)^K$, we have

$$P^{\pi^\lambda}(\underline{d}(T_0 + M + K + 1) = \underline{d}', \underline{i}(T_0 + M + K + 1) = \underline{i}' \mid \underline{d}(T_0) = \underline{d}, \underline{i}(T_0) = \underline{i}, C_0) \geq \rho \quad \forall (\underline{d}, \underline{i}) \in \mathbb{S}. \tag{100}$$

This verifies the analogue of [25, Condition C2].

Lastly, in order to verify the analogue of [5, Assumption A6.1], we show that a condition stronger than [5, Assumption A6.1], one that is the equivalent of Mandl's identifiability condition for countable-state controlled Markov processes, holds in the context of this paper. This is demonstrated in the following lemma.

Lemma 3. *For all arm configurations $C = (h, P_1, P_2)$ and $C' = (h', P'_1, P'_2)$ such that $C \neq C'$ (i.e., C and C' differ in at least one component), there exists $(\underline{d}, \underline{i}) \in \mathbb{S}$, possibly depending on C and C' , such that*

$$Q_C(\cdot \mid \underline{d}, \underline{i}, b) \neq Q_{C'}(\cdot \mid \underline{d}, \underline{i}, b) \quad \text{for all } b \in \mathcal{A}. \tag{101}$$

Proof of Lemma 3: We present the proof of the lemma under various cases.

A. Case 1: $C = (h, P_1, P_2)$, $C' = (h', P_1, P_2)$, where $h' \neq h$

We first consider the case when C and C' differ only in their first components. Recall that the transition probability matrices P_1 and P_2 satisfy the condition $P_2 \neq P_1$. This means that there exist $i^*, j^* \in \mathcal{S}$ such that $P_1(j^*|i^*) \neq P_2(j^*|i^*)$. Fix $a^* = h$.

Let $(\underline{d}, \underline{i}) \in \mathbb{S}$ be such that $d_{a^*} = 1$ and $i_{a^*} = i^*$. Also, let $(\underline{d}', \underline{i}')$ be such that $d'_{a^*} = 1$, $d'_a = d_a + 1$ for all $a \neq a^*$, $i'_{a^*} = j^*$ and $i'_a = i_a$ for all $a \neq a^*$. Then, for the above choices of $(\underline{d}, \underline{i})$ and $(\underline{d}', \underline{i}')$, it follows from (5) that for all $b \in \mathcal{A}$,

$$\begin{aligned} Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{b=a^*\}} \right) P_1(j^* \mid i^*), \\ Q_{C'}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{b=a^*\}} \right) P_2(j^* \mid i^*). \end{aligned} \quad (102)$$

Because $P_1(j^* \mid i^*) \neq P_2(j^* \mid i^*)$, it follows that $Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) \neq Q_{C'}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b)$ for all $b \in \mathcal{A}$. This establishes (101).

B. Case 2: $C = (h, P_1, P_2)$, $C' = (h, P'_1, P_2)$, where $P_1 \neq P'_1$

The proof for this case follows along the lines of that for Case 1, with $a^* = h$ and the following modifications:

$$\begin{aligned} Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{b=a^*\}} \right) P_1(j^* \mid i^*), \\ Q_{C'}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{b=a^*\}} \right) P'_1(j^* \mid i^*) \end{aligned} \quad (103)$$

for all $b \in \mathcal{A}$, where $i^*, j^* \in \mathcal{S}$ are chosen such that $P_1(j^* \mid i^*) \neq P'_1(j^* \mid i^*)$.

C. Case 3: $C = (h, P_1, P_2)$, $C' = (h, P_1, P'_2)$, where $P_2 \neq P'_2$

The proof for this case follows along the lines of that for Case 1, with $a^* = h$ and the following modifications:

$$\begin{aligned} Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{b=a^*\}} \right) P_2(j^* \mid i^*), \\ Q_{C'}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{b=a^*\}} \right) P'_2(j^* \mid i^*) \end{aligned} \quad (104)$$

for all $b \in \mathcal{A}$, where $i^*, j^* \in \mathcal{S}$ are chosen such that $P_2(j^* \mid i^*) \neq P'_2(j^* \mid i^*)$.

D. Case 4: $C = (h, P_1, P_2)$, $C' = (h', P'_1, P_2)$, where $h' \neq h$, $P_1 \neq P'_1$

In this case, because C' is a valid arms configuration, it follows that $P'_1 \neq P_2$. The proof for this case then follows along the lines of that for Case 1, with $a^* = h'$ and the following modifications:

$$\begin{aligned} Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{b=a^*\}} \right) P_2(j^* \mid i^*), \\ Q_{C'}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{b=a^*\}} \right) P'_1(j^* \mid i^*) \end{aligned} \quad (105)$$

for all $b \in \mathcal{A}$, where $i^*, j^* \in \mathcal{S}$ are chosen such that $P'_1(j^* \mid i^*) \neq P_2(j^* \mid i^*)$.

E. Case 5: $C = (h, P_1, P_2)$, $C' = (h', P_1, P'_2)$, where $h' \neq h$, $P_2 \neq P'_2$

The proof for this case follows along the lines of that for Case 1, with $a^* \neq h, h'$ and the following modifications:

$$\begin{aligned} Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{b=a^*\}} \right) P_2(j^* \mid i^*), \\ Q_{C'}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{b=a^*\}} \right) P'_2(j^* \mid i^*) \end{aligned} \quad (106)$$

for all $b \in \mathcal{A}$, where $i^*, j^* \in \mathcal{S}$ are chosen such that $P_2(j^* \mid i^*) \neq P'_2(j^* \mid i^*)$.

F. Case 6: $C = (h, P_1, P_2)$, $C' = (h, P'_1, P'_2)$, where $P_1 \neq P'_1$, $P_2 \neq P'_2$

The proof for this case follows along the lines of that for Case 1, with $a^* = h$ and the following modifications:

$$\begin{aligned} Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{b=a^*\}} \right) P_1(j^* \mid i^*), \\ Q_{C'}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{b=a^*\}} \right) P'_1(j^* \mid i^*) \end{aligned} \quad (107)$$

for all $b \in \mathcal{A}$, where $i^*, j^* \in \mathcal{S}$ are chosen such that $P_1(j^* \mid i^*) \neq P'_1(j^* \mid i^*)$.

G. Case 7: $C = (h, P_1, P_2)$, $C' = (h', P'_1, P'_2)$, where $h' \neq h$, $P_1 \neq P'_1$, $P_2 \neq P'_2$

The proof for this case follows along the lines of that for Case 1, with $a^* = h$ and the following modifications:

$$\begin{aligned} Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{b=a^*\}} \right) P_1(j^* \mid i^*), \\ Q_{C'}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1 - \eta) \mathbb{I}_{\{b=a^*\}} \right) P'_1(j^* \mid i^*) \end{aligned} \quad (108)$$

for all $b \in \mathcal{A}$, where $i^*, j^* \in \mathcal{S}$ are chosen such that $P_1(j^* \mid i^*) \neq P'_1(j^* \mid i^*)$. This completes the proof of the lemma and also the verification of the analogue of [5, Assumption A6.1]. \blacksquare

Finally, with the analogues of [5, Assumptions A1-A5 and A6.1] being verified, we apply [5, Theorem 4.3] (which simply states that under [5, Assumptions A1-A5 and A6.1], the ML estimates converge to their true values almost surely) to deduce that under the arms configuration C_0 ,

$$\hat{P}_{h,1}(n) \longrightarrow P_1, \quad \hat{P}_{h,2}(n) \longrightarrow P_2 \quad \text{as } n \rightarrow \infty, \quad \text{almost surely,}$$

thus proving Proposition 2. We note here that in addition to [5, Assumptions A1-A5 and A6.1], the proof of [5, Theorem 4.3] uses the notion of “ $\{\varepsilon_i\}$ -randomisation” which, for controlled Markov processes, ensures that the probability of selecting any control at any given time is strictly positive. The trembling hand model (1) of our paper guarantees that the probability of sampling any arm at any given time is $\geq \frac{\eta}{K} > 0$, thus alleviating the need to consider $\{\varepsilon_i\}$ randomisation in our work.

APPENDIX C

PROOF OF PROPOSITION 3

We first note the following important points.

- From the exposition in [1, Appendix C], we note that

$$\liminf_{n \rightarrow \infty} \frac{N(n, \underline{d}, \underline{i}, a, j)}{n} > 0, \quad \liminf_{n \rightarrow \infty} \frac{N(n, \underline{d}, \underline{i}, a)}{n} > 0 \quad \text{almost surely} \quad (109)$$

for all $(\underline{d}, \underline{i}) \in \mathbb{S}$, $a \in \mathcal{A}$ and $j \in \mathcal{S}$. Therefore, by the ergodic theorem, it follows that for all $(\underline{d}, \underline{i}) \in \mathbb{S}$, $a \in \mathcal{A}$ and $j \in \mathcal{S}$ and arms configuration C , the following pointwise convergence holds almost surely:

$$\frac{N(n, \underline{d}, \underline{i}, a, j)}{N(n, \underline{d}, \underline{i}, a)} \longrightarrow (P_C^a)^{d_a}(j \mid i_a) \quad \text{as } n \rightarrow \infty. \quad (110)$$

- For each $(\underline{d}, \underline{i}) \in \mathbb{S}$ and $a \in \mathcal{A}$, the sequence

$$\left\{ \frac{N(n, \underline{d}, \underline{i}, a)}{n} \right\}_{n \geq K} \quad (111)$$

is bounded almost surely. This implies that there exists a null set B such that for every $\omega \notin B$ and $(\underline{d}, \underline{i}) \in \mathbb{S}$, there exists a subsequence, say $\{n_k(\omega, \underline{d}, \underline{i}) : k \geq 1\}$, along which the above sequence converges. Because each term of the sequence in (111) lies in the compact interval $[0, 1]$, an application Tychonoff's theorem [27, Theorem 37.3] gives us that $[0, 1]^{\mathbb{S}}$

is compact; here, $[0, 1]^{\mathbb{S}}$ denotes the countable product of $[0, 1]$ with itself. Using an Arzela-Ascoli type diagonalization argument, there exists a non-empty subsequence, call it $\{n_k(\omega) : k \geq 1\}$, along which the sequence in (111) converges for all $(\underline{d}, \underline{i}) \in \mathbb{S}$. Fixing attention to the subsequence $\{n_k(\omega) : k \geq 1\}$, $\omega \notin B$, let

$$\frac{N(n_k(\omega), \underline{d}, \underline{i}, a)(\omega)}{n_k(\omega)} \longrightarrow \alpha(\underline{d}, \underline{i}, a, \omega) \quad \text{as } k \rightarrow \infty,$$

where the limit $\alpha(\underline{d}, \underline{i}, a, \omega) > 0$ for all $(\underline{d}, \underline{i}) \in \mathbb{S}$, $a \in \mathcal{A}$, and $\omega \notin B$, thanks to (109).

Combining the above mentioned points, it follows that for all $\omega \notin B$, when C is the underlying arms configuration,

$$\frac{N(n_k(\omega), \underline{d}, \underline{i}, a, j)(\omega)}{n_k(\omega)} \longrightarrow \alpha(\underline{d}, \underline{i}, a, \omega) \cdot (P_C^a)^{d_a}(j | i_a) \quad \text{as } k \rightarrow \infty. \quad (112)$$

The following lemma shows that the convergence in (112) is, in fact, uniform in $(\underline{d}, \underline{i})$. We omit the proof of the lemma, noting that it is a simple exercise in real analysis.

Lemma 4. *Let \mathcal{S} be a countable set, and let $f_k, f : \mathcal{S} \rightarrow [0, 1]$, $k \geq 1$, be such that*

$$\sum_{s \in \mathcal{S}} f_k(s) < \infty \quad \forall k \geq 1, \quad \sum_{s \in \mathcal{S}} f(s) < \infty.$$

If $f_k \rightarrow f$ pointwise as $k \rightarrow \infty$, then $f_k \rightarrow f$ uniformly as $k \rightarrow \infty$.

We now begin the proof of Proposition 3. Fix an arbitrary null set B . Then, for every $\omega \notin B$, there exists a subsequence $\{n_k(\omega) : k \geq 1\}$ such that

$$\liminf_{n \rightarrow \infty} \frac{M_{hh'}(n)(\omega)}{n} = \lim_{k \rightarrow \infty} \frac{M_{hh'}(n_k(\omega))(\omega)}{n_k(\omega)}.$$

Let us restrict our attention on a further subsequence (obtained as described earlier using Tychonoff's theorem), and without loss of generality, let $\{n_k(\omega) : k \geq 1\}$ be this subsequence. We then have

$$\lim_{k \rightarrow \infty} \frac{M_{hh'}(n_k(\omega))(\omega)}{n_k(\omega)} = \lim_{k \rightarrow \infty} \frac{T_1(n_k(\omega)) + T_2(n_k(\omega)) + T_3(n_k(\omega)) + T_4(n_k(\omega))}{n_k(\omega)}. \quad (113)$$

We shall demonstrate that the right hand side of (113) is strictly positive for all $\omega \notin B$. Fix an arbitrary $\varepsilon \in (0, 1)$.

- From (38), we have

$$\lim_{k \rightarrow \infty} \frac{T_1(n_k(\omega))}{n_k(\omega)} = \lim_{k \rightarrow \infty} \frac{1}{n_k(\omega)} \log \mathbb{E} \left[\exp \left(\sum_{i \in \mathcal{S}} \phi(i) P^{h-1}(X_{h-1}^h | i) \right) \right] \quad (114)$$

$$+ \lim_{k \rightarrow \infty} \frac{1}{n_k(\omega)} \log \mathbb{E} \left[\exp \left(n_k(\omega) \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \log P^{d_h}(j | i_h) \right) \right], \quad (115)$$

where the expectation is with respect to $P \sim D$. Because the expectation term in (114) is finite, the limit in (114) is equal to zero. Hence, it suffices to focus only on the limit in (115). Recall that $C = (h, P_1, P_2)$ is the underlying arms configuration. We write the expectation in (115) as

$$\begin{aligned} & \mathbb{E} \left[\exp \left(n_k(\omega) \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \log P^{d_h}(j | i_h) \right) \right] \\ &= \mathbb{E} \left[\exp \left(n_k(\omega) \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \log \frac{P^{d_h}(j | i_h)}{P_1^{d_h}(j | i_h)} \right) \right. \\ & \quad \cdot \exp \left(n_k(\omega) \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \log P_1^{d_h}(j | i_h) \right) \Big], \end{aligned} \quad (116)$$

and lower bound the right hand side of (116) by

$$\mathbb{E} \left[\exp \left(n_k(\omega) \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \log \frac{P^{d_h}(j | i_h)}{P_1^{d_h}(j | i_h)} \right) \right]$$

$$\cdot \exp \left(n_k(\omega) \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \log P_1^{d_h}(j|i_h) \right) \cdot \mathbb{I} \left(P \in \mathcal{P}(\exp(-\varepsilon/|\mathcal{S}|)) \right) \Big], \quad (117)$$

noting that the expectation of a non-negative (exponential) function only reduces when evaluated over a smaller set. In (117), $\mathcal{P}(\exp(-\varepsilon/|\mathcal{S}|))$ is the set defined in (17) with $\bar{\varepsilon}^* = \exp(-\varepsilon/|\mathcal{S}|)$. When $P \in \mathcal{P}(\exp(-\varepsilon/|\mathcal{S}|))$, it follows that

$$\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \log \frac{P_1^{d_h}(j|i_h)}{P_1^{d_h}(j|i_h)} \geq -\frac{\varepsilon}{|\mathcal{S}|}, \quad (118)$$

as a result of which we have that the expectation in (117) is lower bounded by

$$\exp \left(-n_k(\omega) \cdot \frac{\varepsilon}{|\mathcal{S}|} \right) \cdot \exp \left(n_k(\omega) \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \log P_1^{d_h}(j|i_h) \right) \cdot \mathbb{P}(P \in \mathcal{P}(\exp(-\varepsilon/|\mathcal{S}|))). \quad (119)$$

Combining (116)-(119) and using these in (116), we get that

$$\begin{aligned} & \lim_{k \rightarrow \infty} \frac{T_1(n_k(\omega))}{n_k(\omega)} \\ &= \lim_{k \rightarrow \infty} \frac{1}{n_k(\omega)} \log \mathbb{E} \left[\exp \left(n_k(\omega) \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \log P_1^{d_h}(j|i_h) \right) \right] \\ &\geq \lim_{k \rightarrow \infty} \left(-\frac{\varepsilon}{|\mathcal{S}|} + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \log P_1^{d_h}(j|i_h) + \frac{1}{n_k(\omega)} \log \mathbb{P}(P \in \mathcal{P}(\exp(-\varepsilon/|\mathcal{S}|))) \right). \end{aligned} \quad (120)$$

We now claim that the last probability term in (120), which is also equal to $D(\mathcal{P}(\exp(-\varepsilon/|\mathcal{S}|)))$, is strictly positive.

Indeed, this follows by observing that

- $\mathcal{B}(\exp(-\varepsilon/|\mathcal{S}|)) := \{P \in \mathcal{P}(\mathcal{S}) : P(j|i) \geq \exp(-\varepsilon/|\mathcal{S}|) \text{ for all } i, j \in \mathcal{S}\} \subset \mathcal{P}(\exp(-\varepsilon/|\mathcal{S}|))$,
- the marginal of D along each coordinate is the Lebesgue measure, and
- $D(\mathcal{B}(\exp(-\varepsilon/|\mathcal{S}|))) > 0$ because the Lebesgue measure of the interval $[\exp(-\varepsilon/|\mathcal{S}|), 1]$ is strictly positive.

Therefore, the last term in (120) is equal to zero in the limit as $k \rightarrow \infty$. We thus have

$$\lim_{k \rightarrow \infty} \frac{T_1(n_k(\omega))}{n_k(\omega)} \geq \lim_{k \rightarrow \infty} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \log P_1^{d_h}(j|i_h) - \frac{\varepsilon}{|\mathcal{S}|}. \quad (121)$$

From (112), we know that given any $\varepsilon' > 0$, there exists $K_1 = K_1(\varepsilon', \omega)$ (the dependence of K_1 on $(\underline{d}, \underline{i})$ is removed thanks to Lemma 4) such that for all $k \geq K_1$, under the arms configuration $C = (h, P_1, P_2)$,

$$\frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \leq \alpha(\underline{d}, \underline{i}, h, \omega) P_1^{d_h}(j|i_h) + \varepsilon'.$$

Plugging this in (121), we get

$$\frac{T_1(n_k(\omega))}{n_k(\omega)} \geq \left(\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} (\alpha(\underline{d}, \underline{i}, h, \omega) P_1^{d_h}(j|i_h) + \varepsilon') \log P_1^{d_h}(j|i_h) \right) - \frac{\varepsilon}{|\mathcal{S}|} \quad (122)$$

for all $k \geq K_1$. Let ε' be chosen such that for all $k \geq K_1$,

$$\frac{T_1(n_k(\omega))}{n_k(\omega)} \geq \left(\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \alpha(\underline{d}, \underline{i}, h, \omega) P_1^{d_h}(j|i_h) \log P_1^{d_h}(j|i_h) \right) - \varepsilon. \quad (123)$$

- Similar arguments as above can be used to show that there exists $K_2 = K_2(\varepsilon, \omega)$ such that for all $k \geq K_2$, under the arms configuration $C = (h, P_1, P_2)$,

$$\frac{T_2(n_k(\omega))}{n_k(\omega)} \geq \left(\sum_{a \neq h} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \alpha(\underline{d}, \underline{i}, h, \omega) P_2^{d_a}(j|i_a) \log P_2^{d_a}(j|i_a) \right) - \varepsilon. \quad (124)$$

- We now handle the term $\frac{T_3(n_k(\omega))}{n_k(\omega)}$. Note that under the arms configuration $C = (h, P_1, P_2)$, it follows from (43) that there exists $K_3 = K_3(\varepsilon, \omega)$ such that

$$\frac{T_3(n_k(\omega))}{n_k(\omega)} \geq \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \alpha(\underline{d}, \underline{i}, h', j, \omega) P_2^{d_{h'}}(j|i_{h'}) \log \frac{1}{P_2^{d_{h'}}(j|i_{h'})} - \varepsilon \quad (125)$$

for all $k \geq K_3$ (in arriving at (125), the first term in (40), being non-negative, is lower bounded by 0). Also, there exists $K_4 = K_4(\varepsilon, \omega)$ such that

$$\frac{T_4(n_k(\omega))}{n_k(\omega)} \geq \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \alpha(\underline{d}, \underline{i}, h, \omega) P_1^{d_{h'}}(j|i_{h'}) \log \frac{1}{P_1^{d_{h'}}(j|i_{h'})} + \sum_{a \neq h, h'} \alpha(\underline{d}, \underline{i}, a, \omega) P_2^{d_a}(j|i_a) \log \frac{1}{P_2^{d_a}(j|i_a)} - \varepsilon. \quad (126)$$

for all $k \geq K_4$.

Combining the inequalities in (123)-(126), we see that for all $\varepsilon > 0$ and for all $\omega \notin B$,

$$\frac{M_{hh'}(n_k(\omega))}{n_k(\omega)} \geq \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \alpha(\underline{d}, \underline{i}, h, \omega) \cdot D(P_1^{d_h}(\cdot|i_h) \| P^{d_h}(\cdot|i_h)) + \sum_{a \neq h, h'} \alpha(\underline{d}, \underline{i}, a, \omega) \cdot D(P_2^{d_a}(\cdot|i_a) \| P^{d_a}(\cdot|i_a)) - 4\varepsilon \quad (127)$$

for all $k \geq \max\{K_1, \dots, K_4\}$, from which it follows that

$$\begin{aligned} \liminf_{n \rightarrow \infty} \frac{M_{hh'}(n)(\omega)}{n} &= \lim_{k \rightarrow \infty} \frac{M_{hh'}(n_k(\omega))}{n_k(\omega)} \\ &\geq \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \alpha(\underline{d}, \underline{i}, h, \omega) \cdot D(P_1^{d_h}(\cdot|i_h) \| P^{d_h}(\cdot|i_h)) + \sum_{a \neq h, h'} \alpha(\underline{d}, \underline{i}, a, \omega) \cdot D(P_2^{d_a}(\cdot|i_a) \| P^{d_a}(\cdot|i_a)) - 4\varepsilon. \end{aligned} \quad (128)$$

Letting $\varepsilon \downarrow 0$ in (128), we get

$$\liminf_{n \rightarrow \infty} \frac{M_{hh'}(n)(\omega)}{n} \geq \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \alpha(\underline{d}, \underline{i}, h, \omega) \cdot D(P_1^{d_h}(\cdot|i_h) \| P^{d_h}(\cdot|i_h)) + \sum_{a \neq h, h'} \alpha(\underline{d}, \underline{i}, a, \omega) \cdot D(P_2^{d_a}(\cdot|i_a) \| P^{d_a}(\cdot|i_a)) \quad (129)$$

for all $\omega \notin B$. Because $\alpha(\underline{d}, \underline{i}, a, \omega) > 0$ for all $(\underline{d}, \underline{i}) \in \mathbb{S}$, $a \in \mathcal{A}$, and $\omega \notin B$, the right hand side of (129) is strictly positive. This establishes (44).

APPENDIX D

PROOF OF PROPOSITION 4

The policy $\pi^*(L, \delta)$ 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, when $C = (h, P_1, P_2)$ is the actual (underlying) arms configuration, declares $h' \neq h$ as the true index of the odd arm.

The event in item 1 above has zero probability thanks to Proposition 3. Thus, the probability of error of policy $\pi = \pi^*(L, \delta, \gamma)$, which we denote by P_e^π , may be evaluated as follows: suppose $C = (h, P_1, P_2)$ is the underlying arms configuration. Then,

$$P_e^\pi = P^\pi(\theta(\tau(\pi)) \neq h|C) = P^\pi\left(\exists n \text{ and } h' \neq h \text{ such that } \theta(\tau(\pi)) = h' \text{ and } \tau(\pi) = n \middle| C\right). \quad (130)$$

Let $\mathcal{R}_{h'}(n) := \{\omega : \tau(\pi)(\omega) = n, \theta(n, \omega) = h'\}$ 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, the collection $\{\mathcal{R}_{h'}(n) : h' \neq h, n \geq 0\}$ is a collection of mutually disjoint sets. Therefore, we have

$$P_e^\pi = P^\pi\left(\bigcup_{h' \neq h} \bigcup_{n=0}^{\infty} \mathcal{R}_{h'}(n) \middle| C\right)$$

$$\begin{aligned}
&= \sum_{h' \neq h} \sum_{n=0}^{\infty} P^{\pi}(\tau(\pi) = n, \theta(n) = h' | C) \\
&= \sum_{h' \neq h} \sum_{n=0}^{\infty} \int_{\mathcal{R}_{h'}(n)} dP^{\pi}(\omega | C) \\
&= \sum_{h' \neq h} \sum_{n=0}^{\infty} \int_{\mathcal{R}_{h'}(n)} f(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) | C) d(B^n(\omega), A^n(\omega), \bar{X}^n(\omega)) \\
&\stackrel{(a)}{\leq} \sum_{h' \neq h} \sum_{n=0}^{\infty} \int_{\mathcal{R}_{h'}(n)} \hat{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) | H_h) d(B^n(\omega), A^n(\omega), \bar{X}^n(\omega)) \\
&= \sum_{h' \neq h} \sum_{n=0}^{\infty} \int_{\mathcal{R}_{h'}(n)} \frac{\hat{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) | H_h)}{\bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) | H_{h'})} \cdot \bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) | H_{h'}) d(B^n(\omega), A^n(\omega), \bar{X}^n(\omega)) \\
&= \sum_{h' \neq h} \sum_{n=0}^{\infty} \int_{\mathcal{R}_{h'}(n)} e^{-M_{h'h}(n)(\omega)} \bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) | H_{h'}) d(B^n(\omega), A^n(\omega), \bar{X}^n(\omega)) \\
&\stackrel{(b)}{\leq} \sum_{h' \neq h} \sum_{n=0}^{\infty} \int_{\mathcal{R}_{h'}(n)} \frac{1}{(K-1)L} \bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) | \mathcal{H}_{h'}) d(B^n(\omega), A^n(\omega), \bar{X}^n(\omega)) \\
&= \sum_{h' \neq h} \frac{1}{(K-1)L} \bar{P}^{\pi} \left(\bigcup_{n=0}^{\infty} \mathcal{R}_{h'}(n) \middle| \mathcal{H}_{h'} \right) \\
&\leq \frac{1}{L},
\end{aligned} \tag{131}$$

where (a) follows from the definition of the maximum likelihood \hat{f} , (b) follows from noting that when $\pi^*(L, \delta)$ stops at time n and outputs h' as the odd arm, we must have $M_{h'h}(n) \geq \log((K-1)L)$ almost surely, and \bar{P}^{π} in (131) denotes the probability measure under the average likelihood \bar{f} . Setting $L = 1/\epsilon$ yields the desired result.

APPENDIX E

PROOF OF PROPOSITION 5

The convergences in (50) imply that for each $a \in \mathcal{A}$ and $(\underline{d}, \underline{i}) \in \mathbb{S}$, almost surely,

$$\begin{aligned}
P(A_n = a \mid \underline{d}(n) = \underline{d}, \underline{i}(n) = \underline{i}, \mathcal{F}_{n-1}) &= \frac{\eta}{K} + (1 - \eta) \lambda_{\theta(n), \hat{P}_{\theta(n), 1}(n), \hat{P}_{\theta(n), 2}(n), \delta}(a \mid \underline{d}, \underline{i}) \\
&\longrightarrow \frac{\eta}{K} + (1 - \eta) \lambda_{h, P_1, P_2, \delta}(a \mid \underline{d}, \underline{i}) \quad \text{as } n \rightarrow \infty.
\end{aligned} \tag{132}$$

Recall that for any conditional distribution $\lambda = \lambda(\cdot | \cdot)$, the unique stationary distribution of the ergodic Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ under $\pi^{\lambda} \in \Pi_{\text{SRS}}$ is given by $\mu^{\lambda} = \{\mu^{\lambda}(\underline{d}', \underline{i}') : (\underline{d}', \underline{i}') \in \mathbb{S}\}$. We then have the following important result.

Lemma 5. *For each $(\underline{d}', \underline{i}') \in \mathbb{S}$, the mapping $\lambda \mapsto \mu^{\lambda}(\underline{d}', \underline{i}')$ is continuous.*

Proof: Note that under $\pi^{\lambda} \in \Pi_{\text{SRS}}$, the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is an ergodic Markov process whose transition probabilities under the arms configuration $C = (h, P_1, P_2)$ are given by

$$\begin{aligned}
&P^{\pi^{\lambda}}(\underline{d}(t+1) = \underline{d}', \underline{i}(t+1) = \underline{i}' \mid \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, C) \\
&= \begin{cases} \left(\frac{\eta}{K} + (1 - \eta) \lambda(a \mid \underline{d}, \underline{i}) \right) (P_C^a)^{d_a}(i'_a \mid i_a), & \text{if } d'_a = 1 \text{ and } d'_a = d_a + 1 \text{ for all } \tilde{a} \neq a, \\ & i'_a = i_a \text{ for all } \tilde{a} \neq a, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned} \tag{133}$$

It is clear that the one-step transition probabilities in (133) are continuous in λ for all $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$. As a consequence, it follows that for each $t \geq 1$, the t -step transition probabilities derived from the one-step transition probabilities in (133) are continuous in λ for all $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$. Denoting the t -step transition probability from the state $(\underline{d}, \underline{i})$ to the state $(\underline{d}', \underline{i}')$ under π^λ by $P_{(\underline{d}, \underline{i}), (\underline{d}', \underline{i}')}^t(\lambda)$, we invoke [5, Assumption A5'], an assumption that is equivalent to [5, Assumption A5], to deduce that for each $(\underline{d}', \underline{i}') \in \mathbb{S}$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n P_{(\underline{d}, \underline{i}), (\underline{d}', \underline{i}')}^t(\lambda) = \mu^\lambda(\underline{d}', \underline{i}') \quad \text{uniformly in } (\underline{d}, \underline{i}) \in \mathbb{S} \text{ and } \lambda. \quad (134)$$

Combining (a) the continuity of $P_{(\underline{d}, \underline{i}), (\underline{d}', \underline{i}')}^t(\lambda)$ in λ for all $t \geq 1$ and $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$, and (b) the uniform convergence in (134), we get that $\mu^\lambda(\underline{d}', \underline{i}')$ is continuous in λ for each $(\underline{d}', \underline{i}') \in \mathbb{S}$. This establishes the lemma. \blacksquare

Lemma 5 implies that for every $(\underline{d}, \underline{i}) \in \mathbb{S}$, there exists an open neighbourhood $O = O(\underline{d}, \underline{i})$ around (P_1, P_2) such that for all $(P, Q) \in O$,

$$\mu^{\lambda_{h,P,Q,\delta}}(\underline{d}, \underline{i}) \geq \frac{\mu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d}, \underline{i})}{1 + \delta}. \quad (135)$$

Also, Lemma 5 together with Proposition 2 implies that for each $(\underline{d}, \underline{i}) \in \mathbb{S}$, under the arms configuration $C = (h, P_1, P_2)$,

$$\mu^{\lambda_{\theta(n), \hat{P}_{\theta(n),1}(n), \hat{P}_{\theta(n),2}(n), \delta}}(\underline{d}, \underline{i}) \longrightarrow \mu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d}, \underline{i}) \quad \text{as } n \rightarrow \infty \quad \text{almost surely.} \quad (136)$$

Because $(\theta(n), \hat{P}_{\theta(n),1}(n), \hat{P}_{\theta(n),2}(n)) \longrightarrow (h, P_1, P_2)$ as $n \rightarrow \infty$ almost surely under the policy $\pi_{\text{ns}}^*(L, \delta)$ and under the arms configuration $C = (h, P_1, P_2)$, it follows that for all n sufficiently large, $(\hat{P}_{\theta(n),1}(n), \hat{P}_{\theta(n),2}(n)) = (\hat{P}_{h,1}(n), \hat{P}_{h,2}(n)) \in O$ almost surely. Invoking [5, Lemma 2.10] with the bounded cost function $c(x_m, x_{m+1}, z_m)$ therein set as $c(x_m, x_{m+1}, z_m) = 1 - \mathbb{I}_{\{x_m = (\underline{d}, \underline{i})\}}$, we get

$$\begin{aligned} \liminf_{n \rightarrow \infty} \frac{N(n, \underline{d}, \underline{i})}{n} &\geq \inf_{(P,Q) \in O} \mu^{\lambda_{h,P,Q,\delta}}(\underline{d}, \underline{i}) \\ &\geq \frac{\mu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d}, \underline{i})}{1 + \delta} \quad \text{almost surely.} \end{aligned} \quad (137)$$

Here, $N(n, \underline{d}, \underline{i}) = \sum_{a=1}^K N(n, \underline{d}, \underline{i}, a)$. Invoking the null set B from the proof of Proposition 3 in Appendix C, restricting the almost sure inequality in (137) to outside the null set B , and using the convergence in (132), we get that for all $\omega \notin B$, $(\underline{d}, \underline{i}) \in \mathbb{S}$ and $a \in \mathcal{A}$,

$$\begin{aligned} \alpha(\underline{d}, \underline{i}, a, \omega) &\geq \left(\inf_{(P,Q) \in O} \mu^{\lambda_{h,P,Q,\delta}}(\underline{d}, \underline{i}) \right) \left(\frac{\eta}{K} + (1 - \eta) \lambda_{h,P_1,P_2,\delta}(a \mid \underline{d}, \underline{i}) \right) \\ &\geq \frac{\mu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d}, \underline{i})}{1 + \delta} \left(\frac{\eta}{K} + (1 - \eta) \lambda_{h,P_1,P_2,\delta}(a \mid \underline{d}, \underline{i}) \right) \\ &= \frac{\nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d}, \underline{i}, a)}{1 + \delta}. \end{aligned} \quad (138)$$

Plugging (138) into (129), we see that under the arms configuration $C = (h, P_1, P_2)$, for all $\omega \notin B$,

$$\begin{aligned} &\liminf_{n \rightarrow \infty} \frac{M_{hh'}(n)(\omega)}{n} \\ &\geq \frac{1}{1 + \delta} \left[\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d}, \underline{i}, h) \cdot D(P_1^{d_h}(\cdot \mid i_h) \parallel P^{d_h}(\cdot \mid i_h)) + \sum_{a \neq h, h'} \nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d}, \underline{i}, a) \cdot D(P_2^{d_a}(\cdot \mid i_a) \parallel P^{d_a}(\cdot \mid i_a)) \right] \\ &\geq \frac{1}{1 + \delta} \inf_{P'_1, P'_2} \left[\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d}, \underline{i}, h) \cdot D(P_1^{d_h}(\cdot \mid i_h) \parallel (P'_2)^{d_h}(\cdot \mid i_h)) + \sum_{a \neq h, h'} \nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d}, \underline{i}, a) \cdot D(P_2^{d_a}(\cdot \mid i_a) \parallel (P'_2)^{d_a}(\cdot \mid i_a)) \right] \\ &= \frac{1}{1 + \delta} \inf_{\substack{P'_1, P'_2: \\ P'_1 \neq P'_2}} \left[\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d}, \underline{i}, h) \cdot D(P_1^{d_h}(\cdot \mid i_h) \parallel (P'_2)^{d_h}(\cdot \mid i_h)) + \nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d}, \underline{i}, h') \cdot D(P_2^{d_{h'}}(\cdot \mid i_{h'}) \parallel (P'_1)^{d_{h'}}(\cdot \mid i_{h'})) \right] \end{aligned}$$

$$\begin{aligned}
& + \sum_{a \neq h, h'} \nu^{\lambda_{h, P_1, P_2, \delta}}(\underline{d}, \underline{i}, a) \cdot D(P_2^{d_a}(\cdot | i_a) \| (P'_2)^{d_a}(\cdot | i_a)) \Big] \\
& \geq \frac{1}{1 + \delta} \inf_{\substack{C' = (h', P'_1, P'_2): \\ h' \neq h, \\ P'_1 \neq P'_2}} \left[\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \nu^{\lambda_{h, P_1, P_2, \delta}}(\underline{d}, \underline{i}, h) \cdot D(P_1^{d_h}(\cdot | i_h) \| (P'_2)^{d_h}(\cdot | i_h)) + \nu^{\lambda_{h, P_1, P_2, \delta}}(\underline{d}, \underline{i}, h') \cdot D(P_2^{d_{h'}}(\cdot | i_h) \| (P'_1)^{d_{h'}}(\cdot | i_{h'})) \right. \\
& \quad \left. + \sum_{a \neq h, h'} \nu^{\lambda_{h, P_1, P_2, \delta}}(\underline{d}, \underline{i}, a) \cdot D(P_2^{d_a}(\cdot | i_a) \| (P'_2)^{d_a}(\cdot | i_a)) \right] \\
& \geq \frac{R^*(h, P_1, P_2)}{(1 + \delta)^2}, \tag{139}
\end{aligned}$$

where the last line above follows from the definition of $\lambda_{h, P_1, P_2, \delta}$ in (16). The desired result is thus established.

APPENDIX F

PROOF OF PROPOSITION 6

It suffices to show that under the arms configuration $C = (h, P_1, P_2)$ and under the policy $\pi = \pi^*(L, \delta)$,

$$\limsup_{L \rightarrow \infty} P^\pi(\tau(\pi) \leq m \mid C) = 0 \quad \text{for all } m \geq 1. \tag{140}$$

Because the policy $\pi^*(L, \delta)$ selects arm 1 at time $t = 0$, arm 2 at time $t = 1$ and so on until arm K at time $t = K$, it suffices to show that for each $m \geq K$,

$$\limsup_{L \rightarrow \infty} P^\pi(\tau(\pi) \leq m \mid C) = 0. \tag{141}$$

Note that

$$\begin{aligned}
P^\pi(\tau(\pi) \leq m \mid C) &= P^\pi(\exists K \leq n \leq m \text{ and } \tilde{h} \text{ such that } M_{\tilde{h}}(n) \geq \log((K - 1)L)) \\
&\leq \sum_{\tilde{h} \in \mathcal{A}} \sum_{n=K}^m P^\pi(M_{\tilde{h}}(n) \geq \log((K - 1)L) \mid C) \\
&\leq \frac{1}{\log((K - 1)L)} \sum_{\tilde{h} \in \mathcal{A}} \sum_{n=K}^m E^\pi[M_{\tilde{h}}(n) \mid C], \tag{142}
\end{aligned}$$

where the first inequality above is due to the union bound, and the last line is due to Markov's inequality. We now show that for each $n \in \{K, \dots, m\}$, the expectation term $E^\pi[M_{\tilde{h}}(n) \mid C]$ is finite, from which the desired result follows. We carry out the analysis by considering the cases $\tilde{h} = h$ and $\tilde{h} \neq h$ separately. Here, h is the odd arm index in the arms configuration $C = (h, P_1, P_2)$.

A. Case $\tilde{h} = h$

In this case, we have

$$\begin{aligned}
E^\pi[M_{\tilde{h}}(n) \mid C] &= E^\pi[M_h(n) \mid C] \\
&= E^\pi \left[\min_{h' \neq h} \log \frac{\bar{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_h)}{\hat{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_{h'})} \mid C \right] \\
&\leq E^\pi \left[\log \frac{\bar{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_h)}{\hat{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_{h'})} \mid C \right] \quad \forall h' \neq h \\
&\stackrel{\text{Jensen's}}{\leq} \log E^\pi \left[\frac{\bar{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_h)}{\hat{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_{h'})} \mid C \right] \quad \forall h' \neq h
\end{aligned}$$

$$\begin{aligned}
&= \log \int_{\Omega} \frac{\bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_h)}{\hat{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_{h'})} \cdot f(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid C) dP^\pi(\omega) \quad \forall h' \neq h \\
&= \log \int_{\Omega} \frac{\bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_h)}{\hat{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_{h'})} \cdot \exp(-Z_{hh'}(n)(\omega)) \cdot f(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid C' = (h', P_1, P_2)) dP^\pi(\omega) \\
&\quad \forall h' \neq h \\
&\leq \log \int_{\Omega} \frac{\bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_h)}{\hat{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_{h'})} \cdot \exp(-Z_{hh'}(n)(\omega)) \cdot \hat{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_{h'}) dP^\pi(\omega) \quad \forall h' \neq h \\
&= \log \int_{\Omega} \bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_h) \cdot \exp(-Z_{hh'}(n)(\omega)) \cdot dP^\pi(\omega) \quad \forall h' \neq h, \tag{143}
\end{aligned}$$

where the last line follows from an application of the change of measure technique presented in [13, Lemma 18]. Also, $Z_{hh'}(n)$ in the above set of equations is given by

$$\begin{aligned}
Z_{hh'}(n) &= \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \sum_{a=1}^K N(n, \underline{d}, \underline{i}, a, j) \log \frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)} \\
&= \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h, j) \log \frac{P_1^{d_h}(j|i_h)}{P_2^{d_h}(j|i_h)} + N(n, \underline{d}, \underline{i}, h', j) \log \frac{P_2^{d_{h'}}(j|i_{h'})}{P_1^{d_{h'}}(j|i_{h'})}. \tag{144}
\end{aligned}$$

From Assumption 2, we know that there exists $\bar{\varepsilon}^* \in (0, 1)$ such that

$$\log \frac{1}{\bar{\varepsilon}^*} \geq \log \frac{P_1^d(j|i)}{P_2^d(j|i)} \geq \log \bar{\varepsilon}^* \quad \forall d \geq 1, i, j \in \mathcal{S}. \tag{145}$$

Therefore, it follows that almost surely,

$$\begin{aligned}
Z_{hh'}(n) &\geq (\log \bar{\varepsilon}^*) \cdot \left(\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h, j) + N(n, \underline{d}, \underline{i}, h', j) \right) \\
&\geq (\log \bar{\varepsilon}^*) \cdot \left(\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \sum_{a=1}^K N(n, \underline{d}, \underline{i}, a, j) \right) \tag{146}
\end{aligned}$$

$$\geq n \cdot (\log \bar{\varepsilon}^*). \tag{147}$$

Using (147) in (143), we get

$$\begin{aligned}
E^\pi[M_{\tilde{h}}(n) \mid C] &\leq \log \int_{\Omega} \bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_h) \cdot \exp(-n \cdot (\log \bar{\varepsilon}^*)) \cdot dP^\pi(\omega) \\
&= n \log \left(\frac{1}{\bar{\varepsilon}^*} \right). \tag{148}
\end{aligned}$$

B. Case $\tilde{h} \neq h$

In this case, we note that

$$\begin{aligned}
E^\pi[M_{\tilde{h}}(n) \mid C] &= E^\pi \left[\min_{h' \neq \tilde{h}} \log \frac{\bar{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_{\tilde{h}})}{\hat{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_{h'})} \middle| C \right] \\
&\leq E^\pi \left[\log \frac{\bar{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_{\tilde{h}})}{\hat{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_{\tilde{h}})} \middle| C \right] \\
&\stackrel{\text{Jensen's}}{\leq} \log E^\pi \left[\frac{\bar{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_{\tilde{h}})}{\hat{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_{\tilde{h}})} \middle| C \right] \\
&= \log \int_{\Omega} \frac{\bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_{\tilde{h}})}{\hat{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_{\tilde{h}})} \cdot f(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid C) dP^\pi(\omega)
\end{aligned}$$

$$\begin{aligned}
&\leq \log \int_{\Omega} \bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_{\bar{h}}) dP^\pi(\omega) \\
&= 0.
\end{aligned} \tag{149}$$

From (148) and (149), we have

$$\begin{aligned}
P^\pi(\tau(\pi) \leq m \mid C) &\leq \frac{1}{\log((K-1)L)} \sum_{\bar{h} \in \mathcal{A}} \sum_{n=K}^m E^\pi[M_{\bar{h}}(n) \mid C] \\
&\leq \frac{1}{\log((K-1)L)} \sum_{\bar{h} \in \mathcal{A}} \sum_{n=K}^m n \log\left(\frac{1}{\bar{\varepsilon}^*}\right),
\end{aligned} \tag{150}$$

from which it follows that for each $m \geq K$,

$$\begin{aligned}
\limsup_{L \rightarrow \infty} P(\tau(\pi^*(L, \delta)) \leq m \mid C) &\leq \limsup_{L \rightarrow \infty} \frac{1}{\log((K-1)L)} \sum_{\bar{h} \in \mathcal{A}} \sum_{n=K}^m n \log\left(\frac{1}{\bar{\varepsilon}^*}\right) \\
&= 0.
\end{aligned} \tag{151}$$

This establishes the desired result.

APPENDIX G

PROOF OF PROPOSITION 8

Here, we demonstrate that for each $\delta > 0$, the family $\{\tau(\pi^*(L, \delta))/\log L : L > 1\}$ is uniformly integrable. Clearly, it suffices to show that under the policy $\pi = \pi^*(L, \delta)$ and under the arms configuration $C = (h, P_1, P_2)$,

$$\limsup_{L \rightarrow \infty} E^\pi \left[\left(\frac{\tau(\pi)}{\log L} \right)^2 \mid C \right] < \infty. \tag{152}$$

Before showing that (152) holds, we record the following important result.

Lemma 6. Fix $C = (h, P_1, P_2)$, $L > 1$ and $\delta > 0$. There exists $0 < B < \infty$ independent of L such that for all sufficiently large values of n , under the policy $\pi = \pi^*(L, \delta)$ and under the arms configuration C ,

$$P^\pi(M_h(n) < \log((K-1)L) \mid C) \leq \frac{B}{n^3}. \tag{153}$$

Proof of Lemma 6: Note that

$$\begin{aligned}
P^\pi(M_h(n) < \log((K-1)L) \mid C) &= P^\pi \left(\min_{h' \neq h} M_{hh'}(n) < \log((K-1)L) \mid C \right) \\
&\stackrel{(a)}{\leq} \sum_{h' \neq h} P^\pi \left(M_{hh'}(n) < \log((K-1)L) \mid C \right) \\
&= \sum_{h' \neq h} P^\pi \left(\frac{M_{hh'}(n)}{n} < \frac{\log((K-1)L)}{n} \mid C \right) \\
&= \sum_{h' \neq h} P^\pi \left(\frac{T_1(n)}{n} + \frac{T_2(n)}{n} + \frac{T_3(n)}{n} + \frac{T_4(n)}{n} < \frac{\log((K-1)L)}{n} \mid C \right),
\end{aligned} \tag{154}$$

where (a) above follows from the union bound, and the last line follows from (37). In order to prove (153), it suffices to prove that each term inside the summation in (154) is $O(1/n^3)$.

By virtue of Assumption 2, we know that $P_1, P_2 \in \mathcal{P}(\bar{\varepsilon}^*)$ for some $\bar{\varepsilon}^* \in (0, 1)$. From (38), it follows that

$$\frac{T_1(n)}{n} - \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, h, j)}{n} \log P_1^{d_h}(j \mid i_h)$$

$$\begin{aligned}
&= \frac{1}{n} \log \mathbb{E} \left[\exp \left(\sum_{i \in \mathcal{S}} \phi(i) P^{h-1}(X_{h-1}^h | i) \right) \right] \\
&+ \frac{1}{n} \log \mathbb{E} \left[\exp \left(\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h, j) \log P^{d_h}(j | i_h) \right) \right] - \frac{1}{n} \log \exp \left(\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h, j) \log P_1^{d_h}(j | i_h) \right) \\
&= \frac{1}{n} \log \mathbb{E} \left[\exp \left(\sum_{i \in \mathcal{S}} \phi(i) P^{h-1}(X_{h-1}^h | i) \right) \right] + \frac{1}{n} \log \mathbb{E} \left[\exp \left(\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h, j) \log \frac{P^{d_h}(j | i_h)}{P_1^{d_h}(j | i_h)} \right) \right] \\
&\geq \frac{1}{n} \log \mathbb{E} \left[\exp \left(\sum_{i \in \mathcal{S}} \phi(i) P^{h-1}(X_{h-1}^h | i) \right) \mathbb{I} \left(P \in \mathcal{P}(\bar{\varepsilon}^*) \right) \right] \\
&\quad + \frac{1}{n} \log \mathbb{E} \left[\exp \left(\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h, j) \log \frac{P^{d_h}(j | i_h)}{P_1^{d_h}(j | i_h)} \mathbb{I} \left(P \in \mathcal{P}(\bar{\varepsilon}^*) \right) \right) \right] \\
&\stackrel{(a)}{\geq} \frac{1}{n} \log \bar{\varepsilon}^* + \frac{1}{n} \log \left(\exp \left((\log \bar{\varepsilon}^*) \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h, j) \right) \cdot \mathbb{P}(P \in \mathcal{P}(\bar{\varepsilon}^*)) \right) \\
&\stackrel{(b)}{=} \frac{N(n, h)}{n} (\log \bar{\varepsilon}^*) + \frac{1}{n} \log \bar{\varepsilon}^* + \frac{1}{n} \log \mathbb{P}(P \in \mathcal{P}(\bar{\varepsilon}^*)), \tag{155}
\end{aligned}$$

where (a) above follows by noting that on the set $\{P \in \mathcal{P}(\bar{\varepsilon}^*)\}$,

$$\forall d \geq 1 \text{ and } i, j \in \mathcal{S}, \quad P^d(j | i) \geq \bar{\varepsilon}^* \text{ whenever } P^d(j | i) > 0,$$

and in (b) above,

$$N(n, h) := \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h, j).$$

Let $N(n, a)$ be defined similarly for all $a \neq h$. It follows from (155) that almost surely

$$\frac{T_1(n)}{n} \geq \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, h, j)}{n} \log P_1^{d_h}(j | i_h) + \frac{N(n, h)}{n} (\log \bar{\varepsilon}^*) + \frac{1}{n} \log \bar{\varepsilon}^* + \frac{1}{n} \log \mathbb{P}(P \in \mathcal{P}(\bar{\varepsilon}^*)). \tag{156}$$

Similarly, it can be shown that almost surely,

$$\frac{T_2(n)}{n} \geq \sum_{a \neq h} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, a, j)}{n} \log P_2^{d_a}(j | i_a) + (\log \bar{\varepsilon}^*) \left(\sum_{a \neq h} \frac{N(n, a)}{n} \right) + \frac{K-1}{n} \log \bar{\varepsilon}^* + \frac{1}{n} \log \mathbb{P}(Q \in \mathcal{P}(\bar{\varepsilon}^*)). \tag{157}$$

Next, we note from (40) that for each $h' \neq h$, almost surely,

$$\begin{aligned}
\frac{T_3(n)}{n} &\geq \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, h', j)}{n} \log \frac{1}{(\hat{P}_{h',1}(n))^{d_{h'}}(j | i_{h'})} \\
&= \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, h', j)}{n} \log \frac{1}{P_2^{d_{h'}}(j | i_{h'})} + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, h', j)}{n} \log \frac{P_2^{d_{h'}}(j | i_{h'})}{(\hat{P}_{h',1}(n))^{d_{h'}}(j | i_{h'})} \\
&\geq \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, h', j)}{n} \log \frac{1}{P_2^{d_{h'}}(j | i_{h'})} + (\log \bar{\varepsilon}^*) \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, h', j)}{n} \\
&= \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, h', j)}{n} \log \frac{1}{P_2^{d_{h'}}(j | i_{h'})} + (\log \bar{\varepsilon}^*) \frac{N(n, h')}{n}, \tag{158}
\end{aligned}$$

where the inequality above follows as a consequence of the fact that $P_2 \in \mathcal{P}(\bar{\varepsilon}^*)$.

Lastly, in order to lower bound the term $T_4(n)/n$, let us fix an arbitrary $P_* \in \mathcal{P}(\bar{\varepsilon}^*)$ such that $P \neq P_1, P_2$. Then, for each $h' \neq h$, almost surely,

$$\frac{T_4(n)}{n} \geq \sum_{a \neq h'} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, a, j)}{n} \log \frac{1}{(\hat{P}_{h',2}(n))^{d_a}(j | i_a)}$$

$$\begin{aligned}
&= \sum_{a \neq h'} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, a, j)}{n} \log \frac{1}{P_*^{d_a}(j|i_a)} + \sum_{a \neq h'} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, a, j)}{n} \log \frac{P_*^{d_a}(j|i_a)}{(\hat{P}_{h',2}(n))^{d_a}(j|i_a)} \\
&\geq \sum_{a \neq h'} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, a, j)}{n} \log \frac{1}{P_*^{d_a}(j|i_a)} + (\log \bar{\varepsilon}^*) \sum_{a \neq h'} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, a, j)}{n} \\
&= \sum_{a \neq h'} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, a, j)}{n} \log \frac{1}{P_*^{d_a}(j|i_a)} + (\log \bar{\varepsilon}^*) \left(\sum_{a \neq h'} \frac{N(n, a)}{n} \right), \tag{159}
\end{aligned}$$

Combining the lower bounds in (156)-(159), it follows that for each $h' \neq h$, almost surely,

$$\begin{aligned}
&\frac{M_{hh'}(n)}{n} \\
&\geq 2(\log \bar{\varepsilon}^*) \left(\sum_{a=1}^K \frac{N(n, a)}{n} \right) + \frac{2}{n} \log \mathbb{P}(P \in \mathcal{P}(\bar{\varepsilon}^*)) + \frac{K}{n} \log \bar{\varepsilon}^* \\
&+ \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, h, j)}{n} \log \frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)} + \sum_{a \neq h, h'} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, a, j)}{n} \log \frac{P_2^{d_a}(j|i_a)}{P_*^{d_a}(j|i_a)} \\
&= 2(\log \bar{\varepsilon}^*) \left(\sum_{a=1}^K \frac{N(n, a)}{n} \right) + \frac{2}{n} \log \mathbb{P}(P \in \mathcal{P}(\bar{\varepsilon}^*)) + \frac{K}{n} \log \bar{\varepsilon}^* \tag{160}
\end{aligned}$$

$$+ \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \left(\frac{N(n, \underline{d}, \underline{i}, h, j)}{n} - \frac{N(n, \underline{d}, \underline{i}, h)}{n} P_1^{d_h}(j|i_h) \right) \log \frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)} \tag{161}$$

$$+ \sum_{a \neq h, h'} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \left(\frac{N(n, \underline{d}, \underline{i}, a, j)}{n} - \frac{N(n, \underline{d}, \underline{i}, a)}{n} P_2^{d_a}(j|i_a) \right) \log \frac{P_2^{d_a}(j|i_a)}{P_*^{d_a}(j|i_a)} \tag{162}$$

$$+ \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, h)}{n} D(P_1^{d_h}(\cdot|i_h) \| P_*^{d_h}(\cdot|i_h)) + \sum_{a \neq h, h'} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, a)}{n} D(P_2^{d_a}(\cdot|i_a) \| P_*^{d_a}(\cdot|i_a)). \tag{163}$$

Using (160)-(163), the probability term $P^\pi \left(\frac{M_{hh'}(n)}{n} < \frac{\log((K-1)L)}{n} \mid C \right)$ may be written as

$$P^\pi \left(\frac{M_{hh'}(n)}{n} < \frac{\log((K-1)L)}{n} \mid C \right) \leq U_1(n) + U_2(n) + U_3(n) + U_4(n), \tag{164}$$

where the terms $U_1(n)$ - $U_4(n)$ are as follows:

- The term $U_1(n)$ is given by

$$U_1(n) = P^\pi \left(\frac{K}{n} \log \bar{\varepsilon}^* + \frac{2}{n} \log \mathbb{P}(P \in \mathcal{P}(\bar{\varepsilon}^*)) < -\varepsilon \mid C \right). \tag{165}$$

- The term $U_2(n)$ is given by

$$U_2(n) = P^\pi \left(\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \left(\frac{N(n, \underline{d}, \underline{i}, h, j)}{n} - \frac{N(n, \underline{d}, \underline{i}, h)}{n} P_1^{d_h}(j|i_h) \right) \log \frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)} < -\varepsilon \mid C \right). \tag{166}$$

- The term $U_3(n)$ is given by

$$U_3(n) = P^\pi \left(\sum_{a \neq h, h'} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \left(\frac{N(n, \underline{d}, \underline{i}, a, j)}{n} - \frac{N(n, \underline{d}, \underline{i}, a)}{n} P_2^{d_a}(j|i_a) \right) \log \frac{P_2^{d_a}(j|i_a)}{P_*^{d_a}(j|i_a)} < -\varepsilon \mid C \right). \tag{167}$$

- The term $U_4(n)$ is given by

$$\begin{aligned}
U_4(n) &= P^\pi \left(\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \frac{N(n, \underline{d}, \underline{i}, h)}{n} D(P_1^{d_h}(\cdot|i_h) \| P_*^{d_h}(\cdot|i_h)) \right. \\
&\quad \left. + \sum_{a \neq h, h'} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \frac{N(n, \underline{d}, \underline{i}, a)}{n} D(P_2^{d_a}(\cdot|i_a) \| P_*^{d_a}(\cdot|i_a)) \right)
\end{aligned}$$

$$+ 2(\log \bar{\varepsilon}^*) \left(\sum_{a=1}^K \frac{N(n, a)}{n} \right) + 3\varepsilon < \frac{\log((K-1)L)}{n} \Big| C \Big). \quad (168)$$

The relations in (165)-(168) hold for all $\varepsilon > 0$. We shall shortly demonstrate how to choose $\varepsilon > 0$.

We now show that each of the terms $U_1(n)$ - $U_4(n)$ is either 0 or $O(1/n^3)$. This will then establish (153).

A. Handling $U_1(n)$

We begin by noting that inside the probability term in (165), the left hand side goes to zero whenever $\mathbb{P}(P \in \mathcal{P}(\bar{\varepsilon}^*)) > 0$, whereas the right hand side is strictly negative. Noting that

$$\mathcal{B}(\bar{\varepsilon}^*) := \left\{ P \in \mathcal{P}(\mathcal{S}) : P(j|i) \geq \bar{\varepsilon}^* \text{ for all } i, j \in \mathcal{S} \right\}$$

is a subset of $\mathcal{P}(\bar{\varepsilon}^*)$, and replacing $\bar{\varepsilon}^*$ by $\bar{\varepsilon}^*/|\mathcal{S}|$ if necessary³, we see that

$$\begin{aligned} \mathbb{P}(P \in \mathcal{P}(\bar{\varepsilon}^*)) &\geq \mathbb{P}(\mathcal{B}(\bar{\varepsilon}^*)) \\ &= D(\mathcal{B}(\bar{\varepsilon}^*)) \\ &> 0, \end{aligned} \quad (169)$$

where the strict positivity in the last line above follows from the fact that the marginal of D along each coordinate is the Lebesgue measure, and that the Lebesgue measure of $\mathcal{B}(\bar{\varepsilon}^*)$ is $(1 - \bar{\varepsilon}^*)^{|\mathcal{S}|} > 0$. Therefore, it follows that there exists $N_1 = N_1(\varepsilon)$ such that $U_1(n) = 0$ for all $n \geq N_1$.

B. Handling $U_2(n)$

Next, we note that

$$\left\{ \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \left(N(n, \underline{d}, \underline{i}, h, j) - N(n, \underline{d}, \underline{i}, h) P_1^{d_h}(j|i_h) \right) \log \frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)} \right\}_{n \geq K}$$

is a martingale. Indeed, because $\log \bar{\varepsilon}^* \leq \log \frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)} \leq \log \frac{1}{\bar{\varepsilon}^*}$ (which follows as a consequence of Assumption 2), using the dominated convergence theorem,

$$\begin{aligned} &E^\pi \left[\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \left(N(n, \underline{d}, \underline{i}, h, j) - N(n, \underline{d}, \underline{i}, h) P_1^{d_h}(j|i_h) \right) \log \frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)} \Big| B^{n-1}, A^{n-1}, \bar{X}^{n-1}, C \right] \\ &= E^\pi \left[\sum_{t=K}^n \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \mathbb{I}_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=h\}} \left(\mathbb{I}_{\{\bar{X}_t=j\}} - P_1^{d_h}(j|i_h) \right) \log \frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)} \Big| B^{n-1}, A^{n-1}, \bar{X}^{n-1}, C \right] \\ &= \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \left(N(n-1, \underline{d}, \underline{i}, h, j) - N(n-1, \underline{d}, \underline{i}, h) P_1^{d_h}(j|i_h) \right) \log \frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)} \\ &+ \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \mathbb{I}_{\{\underline{d}(n)=\underline{d}, \underline{i}(n)=\underline{i}\}} \left(P^\pi(A_n = h, \bar{X}_n = j | B^{n-1}, A^{n-1}, \bar{X}^{n-1}, C) \right. \\ &\quad \left. - P^\pi(A_n = h | B^{n-1}, A^{n-1}, \bar{X}^{n-1}, C) \cdot P_1^{d_h}(j|i_h) \right) \log \frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)} \\ &= \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \left(N(n-1, \underline{d}, \underline{i}, h, j) - N(n-1, \underline{d}, \underline{i}, h) P_1^{d_h}(j|i_h) \right) \log \frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)}, \end{aligned} \quad (170)$$

³Note that $\mathcal{P}(\bar{\varepsilon}^*) \subset \mathcal{P}(\bar{\varepsilon}^*/|\mathcal{S}|)$.

where the last line follows by noting that when $(\underline{d}(t), \underline{i}(t)) = (\underline{d}, \underline{i})$, under the arms configuration $C = (h, P_1, P_2)$,

$$P^\pi(\bar{X}_t = j | B^{n-1}, A^{n-1}, \bar{X}^{n-1}, C) = P_1^{d_h}(j | i_h).$$

Further, above martingale is bounded, and its quadratic variation $\langle M_n \rangle$ satisfies

$$\begin{aligned} \langle M_n \rangle &:= \sum_{t=K}^n E^\pi \left[\left(\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \mathbb{I}_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=h\}} \left(\mathbb{I}_{\{\bar{X}_t=j\}} - P_1^{d_h}(j | i_h) \right) \log \frac{P_1^{d_h}(j | i_h)}{P_*^{d_h}(j | i_h)} \right)^2 \middle| B^{t-1}, A^{t-1}, \bar{X}^{t-1}, C \right] \\ &\leq \sum_{t=K}^n E^\pi \left[\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \mathbb{I}_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=h\}} \left(\mathbb{I}_{\{\bar{X}_t=j\}} - P_1^{d_h}(j | i_h) \right)^2 \left(\log \frac{P_1^{d_h}(j | i_h)}{P_*^{d_h}(j | i_h)} \right)^2 \middle| B^{t-1}, A^{t-1}, \bar{X}^{t-1}, C \right] \\ &\stackrel{(a)}{\leq} 4 \left(\log \frac{1}{\bar{\varepsilon}^*} \right)^2 \sum_{t=K}^n E^\pi \left[\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \mathbb{I}_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=h\}} \middle| C \right] \\ &\leq n \left(4 \left(\log \frac{1}{\bar{\varepsilon}^*} \right)^2 |\mathcal{S}| \right) \end{aligned} \tag{171}$$

almost surely. In arriving at (a) above, we use the fact that for all $d \geq 1$ and $i, j \in \mathcal{S}$,

$$\log \frac{P_1^d(j | i)}{P_*^d(j | i)} \leq \log \frac{1}{\bar{\varepsilon}^*}.$$

We then have

$$\begin{aligned} U_2(n) &= P^\pi \left(\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \left(N(n, \underline{d}, \underline{i}, h, j) - N(n, \underline{d}, \underline{i}, h) P_1^{d_h}(j | i_h) \right) \log \frac{P_1^{d_h}(j | i_h)}{P_*^{d_h}(j | i_h)} < -n\varepsilon \middle| C \right) \\ &= P^\pi \left(\left| \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \left(N(n, \underline{d}, \underline{i}, h, j) - N(n, \underline{d}, \underline{i}, h) P_1^{d_h}(j | i_h) \right) \log \frac{P_1^{d_h}(j | i_h)}{P_*^{d_h}(j | i_h)} \right| > n\varepsilon \middle| C \right) \\ &\leq P^\pi \left(\sup_{K \leq t \leq n} \left| \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \left(N(t, \underline{d}, \underline{i}, h, j) - N(t, \underline{d}, \underline{i}, h) P_1^{d_h}(j | i_h) \right) \log \frac{P_1^{d_h}(j | i_h)}{P_*^{d_h}(j | i_h)} \right| > n\varepsilon \middle| C \right) \\ &\stackrel{(a)}{\leq} \frac{1}{n^6 \varepsilon^6} E^\pi \left[\left(\sup_{K \leq t \leq n} \left| \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \left(N(t, \underline{d}, \underline{i}, h, j) - N(t, \underline{d}, \underline{i}, h) P_1^{d_h}(j | i_h) \right) \log \frac{P_1^{d_h}(j | i_h)}{P_*^{d_h}(j | i_h)} \right| \right)^6 \middle| C \right] \\ &\stackrel{(b)}{\leq} \frac{A}{n^6 \varepsilon^6} E^\pi [|\langle M_n \rangle|^3 | C] \\ &\stackrel{(c)}{\leq} \frac{A}{n^6 \varepsilon^6} n^3 \left(4 \left(\log \frac{1}{\bar{\varepsilon}^*} \right)^2 |\mathcal{S}| \right)^3 \\ &= \frac{A'}{n^3}, \end{aligned} \tag{172}$$

where (a) above is due to Markov's inequality, (b) is due to Burkholder inequality [28, p. 414], and (c) follows from (171).

The constant A in (b) above comes from Burkholder inequality. We have thus shown that $U_2(n)$ is $O(1/n^3)$.

C. Handling $U_3(n)$

Following the arguments presented for handling $U_2(n)$, it can be shown that $U_3(n) = O(1/n^3)$. The details are omitted.

D. Handling $U_4(n)$

We first note that

$$U_4(n) \leq P^\pi \left(\sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \frac{N(n, \underline{d}, \underline{i}, h)}{n} D(P_1^{d_h}(\cdot | i_h) \| P_*^{d_h}(\cdot | i_h)) + \sum_{a \neq h, h'} \frac{N(n, \underline{d}, \underline{i}, a)}{n} D(P_2^{d_a}(\cdot | i_a) \| P_*^{d_a}(\cdot | i_a)) \right)$$

$$\begin{aligned}
& + 2(\log \bar{\varepsilon}^*) \left(\sum_{a=1}^K \frac{N(n, a)}{n} \right) + 3\varepsilon < \frac{\log((K-1)L)}{n} \Bigg| C \Bigg) \\
\leq & P^\pi \left(\frac{N(n, \underline{d}, \underline{i}, h)}{n} D(P_1^{d_h}(\cdot|i_h) \| P_*^{d_h}(\cdot|i_h)) + \sum_{a \neq h, h'} \frac{N(n, \underline{d}, \underline{i}, a)}{n} D(P_2^{d_a}(\cdot|i_a) \| P_*^{d_a}(\cdot|i_a)) \right. \\
& \left. + 2(\log \bar{\varepsilon}^*) + 3\varepsilon < \frac{\log((K-1)L)}{n} \Bigg| C \right) \quad (173)
\end{aligned}$$

for all $(\underline{d}, \underline{i}) \in \mathbb{S}$. Going further, let us fix $\underline{d} = (K, K-1, \dots, 1)$ and $\underline{i} = (i^*, i^*, \dots, i^*)$ for some $i^* \in \mathbb{S}$. Let E_n be the event inside the probability term in (173). From (138), we know that almost surely,

$$\liminf_{n \rightarrow \infty} \frac{N(n, \underline{d}, \underline{i}, a)}{n} \geq \frac{\nu^{\lambda_h, P_1, P_2, \delta}(\underline{d}, \underline{i}, a)}{1 + \delta}. \quad (174)$$

Exploiting this, and denoting the right hand side of (174) as $C(\underline{d}, \underline{i}, a)$, the probability term in (173) may be upper bounded as

$$P^\pi(E_n|C) \leq P^\pi \left(E_n, \left\{ \frac{N(n, \underline{d}, \underline{i}, a)}{n} \geq C(\underline{d}, \underline{i}, a)(1 - \epsilon') \quad \forall a \in \mathcal{A} \right\} \Bigg| C \right) \quad (175)$$

$$+ P^\pi \left(E_n, \left\{ \exists a \in \mathcal{A} : \frac{N(n, \underline{d}, \underline{i}, a)}{n} < C(\underline{d}, \underline{i}, a)(1 - \epsilon') \right\} \Bigg| C \right) \quad (176)$$

for all $\epsilon' \in (0, 1)$. We shall soon specify how to choose ϵ' . The probability term in (175) may then be upper bounded by

$$\begin{aligned}
P^\pi \left((1 - \epsilon') \left(C(\underline{d}, \underline{i}, h) D(P_1^{d_h}(\cdot|i_h) \| P_*^{d_h}(\cdot|i_h)) + \sum_{a \neq h, h'} C(\underline{d}, \underline{i}, a) D(P_2^{d_a}(\cdot|i_a) \| P_*^{d_a}(\cdot|i_a)) \right) \right. \\
\left. + 2(\log \bar{\varepsilon}^*) + 3\varepsilon < \frac{\log((K-1)L)}{n} \Bigg| C \right). \quad (177)
\end{aligned}$$

Let $\varepsilon > 0$ be chosen such that

$$3\varepsilon > 2 \log \frac{1}{\bar{\varepsilon}^*} - (1 - \epsilon') \left(C(\underline{d}, \underline{i}, h) D(P_1^{d_h}(\cdot|i_h) \| P_*^{d_h}(\cdot|i_h)) + \sum_{a \neq h, h'} C(\underline{d}, \underline{i}, a) D(P_2^{d_a}(\cdot|i_a) \| P_*^{d_a}(\cdot|i_a)) \right). \quad (178)$$

Such a choice of ε is possible because the right hand side of (178) is strictly positive by virtue of the fact that

$$D(P_1^{d_h}(\cdot|i_h) \| P_*^{d_h}(\cdot|i_h)) \leq \log \frac{1}{\bar{\varepsilon}^*}, \quad D(P_2^{d_a}(\cdot|i_a) \| P_*^{d_a}(\cdot|i_a)) \leq \log \frac{1}{\bar{\varepsilon}^*}$$

for all $a \neq h, h'$. Therefore,

$$\begin{aligned}
& (1 - \epsilon') \left(C(\underline{d}, \underline{i}, h) D(P_1^{d_h}(\cdot|i_h) \| P_*^{d_h}(\cdot|i_h)) + \sum_{a \neq h, h'} C(\underline{d}, \underline{i}, a) D(P_2^{d_a}(\cdot|i_a) \| P_*^{d_a}(\cdot|i_a)) \right) \\
& \leq \left(\log \frac{1}{\bar{\varepsilon}^*} \right) (1 - \epsilon') \sum_{a \neq h'} C(\underline{d}, \underline{i}, a) \\
& \leq \log \frac{1}{\bar{\varepsilon}^*}. \quad (179)
\end{aligned}$$

For instance, it suffices to choose

$$\varepsilon = \frac{2}{3} \left(2 \log \frac{1}{\bar{\varepsilon}^*} - (1 - \epsilon') \left(C(\underline{d}, \underline{i}, h) D(P_1^{d_h}(\cdot|i_h) \| P_*^{d_h}(\cdot|i_h)) + \sum_{a \neq h, h'} C(\underline{d}, \underline{i}, a) D(P_2^{d_a}(\cdot|i_a) \| P_*^{d_a}(\cdot|i_a)) \right) \right). \quad (180)$$

With ε as chosen above, it follows that inside the probability term in (177), the left hand side is strictly positive whereas the right hand side goes to 0 as $n \rightarrow \infty$. Therefore, for all sufficiently large values of n , the probability term in (177) is equal to zero.

We now turn attention to the probability term in (176). Using the union bound, this term may be upper bounded by

$$\sum_{a=1}^K P^\pi \left(\frac{N(n, \underline{d}, \underline{i}, a)}{n} < C(\underline{d}, \underline{i}, a)(1 - \epsilon') \Bigg| C \right) = \sum_{a=1}^K P^\pi \left(N(n, \underline{d}, \underline{i}, a) < n C(\underline{d}, \underline{i}, a)(1 - \epsilon') \Bigg| C \right). \quad (181)$$

In order to complete the proof, it suffices to show that each term inside the summation in (181) is $O(1/n^3)$ for a suitable choice of $\epsilon' > 0$.

Fix an arbitrary $a \in \mathcal{A}$. From the exposition in Appendix B, we know that the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ has the property that for all $T_0 \geq K$, the relation

$$P^\pi(\underline{d}(T_0 + M + K + 1) = \underline{d}, \underline{i}(T_0 + M + K + 1) = \underline{i} \mid \underline{d}(T_0) = \underline{d}'', \underline{i}(T_0) = \underline{i}'', C) \geq \rho \quad \forall (\underline{d}'', \underline{i}'') \in \mathbb{S}, \quad (182)$$

where $\rho \geq \left(\frac{\eta}{K}\right)^{2K} \cdot (\bar{\epsilon}^*)^K > 0$, and M is a positive integer that satisfies

$$P_1^M(j|i) > 0, \quad P_2^M(j|i) > 0 \quad \forall i, j \in \mathcal{S}. \quad (183)$$

The existence of such an integer M is guaranteed by [24, Proposition 1.7]. From (182), it follows that

$$P^\pi(\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} \mid C) \geq \left(\frac{\eta}{K}\right)^{2K} (\bar{\epsilon}^*)^K \quad \forall t \geq M + 3K + 1. \quad (184)$$

Additionally, we know that under the policy $\pi = \pi^*(L, \delta)$,

$$P^\pi(A_t = a \mid \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, B^{t-1}, A^{t-1}, \bar{X}^{t-1}, C) \geq \frac{\eta}{K} \quad \forall t \geq K. \quad (185)$$

As a consequence of (185), it follows that $P^\pi(A_t = a \mid \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, C) \geq \frac{\eta}{K}$ for all $t \geq K$. Combining this with (184), we see that for all $t \geq M + 3K + 1$,

$$P^\pi(\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, A_t = a \mid C) \geq \left(\frac{\eta}{K}\right)^{2K+1} (\bar{\epsilon}^*)^K. \quad (186)$$

Clearly, for all $n \geq M + 3K + 1$,

$$\begin{aligned} N(n, \underline{d}, \underline{i}, a) &= \sum_{t=K}^{M+2K} \mathbb{I}_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a\}} + \sum_{t=K+M+2K+1}^n \mathbb{I}_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a\}} \\ &\geq \sum_{t=K+M+2K+1}^n \mathbb{I}_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a\}} \end{aligned} \quad (187)$$

almost surely. Denoting the right hand side of (187) as $N'(n, \underline{d}, \underline{i}, a)$, it follows from (186) that

$$E^\pi[N'(n, \underline{d}, \underline{i}, a) \mid C] \geq (n - M - 3K) \left(\frac{\eta}{K}\right)^{2K+1} (\bar{\epsilon}^*)^K \quad \forall n \geq M + 3K + 1. \quad (188)$$

Therefore, for all $n \geq M + 3K + 1$, we have

$$\begin{aligned} &P^\pi \left(N(n, \underline{d}, \underline{i}, a) < n C(\underline{d}, \underline{i}, a)(1 - \epsilon') \mid C \right) \\ &\leq P^\pi \left(N'(n, \underline{d}, \underline{i}, a) < n C(\underline{d}, \underline{i}, a)(1 - \epsilon') \mid C \right) \\ &= P^\pi \left(N'(n, \underline{d}, \underline{i}, a) - E^\pi[N'(n, \underline{d}, \underline{i}, a) \mid C] < n C(\underline{d}, \underline{i}, a)(1 - \epsilon') - E^\pi[N'(n, \underline{d}, \underline{i}, a) \mid C] \mid C \right) \\ &\leq P^\pi \left(N'(n, \underline{d}, \underline{i}, a) - E^\pi[N'(n, \underline{d}, \underline{i}, a) \mid C] < n C(\underline{d}, \underline{i}, a)(1 - \epsilon') - (n - M - 3K) \left(\frac{\eta}{K}\right)^{2K+1} (\bar{\epsilon}^*)^K \mid C \right) \\ &\leq P^\pi \left(N'(n, \underline{d}, \underline{i}, a) - E^\pi[N'(n, \underline{d}, \underline{i}, a) \mid C] < n \left(C(\underline{d}, \underline{i}, a)(1 - \epsilon') - \left(\frac{n - M - 3K}{n}\right) \left(\frac{\eta}{K}\right)^{2K+1} (\bar{\epsilon}^*)^K \right) \mid C \right) \\ &\leq P^\pi \left(N'(n, \underline{d}, \underline{i}, a) - E^\pi[N'(n, \underline{d}, \underline{i}, a) \mid C] < n \left(\frac{1 - \epsilon'}{1 + \delta} - \left(\frac{n - M - 3K}{n}\right) \left(\frac{\eta}{K}\right)^{2K+1} (\bar{\epsilon}^*)^K \right) \mid C \right), \end{aligned} \quad (189)$$

where the last line follows by noting that

$$\begin{aligned} C(\underline{d}, \underline{i}, a) &= \frac{\nu^{\lambda_h, P_1, P_2, \delta}(\underline{d}, \underline{i}, a)}{1 + \delta} \\ &< \frac{1}{1 + \delta}. \end{aligned}$$

We now show that for a suitable choice of ϵ' , the right hand side of the probability term in (189) can be made negative. Because $(n - M - 3K)/n \rightarrow 1$ as $n \rightarrow \infty$, it follows that there exists $N_2 = N_2(\delta)$ such that for all $n \geq N_2$,

$$\frac{n - M - 3K}{n} > \frac{1}{1 + \delta},$$

as a consequence of which

$$\left(\frac{n - M - 3K}{n}\right) \left(\frac{\eta}{K}\right)^{2K+1} (\bar{\epsilon}^*)^K > \frac{\left(\frac{\eta}{K}\right)^{2K+1} (\bar{\epsilon}^*)^K}{1 + \delta}$$

for all $n \geq N_2$. Let ϵ' be chosen such that

$$1 - \epsilon' < \left(\frac{\eta}{K}\right)^{2K+1} (\bar{\epsilon}^*)^K. \quad (190)$$

Such a choice of ϵ' is possible since

$$0 < \left(\frac{\eta}{K}\right)^{2K+1} (\bar{\epsilon}^*)^K < 1.$$

For instance, it suffices to choose

$$\epsilon' = 1 - \frac{1}{3} \cdot \left(\frac{\eta}{K}\right)^{2K+1} (\bar{\epsilon}^*)^K. \quad (191)$$

With this choice of ϵ' , it follows that for all $n \geq \max\{N_2, M + 3K + 1\}$,

$$P^\pi \left(N(n, \underline{d}, \underline{i}, a) < n C(\underline{d}, \underline{i}, a)(1 - \epsilon') \mid C \right) \leq P^\pi \left(N'(n, \underline{d}, \underline{i}, a) - E^\pi[N'(n, \underline{d}, \underline{i}, a) | C] < -n\Delta \mid C \right), \quad (192)$$

where Δ is given by

$$\Delta = \frac{2}{3} \cdot \frac{\left(\frac{\eta}{K}\right)^{2K+1} (\bar{\epsilon}^*)^K}{1 + \delta}.$$

We now demonstrate that $N'(n, \underline{d}, \underline{i}, a)$ is sub-gaussian. Subsequently, we use sub-gaussian concentration bounds to show that the right hand side of (192) is bounded above exponentially. Recall that a random variable Z is said to be sub-gaussian with variance factor v [29, Section 2.3] if

$$\log E \left[\exp \left(\lambda(Z - E[Z]) \right) \right] \leq \frac{\lambda^2 v}{2} \quad \forall \lambda \in \mathbb{R}. \quad (193)$$

It is well known that if Z is Bernoulli distributed, then Z is sub-gaussian with variance factor $v = 1/4$. The below lemma demonstrates that given sub-gaussian random variables X and Y (not necessarily independent), their sum is also sub-gaussian.

Lemma 7. *Suppose X is sub-gaussian with variance factor σ_1 and Y (not necessarily independent of X) is sub-gaussian with variance factor σ_2 . Then, $X + Y$ is sub-gaussian with variance factor $\sigma_1 + \sigma_2$.*

Proof of Lemma 7: Using Holder's inequality, for all $p > 1$, we have

$$\begin{aligned} E \left[\exp \left(\lambda(X + Y - (E[X] + E[Y])) \right) \right] &\leq \left(E \left[\exp \left(p \lambda(X - E[X]) \right) \right] \right)^{\frac{1}{p}} \cdot \left(E \left[\exp \left((1 - p) \lambda(Y - E[Y]) \right) \right] \right)^{\frac{1}{1-p}} \\ &\leq \left(\exp \left(\frac{\lambda^2 p^2 \sigma_1}{2} \right) \right)^{\frac{1}{p}} \cdot \left(\exp \left(\frac{\lambda^2 (1 - p)^2 \sigma_2}{2} \right) \right)^{\frac{1}{1-p}} \\ &= \exp \left(\frac{\lambda^2 (\sigma_2 + p(\sigma_1 - \sigma_2))}{2} \right). \end{aligned} \quad (194)$$

In particular, for $p = 1 + \sigma_2/\sigma_1$, we have

$$\begin{aligned} E \left[\exp \left(\lambda(X + Y - (E[X] + E[Y])) \right) \right] &\leq \exp \left(\frac{\lambda^2}{2} \left(\sigma_2 + \frac{\sigma_1^2 - \sigma_2^2}{\sigma_1} \right) \right) \\ &\leq \exp \left(\frac{\lambda^2 (\sigma_1 + \sigma_2)}{2} \right). \end{aligned} \quad (195)$$

This establishes the desired result. \blacksquare

Using Lemma 7 and the sub-gaussian property of a Bernoulli random variable mentioned earlier, it follows that $N'(n, \underline{d}, \underline{i}, a)$ is sub-gaussian with variance factor

$$v^* = \frac{n - M - 3K}{4}.$$

Using concentration bounds for sub-gaussian random variables [29, p. 25], we get

$$P^\pi \left(N'(n, \underline{d}, \underline{i}, a) - E^\pi[N'(n, \underline{d}, \underline{i}, a)|C] < -n\Delta \mid C \right) \leq \exp \left(-\frac{n^2 \Delta^2}{2v^*} \right) \leq \exp(-2n\Delta^2). \quad (196)$$

We know that there exists N_3 such that for all $n \geq N_3$,

$$\exp(-2n\Delta^2) \leq 1/n^3.$$

Therefore, it follows that

$$P^\pi \left(N'(n, \underline{d}, \underline{i}, a) - E^\pi[N'(n, \underline{d}, \underline{i}, a)|C] < -n\Delta \mid C \right) \leq \frac{1}{n^3} \quad \forall n \geq \max\{N_2, N_3, M + 3K + 1\}.$$

Combining (172) and the above result, and choosing A' in (172) large if needed, establishes the desired result. This completes the proof of the lemma. \blacksquare

Proof of (152): We now use the result established in Lemma 6 to show (152). Let

$$u(L) := \frac{\log((K-1)L)}{2 \log \frac{1}{\epsilon'} - (1 - \epsilon') \left(C(\underline{d}, \underline{i}, h) D(P_1^{d_h}(\cdot|i_h) \| P_*^{d_h}(\cdot|i_h)) + \sum_{a \neq h, h'} C(\underline{d}, \underline{i}, a) D(P_2^{d_a}(\cdot|i_a) \| P_*^{d_a}(\cdot|i_a)) \right)}, \quad (197)$$

where ϵ' is as given in (191). Let $\psi(L) = \max\{u(L), N_1, N_2, N_3, M + 3K + 1\}$, where the constants N_1, N_2, N_3 are as determined in the proof of Lemma 6. Then, we have

$$\begin{aligned} & E^\pi \left[\left(\frac{\tau(\pi)}{\log L} \right)^2 \mid C \right] \\ &= \int_0^\infty P^\pi \left(\left(\frac{\tau(\pi)}{\log L} \right)^2 > x \mid C \right) dx \\ &= \int_0^\infty P^\pi \left(\frac{\tau(\pi)}{\log L} > \sqrt{x} \mid C \right) dx \\ &= \int_0^\infty P^\pi \left(\tau(\pi) \geq \lceil \sqrt{x} \log L \rceil \mid C \right) dx \\ &= \int_0^{\psi(L)} P^\pi \left(\tau(\pi) \geq \lceil \sqrt{x} \log L \rceil \mid C \right) dx + \int_{\psi(L)}^\infty P^\pi \left(\tau(\pi) \geq \lceil \sqrt{x} \log L \rceil \mid C \right) dx \\ &= \int_0^{\sqrt{\psi(L) \log L}} P^\pi \left(\tau(\pi) > \lceil t \log L \rceil \mid C \right) 2t dt + \int_{\psi(L)}^\infty P^\pi \left(\tau(\pi) \geq \lceil \sqrt{x} \log L \rceil \mid C \right) dx \\ &= \frac{1}{(\log L)^2} \int_0^{\sqrt{\psi(L) \log L}} P^\pi \left(\tau(\pi) > \lceil u \rceil \mid C \right) 2u du + \int_{\psi(L)}^\infty P^\pi \left(\tau(\pi) \geq \lceil \sqrt{x} \log L \rceil \mid C \right) dx \\ &\leq \frac{1}{(\log L)^2} \int_0^{\sqrt{\psi(L) \log L}} 2u du + \sum_{n \geq \lceil \sqrt{\psi(L) \log L} \rceil} \left(\frac{(n+1)^2}{\log L} - \frac{n^2}{\log L} \right) P^\pi \left(\tau(\pi) > n \mid C \right) \end{aligned}$$

$$\begin{aligned}
&\stackrel{(a)}{\leq} \frac{\psi(L)}{\log L} + \sum_{n \geq \lceil \sqrt{\psi(L)} \log L \rceil} \frac{2n+1}{\log L} P^\pi \left(M_h(n) < \log((K-1)L) \mid C \right) \\
&\stackrel{(b)}{\leq} \frac{\psi(L)}{\log L} + \frac{1}{\log L} \sum_{n \geq \lceil \sqrt{\psi(L)} \log L \rceil} (2n+1) \frac{B}{n^3} \\
&\leq \frac{\psi(L)}{\log L} + \frac{1}{\log L} \sum_{n \geq 1} (2n+1) \frac{B}{n^3},
\end{aligned} \tag{198}$$

where (a) above follows by noting that $\{\tau(\pi) < n\} \subset \{M_h(n) < \log((K-1)L)\}$, and (b) is due to Lemma 6. Noting that the summation (198) is finite, we get

$$\begin{aligned}
\limsup_{L \rightarrow \infty} E^\pi \left[\left(\frac{\tau(\pi)}{\log L} \right)^2 \mid C \right] &\leq \limsup_{L \rightarrow \infty} \frac{\psi(L)}{\log L} \\
&= \limsup_{L \rightarrow \infty} \frac{u(L)}{\log L} \\
&< \infty.
\end{aligned} \tag{199}$$

This establishes (152) and completes the proof of the proposition. ■

REFERENCES

- [1] P. N. Karthik and R. Sundaresan, “Detecting an odd restless markov arm with a trembling hand,” *arXiv preprint arXiv:2005.06255*, 2020.
- [2] M. Naghshvar and T. Javidi, “Two-dimensional visual search,” in *2013 IEEE International Symposium on Information Theory*. IEEE, 2013, pp. 1262–1266.
- [3] P. Whittle, “Restless bandits: Activity allocation in a changing world,” *Journal of applied probability*, vol. 25, no. A, pp. 287–298, 1988.
- [4] P. Mandl, “Estimation and control in markov chains,” *Advances in Applied Probability*, pp. 40–60, 1974.
- [5] V. Borkar and P. Varaiya, “Identification and adaptive control of markov chains,” *SIAM Journal on Control and Optimization*, vol. 20, no. 4, pp. 470–489, 1982.
- [6] 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.
- [7] 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.
- [8] G. R. Prabhu, S. Bhashyam, A. Gopalan, and R. Sundaresan, “Optimal odd arm identification with fixed confidence,” *arXiv preprint arXiv:1712.03682*, 2017.
- [9] 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.
- [10] A. Deshmukh, S. Bhashyam, and V. V. Veeravalli, “Controlled sensing for composite multihypothesis testing with application to anomaly detection,” in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2018, pp. 2109–2113.
- [11] —, “Sequential controlled sensing for composite multihypothesis testing,” *arXiv preprint arXiv:1910.12697*, 2019.
- [12] G. R. Prabhu, S. Bhashyam, A. Gopalan, and R. Sundaresan, “Sequential multi-hypothesis testing in multi-armed bandit problems: An approach for asymptotic optimality,” *arXiv preprint arXiv:2007.12961*, 2020.
- [13] 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.
- [14] V. Moulos, “Optimal best markovian arm identification with fixed confidence,” in *Advances in Neural Information Processing Systems*, 2019, pp. 5606–5615.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] K. J. Åström and B. Wittenmark, “On self tuning regulators,” *Automatica*, vol. 9, no. 2, pp. 185–199, 1973.

- [19] B. Doshi and S. E. Shreve, "Strong consistency of a modified maximum likelihood estimator for controlled markov chains," *Journal of Applied Probability*, pp. 726–734, 1980.
- [20] V. Borkar and P. Varaiya, "Adaptive control of markov chains, i: Finite parameter set," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 953–957, 1979.
- [21] P. R. Kumar, "A survey of some results in stochastic adaptive control," *SIAM Journal on Control and Optimization*, vol. 23, no. 3, pp. 329–380, 1985.
- [22] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- [23] L. M. Ausubel and R. J. Deneckere, "A generalized theorem of the maximum," *Economic Theory*, vol. 3, no. 1, pp. 99–107, 1993.
- [24] D. A. Levin and Y. Peres, *Markov chains and mixing times*. American Mathematical Soc., 2017, vol. 107.
- [25] A. Federgruen, A. Hordijk, and H. C. Tijms, "A note on simultaneous recurrence conditions on a set of denumerable stochastic matrices," *Journal of Applied Probability*, pp. 842–847, 1978.
- [26] P. Milgrom and I. Segal, "Envelope theorems for arbitrary choice sets," *Econometrica*, vol. 70, no. 2, pp. 583–601, 2002.
- [27] J. Munkres, *Topology*, ser. Featured Titles for Topology. Prentice Hall, Incorporated, 2000. [Online]. Available: <https://books.google.co.in/books?id=XjoZAQAIAAJ>
- [28] Y. S. Chow and H. Teicher, *Probability theory: independence, interchangeability, martingales*. Springer Science & Business Media, 2012.
- [29] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.