Best Arm Identification in Restless Markov Multi-Armed Bandits

P. N. Karthik, *Member, IEEE*, Kota Srinivas Reddy, *Member, IEEE*, and Vincent Y. F. Tan, *Senior Member, IEEE*

***Abstract*— We study the problem of identifying the best arm in a multi-armed bandit environment when each arm is a time-homogeneous and ergodic discrete-time Markov process on a common, finite state space. The state evolution on each arm is governed by the arm's transition probability matrix (TPM). A decision entity that knows the set of arm TPMs but not the exact mapping of the TPMs to the arms, wishes to find the index of the best arm as quickly as possible, subject to an upper bound on the error probability. The decision entity selects one arm at a time sequentially, and all the unselected arms continue to undergo state evolution (*restless* arms). For this problem, we derive the first-known problem instance-dependent asymptotic lower bound on the growth rate of the expected time required to find the index of the best arm, where the asymptotics is as the error probability vanishes. Further, we propose a sequential policy that, for an input parameter R, forcibly selects an arm that has not been selected for R consecutive time instants. We show that this policy achieves an upper bound that depends on R and is monotonically non-increasing as $R \to \infty$. The question of whether, in general, the limiting value of the upper bound as $R \to \infty$ matches with the lower bound, remains open. We identify a special case in which the upper and the lower bounds match. Prior works on best arm identification have dealt with (a) independent and identically distributed observations from the arms, and (b) rested Markov arms, whereas our work deals with the more difficult setting of restless Markov arms.**

***Index Terms*— Restless bandit, Markov chain, best arm identification, Markov decision problem, transition matrix.**

I. Introduction

CONSIDER a multi-armed bandit with $K \geq 2$ arms in which each arm is associated with a time-homogeneous and ergodic discrete-time Markov process evolving on a common, finite state space. The state evolutions on each arm are governed by the arm's transition probability matrix (TPM). Given a function f on the common state space of the arms, we define the *best arm* as the arm with the largest average value of f, averaged over the arm's stationary distribution. A decision entity that has knowledge of the set of TPMs of the arms, but does not know the exact mapping of the TPMs to the arms, wishes to find the index of the best arm as quickly as possible, subject to an upper bound on the error probability. Our interest is in the asymptotics as the error probability vanishes.

The above problem, known popularly in the literature as *best arm identification*, is an instance of an optimal stopping problem in decision theory, and can be embedded within the framework of active sequential hypothesis testing studied in the classical works of Chernoff [2] and Albert [3]. Prior works on best arm identification have dealt with (a) independent and identically distributed (*i.i.d.*) observations from the arms, as in [4], and (b) *rested* Markov arms [5] in which the arms yield Markovian observations and an arm evolves only when selected, and remains frozen otherwise. In this work, we extend the results of [4], [5], and [6] to the more difficult setting when the unselected arms continue to evolve (*restless arms*). An examination of the results in [4], [5], and [6] shows that given an error probability threshold $\epsilon > 0$, the minimum expected expected time required to find the best arm (or, expected stopping time) with an error probability no more than ϵ grows as $\Theta(\log(1/\epsilon))$ in the limit as $\epsilon \downarrow 0$. We anticipate a similar growth rate for the expected stopping time in the setting of restless arms. Our goal is to characterise or bound the exact constant multiplying $\log(1/\epsilon)$ for the setting of restless arms and attempt to achieve the lower bound of $\Omega(\log(1/\epsilon))$ in the limit as $\epsilon \downarrow 0$. Additionally, we aim to devise an efficient policy that achieves the fundamental limit.

A. A Preliminary Trembling Hand Model

The continued evolution of the unselected arms in our work makes it necessary for the decision entity to keep a record of (a) the time elapsed since each arm was previously selected (the arm's *delay*), and (b) the state of each arm recorded at its previous selection instant (the arm's *last observed state*). The notion of arm delays is superfluous when the arms are rested because the unobserved arms remain frozen. They are also redundant when the arms yield *i.i.d.* observations because the observation from an arm at any given time is independent of all its previous observations. When the arms are restless, the arm delays are non-negative, integer-valued,

Manuscript received 31 March 2022; accepted 10 December 2022. Date of publication 20 December 2022; date of current version 21 April 2023. This work was supported in part by the National Research Foundation (NRF) Singapore and DSO National Laboratories through the AI Singapore Program (AISG) under Award AISG2-RP-2020-018 and in part by NRF Fellowship under Grant A-0005077-01-00. An earlier version of this paper was presented in part at the 2022 IEEE Information Theory Workshop [DOI: 10.1109/ITW54588.2022.9965908]. *(Corresponding author: P. N. Karthik.)*

P. N. Karthik is with the Institute of Data Science, National University of Singapore, Singapore 119077 (e-mail: karthik@nus.edu.sg).

Kota Srinivas Reddy is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: ksvr1532@gmail.com).

Vincent Y. F. Tan is with the Department of Mathematics and the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: vtan@nus.edu.sg).

Communicated by A. Krishnamurthy, Associate Editor for Machine Learning and Statistics.

Digital Object Identifier 10.1109/TIT.2022.3230939

and introduce a countably-infinite dimension to the problem. In a related problem of identifying an anomalous or *odd* arm in a restless multi-armed bandit, Karthik and Sundaresan [7], [8] demonstrated that the arm delays and the last observed states constitute a controlled Markov process. A key aspect of the works [7] and [8] is the notion of a *trembling hand* for selecting the arms that is defined in the system model. Probabilistically, a trembling hand with parameter $\gamma \in [0,1]$ selects an arm uniformly at random with probability γ, and selects the intended arm with probability $1-\gamma$. As the authors note in [7], when $\gamma > 0$, the controlled Markov process (of arm delays and last observed states) satisfies a key ergodicity property (see [7, Lemma 1]) which is pivotal in establishing matching upper and lower bounds on the expected time to find the odd arm. An understanding of whether the lower bound for the case $\gamma = 0$ admits a matching upper bound remains open.

B. Forgoing the Trembling Hand Model and Constraining the Maximum Delay of Each Arm

The trembling hand model of [7] and [8] implies that at any given time, the probability of selecting any arm is at least γ/K, irrespective of the arm selection scheme used. In this work, we forgo the trembling hand assumption and allow for arm selection schemes to put zero mass on certain arms. We derive a lower bound on the expected time to find the best arm over *all* arm selection schemes. However, such a generic lower bound may not be achievable. In order to analyse the achievability of the lower bound, we restrict attention to those arm selection policies which, for an input parameter $R > K$, forcefully select an arm if its delay is equal to R. For this class of policies, we derive an upper bound in terms of R and show that the sequence of upper bounds, one for each value of R, is monotonically non-increasing as R increases. An advantage of our method over the methods of [7] and [8] is that while the arm selection schemes in [7] and [8] are not practically implementable, the arm selection schemes we propose in this paper are easy to implement.

The question of whether, in general, the limiting value of the upper bounds matches the generic lower bound appears to be a difficult problem and remains open. Notwithstanding this, we identify some special cases when the limiting value of the upper bounds matches with the generic lower bound.

C. Prior Works on Multi-Armed Bandits and Best Arm Identification

The problem of minimising regret for multi-armed bandits was introduced in the seminal work of Lai and Robbins [9] when each arm yields *i.i.d.* observations. Anantharam et al. [10] and Tekin and Liu [11] extended the results of Lai and Robbins to the setting of rested arms and restless arms respectively. We refer the reader to [12] for an extensive survey of regret minimization problems. In [13] and [4], the authors derived a lower bound on the sample complexity of best arm identification in multi-armed bandits with *i.i.d.* observations from arms. Many follow-up works proposed algorithms to achieve the lower bounds; notable among them are action-elimination algorithms [14], [15], upper confidence bound (UCB) algorithms [16], [17], and lower and upper confidence bound (LUCB) algorithms [18], [19]. Several extensions to the classical setup of best-arm identification in multi-armed bandits have appeared in the literature; notable among them are correlated bandits [20], cascading bandits [21], bandits with switching costs [22], and bandits with corrupted rewards [23].

For the problem of finding the best arm as quickly as possible subject to an upper bound on the error probability, the paper [4] provided a problem instance-dependent lower bound on the expected time required to find the best arm for the setting of *i.i.d.* observations from the arms, and a *track-and-stop* scheme that meets the lower bound asymptotically as the error probability vanishes. Moulos [5] extended the results of [4] to the setting of rested arms with hidden Markov observations. However, the lower and the upper bounds in [5] differ by a constant multiplicative factor; the achievability analysis therein uses novel concentration inequalities for Markov processes derived by the author. For a related problem of finding the anomalous (or odd) arm in multi-armed bandits, the works [7], [8], [24] obtain matching upper and lower bounds on the expected time required to find the odd arm subject to an upper bound on the error probability. While [24] studies the setting of rested arms, the works [7], [8] focus on the setting of restless arms. A key aspect of the works [7], [8] is a certain trembling hand model for selecting the arms, motivated from a certain visual science experiment. The paper [7] assumes that the TPMs of the odd arm and the non-odd arms are known beforehand to the decision entity, whereas the works [8], [24] deal with the case when the arm TPMs are unknown.

The recent works [25], [26], [27] study a more general problem of sequential hypothesis testing in multi-armed bandits, some special cases of which are the problems of best arm identification and odd arm identification, in the context of *i.i.d.* observations from each arm. An extension of the results of [25], [26], and [27] to the settings of rested and restless arms is a possible direction of future work.

D. Contributions

In this paper, we study the problem of finding the best arm in a restless multi-armed bandit as quickly as possible, subject to an upper bound on the error probability. Our technical contributions are as follows:

- In Section V, we show that given any $\epsilon > 0$, the expected time required to find the best arm with an error probability no more than ϵ grows as $\Omega(\log(1/\epsilon))$ in the limit as $\epsilon \downarrow 0$. We explicitly characterise the problem instance-dependent constant multiplying $\log(1/\epsilon)$ that is a function of the arm TPMs.
- For a problem instance C, the constant $T^\star(C)$ appearing in our lower bound is the solution of a sup-min optimisation problem, where the supremum is over all possible state-action occupancy measures associated with the countable-state controlled Markov process of arm delays and last observed states; see (20) for details. The question of whether the supremum in the expression for $T^\star(C)$ lower bound is attained is still open. The key difficulty

in showing this is the presence of the countably-infinite valued arm delays appearing in the expression for $T^\star(C)$, which makes further simplifications of the expression for $T^\star(C)$ difficult. In the prior works [4], [5], [6], further simplification of the expression for the constants governing the lower bound is possible because the notion of arm delays is superfluous in the settings of those works.

In a related problem of odd arm identification in restless multi-armed bandits, Karthik and Sundaresan [7], [8] show that under a trembling hand model for arms selection, the supremum in the expression for the lower bounds in these works may be further simplified by restricting the supremum to the class of all stationary arm selection policies; see [7], [8] for more details. However, it is unclear if such a simplification is possible in the absence of the trembling hand (as is the setting of this paper).

– In our achievability analysis presented in Section VI, to ameliorate the difficulty arising from the countably infinite-valued arm delays, we constrain the maximum delay of each arm to be no more than R, and focus our attention on those policies which select an arm forcibly if its delay equals R. This constraint on the maximum delay of each arm makes the state-action space finite and amenable to further analyses. To the best of our knowledge, this is the first work to analyse delay-constrained policies for restless multi-armed bandits.

  It is worth noting that the achievability analyses in the works [7], [8] rely crucially on a key ergodicity property (see [7, Lemma 1]) for the controlled Markov process of arm delays and last observed states that is satisfied under a *trembling hand* model for arms selection. This ergodicity property is pivotal to the achievability analysis in [7] and [8], and it is unclear if the same property holds in the absence of the trembling hand.
– We devise a policy that, for an input parameter R, forcibly pulls an arm if its delay equals R, and achieves an upper bound of the order $\Theta(\log(1/\epsilon))$. We show that the best (smallest) constant multiplying $\log(1/\epsilon)$ is equal to $1/T^\star_R(C)$ under the problem instance C; see (26) for the exact expression of $T^\star_R(C)$. Our results imply that $1/T^\star_R(C)$ is a valid asymptotic growth rate for the expected stopping time for all integers R.
– We show that $T^\star_R(C)$ is non-decreasing in R, and that $T^\star_R(C) \leq T^\star(C)$ for all R, thus implying that $\lim_{R\to\infty} T^\star_R(C)$ exists. Thus, the lower bound on the limiting value of the expected stopping time normalised by $\log(1/\epsilon)$ as $\epsilon \downarrow 0$, is given by $T^\star(C)^{-1}$, whereas the upper bound is governed by $\lim_{R\to\infty} T^\star_R(C)^{-1}$. While it is certainly true that $\lim_{R\to\infty} T^\star_R(C) \leq T^\star(C)$, showing that, in general, this inequality is an equality seems to be a difficult problem and remains open. In the special case when the TPM of each arm has identical rows, which is akin to obtaining *i.i.d.* observations from the arms, the above inequality is an equality, thus leading to matching upper and lower bounds in this special setting.

E. Paper Organisation

The rest of this paper is organised as follows. In Section II, we setup the notations and state our central goal. In Section III, we introduce the notions of arm delays, last observed states, and the Markov decision problem arising from the arm delays and the last observed states. We provide expressions for the log-likelihoods and the log-likelihood ratios, the basic quantities of analysis, in Section IV, and present the asymptotic lower bound on the growth rate of the expected time required to find the best arm in Section V. In Section VI, we constrain the maximum delay of each arm to be no more than R for some $R \in \mathbb{N} \cap (K, \infty)$, and describe a policy that forcibly samples an arm whose delay equals R. We provide results on the performance of the policy in Section VII, and state the main result in Section VIII. We discuss the convergence of $T^\star_R(C)$ to $T^\star(C)$ as $R \to \infty$ in Section IX, where we show that the convergence takes place in the special case when the TPM of each arm has identical rows. We conclude the paper in Section X. The proofs of all the results are contained in Appendices A-I.

II. Notations and Preliminaries

We consider a multi-armed bandit with $K \geq 2$ arms, and define $\mathcal{A} := \{1, \ldots, K\}$ to be the set of arms. We associate with each arm an ergodic discrete-time Markov process on a common, finite state space $\mathcal{S}$. We assume that the Markov process of each arm is independent of those of the other arms. We write $\{X^a_t : t \geq 0\}$ denote the Markov process of arm $a \in \mathcal{A}$.[1] The state evolution on each arm is governed by its transition probability matrix (TPM). Given TPMs $P_1, \ldots, P_K$ and a permutation $\sigma : \{1, \ldots, K\} \to \{1, \ldots, K\}$, let $C = (P_{\sigma(1)}, \ldots, P_{\sigma(K)})$ denote an assignment of the TPMs to the arms in which the TPM assigned to arm a is $P_{\sigma(a)}$. In the sequel, we refer to C as an *assignment of the TPMs*, and we let $\mathcal{C}$ denote the collection of all possible assignments of the TPMs, i.e.,

$$\mathcal{C} = \{(P_{\sigma(1)}, \ldots, P_{\sigma(K)}) : \ \sigma \text{ is a permutation on } \mathcal{A}\}. \quad (1)$$

For each $k = 1, \ldots, K$, let $\mu_k = \{\mu_k(i) : \ i \in \mathcal{S}\}$ denote the unique stationary distribution of the TPM P_k. Given a function $f : \mathcal{S} \to \mathbb{R}$, let

$$\nu_a := \sum_{i\in\mathcal{S}} f(i)\, \mu_{\sigma(a)}(i), \quad a \in \mathcal{A}, \quad (2)$$

denote the average value of f under μ_a. Define the *best arm* $a^\star \in \mathcal{A}$ as

$$a^\star := \arg\max_{a\in\mathcal{A}} \ \nu_a = \arg\max_{a\in\mathcal{A}} \sum_{i\in\mathcal{S}} f(i)\, \mu_{\sigma(a)}(i). \quad (3)$$

We assume throughout the paper that $a^\star$ is unique. Without loss of generality, let $P_{\sigma(a^\star)} = P_1$.[2]

For an integer $d \geq 1$ and a matrix P, let P^d denote the matrix obtained by multiplying P with itself d times. For $i, j \in \mathcal{S}$ and $d \geq 1$, let $P^d(j|i)$ to denote the (i, j)th element of P^d. For $a \in \mathcal{A}$, let each row of P_a be mutually absolutely

[1] Throughout the paper, time $t \in \{0, 1, 2, \ldots\}$.

[2] The reader may recognise that this does not necessarily imply $a^\star = 1$.

continuous with the corresponding row of $P_{a'}$ for all $a' \neq a$. It is easy to see that this implies that for all $d \geq 1$, each row of P_a^d is mutually absolutely continuous with the corresponding row of $P_{a'}^d$. The above assumption implies that the decision entity cannot infer the best arm merely by observing certain specific state(s) or state-transition(s) on the arm.

A decision entity that knows $P_1, \ldots, P_K$ only up to a permutation wishes to find the index of the best arm (i.e., the arm whose TPM is P_1) as quickly as possible, subject to an upper bound on the error probability. Clearly, this is accomplished if the decision entity finds $C \in \mathcal{C}$ that defines the problem instance. Given $C \in \mathcal{C}$, we write $\text{Alt}(C)$ to denote the set of all assignments of the TPMs alternative to C, i.e., those assignments of the TPMs in which the location of the best arm is different from the one in C. In order to find the index of the best arm in a problem instance C, the decision entity selects the arms sequentially, one at each time t. Let A_t be the arm selected at time t. The decision entity observes the state of the arm A_t, denoted by $\bar{X}_t$. In contrast to the previous works [4], [5], [6] that deal with *i.i.d.* observations from the arms and rested arms, we assume that the unobserved arms continue to undergo state evolution whether or not they are selected (*restless* arms). Let $(A_{0:t}, \bar{X}_{0:t}) := (A_0, \bar{X}_0, \ldots, A_t, \bar{X}_t)$ denote the history of all the arm selections and observations seen up to time t. All random variables are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Define the filtration

$$\mathcal{F}_0 := \{\Omega, \emptyset\}, \quad \mathcal{F}_t := \sigma(A_{0:t-1}, \bar{X}_{0:t-1}), \quad t \geq 1. \tag{4}$$

A. Policy and Problem Definition

A policy π is defined by a collection of functions $\{\pi_t : t \geq 0\}$. At each time t, π_t does one of the following based on the history $\mathcal{F}_t$:

- stop and declare the index of the best arm;
- choose to pull arm A_t according to a deterministic or a randomised rule.

Let π denote a generic policy, and let $\tau(\pi)$ denote the stopping time of policy π (defined with respect to the filtration (4)). Let $\theta(\pi)$ denote the index of the best arm declared by the policy π at the stopping time.

Let $P_C^\pi(\cdot)$ denote the probability computed under the assignment of the TPMs C and under the policy π. For $a \in \mathcal{A}$, let $\mathcal{C}_a \subset \mathcal{C}$ denote the collection of all permutations in which $P_{\sigma(a)} = P_1$. Clearly, the collection $\{\mathcal{C}_a : a \in \mathcal{A}\}$ is a partition of $\mathcal{C}$. Given an error probability threshold $\epsilon > 0$, let

$$\Pi(\epsilon) := \{\pi : \text{ for all } a \in \mathcal{A}, \quad P_C^\pi(\theta(\pi) \neq a) \leq \epsilon \ \forall C \in \mathcal{C}_a\} \tag{5}$$

denote the collection of all policies whose error probability at the stopping time is no more than ϵ for all possible assignments of TPMs. We anticipate from similar results in the prior works [4], [5], [6] that

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

Here, $\mathbb{E}_C^\pi[\cdot]$ denotes the expectation under the assignment of the TPMs C and policy π. Our interest is in characterising, or at least bounding, the value of

$$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{\mathbb{E}_C^\pi[\tau(\pi)]}{\log(1/\epsilon)}. \tag{6}$$

For simplicity, we assume that every policy starts by selecting arm 1 at time $t = 0$, arm 2 at time $t = 1$, etc., and arm K at time $t = K - 1$. This ensures that the Markov process of each arm is observed at least once.

III. Delays, Last Observed States, and a Markov Decision Problem

With an intent to keep the material in the paper self-contained, we reproduce the contents of [7, Section II-B, II-C] here, but with suitable modifications to reflect the absence of the *trembling hand* assumption of [7]. Recall that the decision entity observes only one of the arms at each time t, while the unobserved arms continue to undergo state evolution. This means that at any time t, the probability of the observation $\bar{X}_t$ on the selected arm A_t given the history $\mathcal{F}_t$ is a function of (a) the time elapsed since the previous time instant of selection of arm A_t (called the *delay* of arm A_t), and (b) the state of arm A_t at its previous selection time instant (called the *last observed state* of arm A_t). Notice that when the arms are *rested*, the notion of arm delays is superfluous since each arm remains frozen at its previously observed state until its next selection time instant. Also, the notion of arm delays is redundant in the setting of iid observations because the current state of the arm selected is independent of the state at its previous selection. Thus, arm delays are a key distinguishing feature of the setting of restless arms.

For $t \geq K$, let $d_a(t)$ and $i_a(t)$ respectively denote the delay and the last observed state of arm a at time t. Let $\underline{d}(t) := (d_1(t), \ldots, d_K(t))$ and $\underline{i}(t) := (i_1(t), \ldots, i_K(t))$ denote the vectors of arm delays and the last observed states at time t. Note that arm delays and last observed states are defined only for $t \geq K$ as these quantities are well-defined only when at least one observation is available from each arm. Set $\underline{d}(K) = (K, K-1, \ldots, 1)$. Thus, $d_a(t) \geq 1$ for all $t \geq K$, and that $d_a(t) = 1$ if and only if arm a is selected at time $t-1$.

The rule for updating the arm delays and last observed states is as follows: if $A_t = a'$, then

$$d_a(t+1) = \begin{cases} d_a(t) + 1, & a \neq a', \\ 1, & a = a', \end{cases} \tag{7}$$

$$i_a(t+1) = \begin{cases} i_a(t), & a \neq a', \\ X_t^a, & a = a', \end{cases} \tag{8}$$

where $X_t^a = \bar{X}_t$ is the state of the arm $A_t = a'$ at time t. Thus, for all $t \geq K$, based on $\{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}$, the decision entity chooses to pull A_t, observes $\bar{X}_t$,[3] and then constructs $(\underline{d}(t+1), \underline{i}(t+1))$. This repeats until the stopping time, at which time the decision entity declares $\theta(\pi)$ (under policy π) as the candidate best arm.

[3]Note that specifying $\{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}$ is equivalent to specifying $(A_{0:t-1}, \bar{X}_{0:t-1})$ for all $t \geq K$.

From the update rule in (8), it is clear that the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ takes values in a subset $\mathbb{S}$ of the countable set $\mathbb{N}^K \times \mathcal{S}^K$, where $\mathbb{N} = \{1, 2, \ldots\}$ denotes the set of natural numbers. The subset $\mathbb{S}$ is formed based on the constraint that at any time $t \geq K$, exactly one of the components of $\underline{d}(t)$ is equal to 1, and all the other components are > 1. Given any assignment of the TPMs $C \in \mathcal{C}$ and policy π, note that for all $(\underline{d}', \underline{i}') \in \mathbb{S}$ and $t \geq K$,

$$P_C^\pi(\underline{d}(t+1) = \underline{d}', \underline{i}(t+1) = \underline{i}' | (\underline{d}(s), \underline{i}(s)), K \leq s \leq t, A_{0:t}) = P_C^\pi(\underline{d}(t+1) = \underline{d}', \underline{i}(t+1) = \underline{i}' | (\underline{d}(t), \underline{i}(t)), A_t). \quad (9)$$

On account of (9) being satisfied, we say that under any policy π, the evolution of the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is *controlled* by the sequence $\{A_t\}_{t \geq 0}$ of intended arm selections under policy π. Alternatively, $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is a controlled Markov process, with $\{A_t\}_{t \geq 0}$ being the sequence of controls.[4] Thus, we are in a Markov decision problem (MDP) setting. We now make precise the state space, the action space, the transition probabilities and our objective.

The state space of the MDP is $\mathbb{S}$, with the state at time t denoted $(\underline{d}(t), \underline{i}(t))$. The action space of the MDP is $\mathcal{A}$, with action A_t at time t possibly depending on the history $\mathcal{F}_t$. The transition probabilities for the MDP under an assignment of the TPMs $C \in \mathcal{C}$ and a policy π are given by

$$P_C^\pi(\underline{d}(t+1) = \underline{d}', \underline{i}(t+1) = \underline{i}' \mid \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, A_t = a) = \begin{cases} (P_C^a)^{d_a}(i'_a | i_a), & \text{if } d'_a = 1 \text{ and } d'_{\tilde{a}} = d_{\tilde{a}} + 1 \ \forall \tilde{a} \neq a, \\ & i'_{\tilde{a}} = i_{\tilde{a}} \ \forall \tilde{a} \neq a, \\ 0, & \text{otherwise}, \end{cases} \quad (10)$$

where P_C^a denotes the TPM of arm a under the assignment of the TPMs C. For instance, if $C = (P_1, \ldots, P_K)$, then $P_C^a = P_a$ for all $a \in \mathcal{A}$. Note that the right hand side of (10) is not a function of t or π. Let $Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, a)$ denote the transition probabilities in (10).

IV. Log-Likelihoods and Log-Likelihood Ratios

Given $C \in \mathcal{C}$, let

$$Z_C^\pi(n) := \log P_C^\pi(A_{0:n}, \bar{X}_{0:n}) \quad (11)$$

denote the log-likelihood of all the controls and observations seen up to time n under the policy π when C is the assignment of the TPMs. Under the assumption that π selects arm 1 at time $t = 0$, arm 2 at time $t = 1$, etc., and arm K at time $K - 1$, (11) may be expressed as

$$Z_C^\pi(n) = \sum_{a=1}^{K} \log P_C^\pi(X_{a-1}^a) \quad (12a)$$

$$+ \sum_{t=K}^{n} \log P_C^\pi(A_t | A_{0:t-1}, \bar{X}_{0:t-1}) \quad (12b)$$

$$+ \sum_{t=K}^{n} \log P_C^\pi(\bar{X}_t | A_{0:t}, \bar{X}_{0:t-1}). \quad (12c)$$

[4]The terminology used here follows that of Borkar [28].

Because π is oblivious to the assignment of the TPMs C, (12b) does not depend on C. Furthermore, (12c) may be expressed as

$$\sum_{t=K}^{n} \log P_C^\pi(\bar{X}_t | A_{0:t}, \bar{X}_{0:t-1}) = \sum_{t=K}^{n} \sum_{a=1}^{K} \mathbb{I}_{\{A_t = a\}} \log P_C^\pi(\bar{X}_t | A_{0:t-1}, A_t = a, \bar{X}_{0:t-1}) = \sum_{t=K}^{n} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \mathbb{I}_{\{\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, A_t = a\}} \log (P_C^a)^{d_a}(\bar{X}_t | i_a) = \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, a, j) \log (P_C^a)^{d_a}(j | i_a), \quad (13)$$

where

$$N(n, \underline{d}, \underline{i}, a, j) := \sum_{t=K}^{n} \mathbb{I}_{\{\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, A_t = a, \bar{X}_t = j\}} \quad (14)$$

denotes the number of times up to time n the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is in the state $(\underline{d}, \underline{i})$, arm a is selected subsequently, and the state of arm a is observed to be j. Let

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

$$N(n, \underline{d}, \underline{i}) := \sum_{a=1}^{K} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, a, j). \quad (15)$$

Given $C, C' \in \mathcal{C}$, let

$$Z_{CC'}^\pi(n) := Z_C^\pi(n) - Z_{C'}^\pi(n) = \log \frac{P_C^\pi(A_{0:n}, \bar{X}_{0:n})}{P_{C'}^\pi(A_{0:n}, \bar{X}_{0:n})} \quad (16)$$

denote the log-likelihood ratio (LLR) of the controls and observations seen under the policy π when the assignment of the TPMs is C with respect to that when the assignment of the TPMs is C'. Using (13) in conjunction with the fact that (12b) does not depend on C, we get

$$Z_{CC'}^\pi(n) = \sum_{a=1}^{K} \log \frac{P_C^\pi(X_{a-1}^a)}{P_{C'}^\pi(X_{a-1}^a)} + \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, a, j) \log \frac{(P_C^a)^{d_a}(j | i_a)}{(P_{C'}^a)^{d_a}(j | i_a)}. \quad (17)$$

V. Lower Bound

We now present a lower bound for (6). Given two probability distributions μ and ν on the finite state space $\mathcal{S}$, the *Kullback–Leibler (KL) divergence* (also called the *relative entropy*) between μ and ν is defined as

$$D_{\mathrm{KL}}(\mu \| \nu) := \sum_{i \in \mathcal{S}} \mu(i) \log \frac{\mu(i)}{\nu(i)}, \quad (18)$$

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

Proposition 1: Suppose $C \in \mathcal{C}$ is the underlying assignment of the TPMs. Then,

$$\liminf_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{\mathbb{E}_C^\pi[\tau(\pi)]}{\log(1/\epsilon)} \geq \frac{1}{T^\star(C)}, \quad (19)$$

where $T^\star(C)$ is given by

$$T^\star(C) := \sup_{\nu} \min_{C' \in \text{Alt}(C)} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a)\ k_{CC'}(\underline{d},\underline{i},a), \tag{20}$$

where $k_{CC'}(\underline{d},\underline{i},a) = D_{\text{KL}}((P_C^a)^{d_a}(\cdot|i_a)\|(P_{C'}^a)^{d_a}(\cdot|i_a))$. In (20), the supremum is over all $\nu = \{\nu(\underline{d},\underline{i},a) : (\underline{d},\underline{i}) \in \mathbb{S}, a \in \mathcal{A}\}$ satisfying

$$\sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a) = 1, \tag{21}$$

$$\nu(\underline{d},\underline{i},a) \geq 0 \quad \text{for all } (\underline{d},\underline{i},a) \in \mathbb{S} \times \mathcal{A}. \tag{22}$$

Proof: The proof proceeds in several steps. First, we derive an analogue of the ubiquitous change-of-measure result [6, Lemma 18] for the setting of restless arms. We then lower bound the expected LLR of any policy whose stoppage error probability is at most ϵ by $d(\epsilon, 1-\epsilon)$, where $d(x,y)$ denotes the relative entropy between two Bernoulli distributions with parameters x and y. Next, we derive an upper bound on the expected LLR in terms of the expected stopping time. This involves deriving an analogue of Wald's identity for the setting of restless arms. Combining the upper and the lower bounds for the expected LLR, we get (19). The details are in Appendix A. ∎

A. A Flow Constraint

Notice that $T^\star(C)$ is the optimal value of an infinite-dimensional linear program (LP). The question of whether there exists ν attaining the supremum in (20) remains open. The key difficulty in showing this is that because the set $\mathbb{S}$ is countably infinite, it is not clear whether the set of all ν satisfying (21)-(22) (which is akin to the space of all probability distributions on $\mathbb{S} \times \mathcal{A}$) is compact. Also, because this supremum is over *all* probability distributions on $\mathbb{S} \times \mathcal{A}$ which is a large class of distributions, the lower bound in (19) may not be achievable. It seems necessary to introduce additional constraints on ν to render the lower bound achievable. Indeed, given $\delta > 0$, suppose ν_δ is a probability distribution on $\mathbb{S} \times \mathcal{A}$ such that

$$\min_{C' \in \text{Alt}(C)} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu_\delta(\underline{d},\underline{i},a)\ k_{CC'}(\underline{d},\underline{i},a) \geq T^\star(C) - \delta.$$

One way to achieve the quantity on the left hand side of the above equation is to ensure that for all $(\underline{d},\underline{i},a) \in \mathbb{S} \times \mathcal{A}$, the value of the fraction $N(n,\underline{d},\underline{i},a)/n$ is close to $\nu_\delta(\underline{d},\underline{i},a)$ for all n large (the regime of large n is akin to the regime of vanishing error probabilities). It seems difficult to accomplish this in the absence of more structure on ν_δ. It is worth noting here that in a related problem of odd arm identification, the authors of [7] are confronted with a similar difficulty in showing the achievability of the lower bound therein. To ameliorate the difficulty, they introduce a version of the following *flow* constraint on ν:

$$\text{flow constraint}: \ \forall (\underline{d}',\underline{i}') \in \mathbb{S},$$
$$\sum_{a=1}^{K} \nu(\underline{d}',\underline{i}',a) = \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a)\ Q_C(\underline{d}',\underline{i}'|\underline{d},\underline{i},a). \tag{23}$$

In (23), Q_C is the MDP transition matrix defined in (10). The flow constraint is, in fact, a *global balance* equation, and dictates that for any $(\underline{d}',\underline{i}') \in \mathbb{S}$, the long-term probability of a transition from $(\underline{d}',\underline{i}')$ (the *flow* out of $(\underline{d}',\underline{i}')$, captured by the left hand side of (23)) is equal to the probability of a transition to $(\underline{d}',\underline{i}')$ (the *flow* into $(\underline{d}',\underline{i}')$, captured by the right hand side of (23)). The authors of [7] show that their lower bound, after including the flow constraint, can be achieved by a certain *trembling hand*-based policy; see [7, Section V] for a description of the policy. However, the policy in [7] is not practically implementable.

With an end goal of showing achievability of our lower bound, we take (23) into consideration along with (21)-(22) when evaluating the supremum in (20), and wish to design a policy that (a) is computationally feasible/tractable and easy-to-implement, and (b) achieves the lower bound in (19). This forms the content of the next section.

VI. Achievability

As alluded to in the previous section, the *trembling hand*-based policy of [7] is not practically implementable. One reason for this is that the arm delays (which appear in the policy of [7]) take countably infinitely many values and therefore cannot be handled on a machine with finite-size memory. To alleviate the difficulty arising from the countably infinite-valued arm delays, we study a simplified setting where the maximum delay of each arm is restricted to be at most R for some $R \in \mathbb{N} \cap (K, \infty)$,[5] and an arm whose delay is equal to R at any given time is forcibly selected in the following time instant. Let $\mathbb{S}_R$ denote the subset of $\mathbb{S}$ in which the delay of each arm is no more than R. Further, for $a \in \mathcal{A}$, let $\mathbb{S}_{R,a}$ denote the subset of $\mathbb{S}_R$ in which the delay of arm a is equal to R. Notice that $\mathbb{S}_{R,a}$ is a finite set for each a and that $\mathbb{S}_{R,a}$ and $\mathbb{S}_{R,a'}$ are disjoint for all $a' \neq a$.

A. Modifications to the MDP Transition Probabilities Under Maximum Delay Constraint

Recall that $\{(\underline{d}(t),\underline{i}(t)) : t \geq K\}$ and $\{A_t : t \geq 0\}$ together define a Markov decision problem (MDP) whose state space is $\mathbb{S}$, the action space is $\mathcal{A}$, the state at time t is $(\underline{d}(t),\underline{i}(t))$, and the control (or action) at time t is A_t. From Section III, we know that the transition probabilities of the MDP are given by (10). When the delay of each arm is constrained to be no more than R, the modified state space of the MDP is $\mathbb{S}_R$, and the modified transition probabilities for the MDP are as follows: leftmargin=*

- *Case 1:* $(\underline{d},\underline{i}) \notin \bigcup_{a=1}^{K} \mathbb{S}_{R,a}$. In this case, the transition probabilities are as in (10).

[5]We consider $R > K$ to be consistent with our assumption that each of the arms is sampled once in the first K time instants.

- *Case 2:* $(\underline{d},\underline{i}) \in \mathbb{S}_{R,a}$ for some $a \in \mathcal{A}$. In this case, when $A_t = a$,

$$P_C^{\pi}(\underline{d}(t+1)=\underline{d}', \underline{i}(t+1)=\underline{i}' | \underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a) = \begin{cases} (P_C^a)^R(i_a'|i_a), & \text{if } d_a'=1 \text{ and } d_{\tilde{a}}' = d_{\tilde{a}}+1 \ \forall \tilde{a} \neq a, \\ & i_{\tilde{a}}' = i_{\tilde{a}} \ \forall \tilde{a} \neq a, \\ 0, & \text{otherwise}, \end{cases} \tag{24}$$

and when $A_t \neq a$, the transition probabilities are undefined.

We write $Q_{C,R}(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, a)$ to denote the transition probabilities in (24).

B. Capturing the Maximum Delay Constraint and a Finite Dimensional Linear Program

Recall that in the absence of any constraints on the maximum delay of each arm, the lower bound is as in (19), with the constant $T^\star(C)$ in (19) as given in (20). Further, the supremum in (20) is over all ν satisfying (21)-(23). When the delay of each arm is constrained to be no more than R, the following additional constraint on ν comes into play:

$$R\text{-max-delay-constraint}: \quad \nu(\underline{d},\underline{i},a) = \sum_{a'=1}^{K} \nu(\underline{d},\underline{i},a') \quad \forall (\underline{d},\underline{i}) \in \mathbb{S}_{R,a}. \tag{25}$$

The condition in (25) captures the observation that any occurrence of the state $(\underline{d},\underline{i}) \in \mathbb{S}_{R,a}$ is followed by selecting arm a forcibly (i.e., with probability 1), thus implying that $\nu(\underline{d},\underline{i},a') = 0$ for all $a' \neq a$ which is equivalent to (25). Let $T_R^\star(C)$ be the optimal value of the following optimisation problem:

$$\sup_{\nu} \min_{C' \in \text{Alt}(C)} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}_R} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a) \ k_{CC'}(\underline{d},\underline{i},a), \tag{26}$$

subject to

$$\sum_{a=1}^{K} \nu(\underline{d}',\underline{i}',a) = \sum_{(\underline{d},\underline{i}) \in \mathbb{S}_R} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a) \ Q_{C,R}(\underline{d}',\underline{i}'|\underline{d},\underline{i},a) \quad \forall (\underline{d}',\underline{i}') \in \mathbb{S}_R, \tag{27}$$

$$\sum_{(\underline{d},\underline{i}) \in \mathbb{S}_R} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a) = 1, \tag{28}$$

$$\nu(\underline{d},\underline{i},a) \geq 0 \quad \forall (\underline{d},\underline{i},a) \in \mathbb{S}_R \times \mathcal{A}, \tag{29}$$

$$\nu(\underline{d},\underline{i},a) = \sum_{a'=1}^{K} \nu(\underline{d},\underline{i},a') \quad \forall (\underline{d},\underline{i}) \in \mathbb{S}_{R,a}, \ a \in \mathcal{A}. \tag{30}$$

Notice that the above optimisation problem is a finite-dimensional LP, and is the analogue of the infinite-dimensional LP in (20) with the additional condition in (30) to account for the case when an arm is forcibly selected if its delay equals R. Because (a) $\mathbb{S}_R \times \mathcal{A}$ is finite, (b) the space of all probability distributions on $\mathbb{S}_R \times \mathcal{A}$ (say, $\mathscr{P}(\mathbb{S}_R \times \mathcal{A})$) is compact with respect to the topology arising from the Euclidean metric in $\mathbb{R}^{|\mathbb{S}_R \times \mathcal{A}|}$, (c) the set of all ν satisfying (27)-(30) is a closed subset of $\mathscr{P}(\mathbb{S}_R \times \mathcal{A})$ (and therefore compact), and (d) the expression

$$\min_{C' \in \text{Alt}(C)} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}_R} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a) \ k_{CC'}(\underline{d},\underline{i},a)$$

is continuous in ν, it follows by Weierstrass extreme value theorem that there exists $\nu_{C,R}^\star = \{\nu_{C,R}^\star(\underline{d},\underline{i},a) : (\underline{d},\underline{i},a) \in \mathbb{S}_R \times \mathcal{A}\}$ that attains the supremum in (26). Although a closed-form expression for $\nu_{C,R}^\star$ is not currently available, it can easily be evaluated numerically. In the next section, we fix $R \in \mathbb{N} \cap (K,\infty)$ and design a policy for finding the best arm that samples an arm forcibly if its delay is equal to R. Additionally, we demonstrate that our policy (a) stops in finite time almost surely, (b) satisfies the desired error probability, and (c) achieves an upper bound of $1/T_R^\star(C)$ asymptotically as the error probability vanishes. We shall see that our policy is easy to implement as it operates on the finite set $\mathbb{S}_R$ instead of the countable set $\mathbb{S}$.

C. A Policy for Finding the Best Arm Under a Maximum Delay Constraint

Fix an $R \in \mathbb{N} \cap (K,\infty)$. In this section, we design a policy for finding the best arm when the delay of each arm is constrained to be no more than R. Towards this, we first analyse a uniform arm selection policy that, for all $t \geq K$, selects the arms uniformly whenever the delay of each arm is $< R$, and forcibly selects an arm whose delay is equal to R. Let this policy be denoted π_R^{unif}. It is clear that $\{(\underline{d}(t),\underline{i}(t)) : t \geq K\}$ is a Markov process under π_R^{unif} with $\mathbb{S}_R$ as its state space. The following result shows that this Markov process is, in fact, ergodic.

Lemma 1: Fix $R \in \mathbb{N} \cap (K,\infty)$. Under every assignment of the TPMs $C \in \mathcal{C}$, the process $\{(\underline{d}(t),\underline{i}(t)) : \ t \geq K\}$ is ergodic under the policy π_R^{unif}.

Proof: See Appendix B. ∎

As a consequence of Lemma 1, the process $\{(\underline{d}(t),\underline{i}(t)) : t \geq K\}$ has a unique stationary distribution, say $\mu_{C,R}^{\text{unif}} = \{\mu_{C,R}^{\text{unif}}(\underline{d},\underline{i}) : (\underline{d},\underline{i}) \in \mathbb{S}_R\}$, under the policy π_R^{unif} and under the assignment of the TPMs C. We note that $\mu_{C,R}^{\text{unif}}(\underline{d},\underline{i}) > 0$ for all $(\underline{d},\underline{i}) \in \mathbb{S}_R$. Let

$$\nu_{C,R}^{\text{unif}}(\underline{d},\underline{i},a) := \begin{cases} \frac{\mu_{C,R}^{\text{unif}}(\underline{d},\underline{i})}{K}, & (\underline{d},\underline{i}) \notin \bigcup_{a'=1}^{K} \mathbb{S}_{R,a'}, \\ \mu_{C,R}^{\text{unif}}(\underline{d},\underline{i}), & (\underline{d},\underline{i}) \in \mathbb{S}_{R,a}, \\ 0, & (\underline{d},\underline{i}) \in \bigcup_{a' \neq a} \mathbb{S}_{R,a'}. \end{cases} \tag{31}$$

denote the corresponding *ergodic* state-action occupancy measure. Observe that $\nu_{C,R}^{\text{unif}}$ satisfies (27)-(30).

For $\eta \in (0,1]$ and $C \in \mathcal{C}$, let

$$\nu_{\eta,R,C}(\underline{d},\underline{i},a) := \eta \ \nu_{C,R}^{\text{unif}}(\underline{d},\underline{i},a) + (1-\eta) \ \nu_{C,R}^\star(\underline{d},\underline{i},a), \tag{32}$$

$$\mu_{\eta,R,C}(\underline{d},\underline{i}) := \sum_{a=1}^{K} \nu_{\eta,R,C}(\underline{d},\underline{i},a). \tag{33}$$

Observe that $\mu_{\eta,R,C}(\underline{d},\underline{i}) \geq \frac{\eta}{K} \ \mu_{C,R}^{\text{unif}}(\underline{d},\underline{i}) > 0$ for all $(\underline{d},\underline{i}) \in \mathbb{S}_R$. Also, $\nu_{\eta,R,C}$ satisfies (27)-(30) by virtue of the

fact that both $\nu^{\text{unif}}_{C,R}$ and $\nu^{\star}_{C,R}$ satisfy (27)-(30). Let $\lambda_{\eta,R,C} = \{\lambda_{\eta,R,C}(a|\underline{d},\underline{i}) : a \in \mathcal{A},\ (\underline{d},\underline{i}) \in \mathbb{S}_R\}$ be defined as

$$\lambda_{\eta,R,C}(a|\underline{d},\underline{i}) = \frac{\nu_{\eta,R,C}(\underline{d},\underline{i},a)}{\mu_{\eta,R,C}(\underline{d},\underline{i})} \quad (\underline{d},\underline{i}) \in \mathbb{S}_R,\ a \in \mathcal{A}. \quad (34)$$

Notice that for all $(\underline{d},\underline{i}) \notin \bigcup_{a'=1}^{K} \mathbb{S}_{R,a'}$,

$$\lambda_{\eta,R,C}(a|\underline{d},\underline{i}) \geq \frac{\eta}{K}\ \mu^{\text{unif}}_{C,R}(\underline{d},\underline{i}) \geq \frac{\eta}{K}\ \mu^{\min}_{R}, \quad (35)$$

where

$$\mu^{\min}_{R} := \min_{C\in\mathcal{C}} \min_{(\underline{d},\underline{i})\in\mathbb{S}_R} \mu^{\text{unif}}_{C,R}(\underline{d},\underline{i}) > 0. \quad (36)$$

That is, whenever the delay of each arm is $< R$, the distribution $\lambda_{\eta,R,C}(\cdot|\underline{d},\underline{i})$ puts a strictly positive mass on each of the arms. Using the arm selection rule in (34) instead of the uniform sampling rule and following the proof template in Appendix B, it can be shown that the policy $\pi^{\lambda_{\eta,R,C}}$, which selects the arms at each time instant according to the rule in (34), renders the process $\{(\underline{d}(t),\underline{i}(t)) : t \geq K\}$ ergodic. We now claim that the stationary distribution of the process $\{(\underline{d}(t),\underline{i}(t)) : t \geq K\}$ under $\pi^{\lambda_{\eta,R,C}}$ is $\mu_{\eta,R,C}$. Indeed, suppose $Q^{C,R} = \{Q^{C,R}(\underline{d}',\underline{i}'|\underline{d},\underline{i}) : (\underline{d}',\underline{i}'),(\underline{d},\underline{i}) \in \mathbb{S}_R\}$ denotes the transition probability matrix of the Markov process $\{(\underline{d}(t),\underline{i}(t)) : t \geq K\}$ under the policy $\pi^{\lambda_{\eta,R,C}}$. Then,

$$Q^{C,R}(\underline{d}',\underline{i}'|\underline{d},\underline{i}) = \sum_{a=1}^{K} \lambda_{\eta,R,C}(a|\underline{d},\underline{i})\ Q_{C,R}(\underline{d}',\underline{i}'|\underline{d},\underline{i},a), \quad (37)$$

from which it follows that for all $(\underline{d}',\underline{i}') \in \mathbb{S}_R$,

$$\begin{aligned}
&\sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \mu_{\eta,R,C}(\underline{d},\underline{i})\ Q^{C,R}(\underline{d}',\underline{i}'|\underline{d},\underline{i}) \\
&= \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \mu_{\eta,R,C}(\underline{d},\underline{i})\ \lambda_{\eta,R,C}(a|\underline{d},\underline{i})\ Q_{C,R}(\underline{d}',\underline{i}'|\underline{d},\underline{i},a) \\
&= \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \nu_{\eta,R,C}(\underline{d},\underline{i},a) Q_{C,R}(\underline{d}',\underline{i}'|\underline{d},\underline{i},a) \\
&\overset{(a)}{=} \sum_{a=1}^{K} \nu_{\eta,R,C}(\underline{d}',\underline{i}',a) \\
&= \mu_{\eta,R,C}(\underline{d}',\underline{i}'), \quad (38)
\end{aligned}$$

thus proving the claim. In the above set of equations, (a) follows from the fact that $\nu_{\eta,R,C}$ satisfies (27).

Remark 1: Observe that $\nu_{\eta,R,C}$ in (32) is a *mixture* of two terms, one arising from the uniform arm selection rule under maximum delay constraints (the term corresponding to $\nu^{\text{unif}}_{C,R}$), and the other arising from the optimal solution to the finite-dimensional LP under a constraint on the delay of each arm (the term corresponding to $\nu^{\star}_{C,R}$). A similar, *trembling hand*-based mixture term appears in the works [7], [8]. While the mixtures in [7] and [8] arise from a restrictive system model that forces each arm to be selected with a strictly positive probability, the mixture in (32) ensures that each arm is selected with a strictly positive probability (except when the delay of an arm is equal to the maximum allowed delay R in which case it is forcibly selected) without any restrictions on the system model. It is also worth noting that the mixtures in [7] and [8] give rise to a conditional probability distribution on the arms, conditioned on the arm delays and the last observed states (a quantity akin to $\lambda(a|\underline{d},\underline{i})$), whereas the mixture in (32) gives rise to a joint probability distribution on the set $\mathbb{S}_R \times \mathcal{A}$ (a quantity akin to $\nu(\underline{d},\underline{i},a)$).

For $n \geq 0$, let

$$M^{\pi^\star(L,\eta,R)}_{C}(n) = \min_{C'\in\text{Alt}(C)} Z^{\pi^\star(L,\eta,R)}_{CC'}(n), \quad C \in \mathcal{C}.$$

Our policy, which we call *R-Delay-Constrained-Restless-BAI* (or R-DCR-BAI in short) or alternatively as $\pi^\star(L,\eta,R)$, is described in Algorithm 1. Here, $L > 1$, $\eta \in (0,1]$ and $R \in \mathbb{N} \cap (K,\infty)$ are parameters of the policy. As we shall see, the parameter L controls the policy's error probability. In fact, we will see from Lemma 3 that if we set $L = 1/\epsilon$, the error probability is bounded above by ϵ.

Algorithm 1 R-DCR-BAI or $\pi^\star(L,\eta,R)$

Input: $L > 1$, $\eta \in (0,1]$, and $R \in \mathbb{N} \cap (K,\infty)$
Output: Best arm $\theta(\pi^\star(L,\eta,R))$
Initialisation: `stop` $=$ `false`, $n = K-1$,
$A_0 = 1, A_1 = 2, \ldots, A_{K-1} = K$
while `stop` $==$ `false` **do**
  Compute $\bar{C}(n) \in \arg\max_{C\in\mathcal{C}}\ M^{\pi^\star(L,\eta,R)}_{C}(n)$.
  Resolve ties uniformly at random.
  if $M^{\pi^\star(L,\eta,R)}_{\bar{C}(n)}(n) \geq \log(L(K-1)(K-1)!)$ **then**
    `stop` $\leftarrow$ `true`
    $\theta(\pi^\star(L,\eta,R)) = a^\star(\bar{C}(n))$.
  else
    $n \leftarrow n+1$
    Pull $A_n \sim \lambda_{\eta,R,\bar{C}(n-1)}(\cdot \mid \underline{d}(n),\underline{i}(n))$.
  end if
end while
return $\theta(\pi^\star(L,\eta,R))$

In Algorithm 1, $\bar{C}(n)$ denotes the estimate of the underlying assignment of the TPMs based on all the controls (arm selections) and observations seen up to time n. If the LLR between $\bar{C}(n)$ and its nearest alternative assignment of the TPMs exceeds a certain threshold (i.e., $\geq \log(L(K-1)(K-1)!)$), then the policy is sufficiently confident that $\bar{C}(n)$ is indeed the underlying assignment of the TPMs, and therefore stops and declares the index of the best arm in $\bar{C}(n)$. Else, it samples the next arm based on the value of $(\underline{d}(n),\underline{i}(n))$ according to the distribution $\lambda_{\eta,R,\bar{C}(n-1)}(\cdot|\underline{d}(n),\underline{i}(n))$.

Remark 2: The function f in (3) used to define the best arm appears implicitly in the description of the policy R-DCR-BAI, at the place where $\text{Alt}(C)$ is evaluated. By defining $\text{Alt}(C)$ appropriately, the above policy can be used to solve a host of other closely related problems such as finding the second-best arm, finding the exact association of the TPMs to the arms (i.e., the permutation σ such that the underlying assignment of the TPMs is $C = (P_{\sigma(1),\ldots,\sigma(K)})$), finding the top-$M$ arms for $M \leq K$, etc. See Section X for a discussion on this.

In the following section, we demonstrate that for a suitable choice of L, the policy R-DCR-BAI achieves the desired error

probability. Further, letting $L \to \infty$, we show that the growth rate of its expected stopping time satisfies an asymptotic upper bound that is arbitrarily close to $1/T_R^\star(C)$ under the assignment of the TPMs C for a suitable choice of η.

VII. Results on the Performance of Policy R-DCR-BAI

This section is organised as follows. In Section VII-A, we establish that for a given $C \in \mathcal{C}$ and any $C' \in \text{Alt}(C)$, the LLR $Z^\pi_{CC'}(n)$ has a strictly positive drift a.s. under the assignment of the TPMs C (Lemma 2), and therefore the policy R-DCR-BAI stops in finite time a.s.. In Section VII-B, we show that R-DCR-BAI satisfies any desired error probability for a suitable choice of L (Lemma 3). In Section VII-C, we strengthen the result of Section VII-A by showing that the LLR of C with respect to its nearest alternative $C' \in \text{Alt}(C)$ is a certain constant that, in the limit as $\eta \downarrow 0$, converges to $T_R^\star(C)$ (Proposition 2). In Section VII-D, we show that the stopping time of R-DCR-BAI grows a.s. as $L \to \infty$ (Lemma 4). In Section VII-E, we derive an almost sure upper bound on the stopping time of R-DCR-BAI that, in the limit as $\eta \downarrow 0$, converges to $1/T_R^\star(C)$. In Section VII-F, we show that the expected stopping time of R-DCR-BAI satisfies the same upper bound as that derived in Section VII-E. This is based on (a) showing that the family $\{\tau(\pi^\star(L,\eta,R))/\log L : L > 1\}$ is uniformly integrable, and (b) combining the almost sure upper bound with the uniform integrability result to obtain an upper bound in expectation. The proofs of all the results are relegated to the appendices.

A. Strictly Positive Drift for the LLRs

Let $\pi^\star_{\text{NS}}(L,\eta,R)$ denote a version of R-DCR-BAI that never stops, i.e., it does not check the second step in R-DCR-BAI and continues to the last step indefinitely. We now show that under $\pi^\star_{\text{NS}}(L,\eta,R)$, the LLRs have a strictly positive drift as the number of rounds of arm selection $n \to \infty$.

Lemma 2: Fix $L > 1$, $\eta \in (0,1]$, $R \in \mathbb{N} \cap (K,\infty)$, and $C \in \mathcal{C}$. Under the assignment of the TPMs C and the policy $\pi = \pi^\star_{\text{NS}}(L,\eta,R)$,

$$\liminf_{n\to\infty} \frac{Z^\pi_{CC'}(n)}{n} > 0 \quad \text{a.s.} \quad \text{for all } C' \in \text{Alt}(C). \tag{39}$$

Proof: See Appendix C. ∎

Lemma 2 asserts that under the TPMs assignment C,

$$\liminf_{n\to\infty} \frac{M^\pi_C(n)}{n} > 0 \quad \text{a.s.} \tag{40}$$

when $\pi = \pi^\star_{\text{NS}}(L,\eta,R)$. This means that $M^\pi_C(n) \geq \log(L(K-1)(K-1)!)$ for all n large, a.s.. This proves that R-DCR-BAI stops in finite time a.s..

B. Desired Error Probability

In this section, we show that for an appropriate choice of the parameter L, the policy R-DCR-BAI achieves any desired error probability.

Lemma 3: Fix an error probability threshold $\epsilon > 0$. If $L = 1/\epsilon$, then $\pi^\star(L,\eta,R) \in \Pi(\epsilon)$ for all $\eta \in (0,1]$ and $R \in \mathbb{N} \cap (K,\infty)$. Here, $\Pi(\epsilon)$ is as defined in (5).

Proof: The proof uses the fact that the policy stops in finite time a.s., and is given in Appendix D. ∎

C. The Correct Asymptotic Drift of the LLRs

In this section, we strengthen the result of Section VII-A by showing that under the constraint that the delay of each arm is at most R, the asymptotic drift of the LLRs is arbitrarily close to $T_R^\star(C)$.

Proposition 2: Fix $L > 1$, $\eta \in (0,1]$, $R \in \mathbb{N} \cap (K,\infty)$, and $C \in \mathcal{C}$. Consider the policy $\pi = \pi^\star_{\text{NS}}(L,\eta,R)$. Under the assignment of the TPMs C, for all $C' \in \text{Alt}(C)$,

$$\lim_{n\to\infty} \frac{Z^\pi_{CC'}(n)}{n} = \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^K \nu_{\eta,R,C}(\underline{d},\underline{i},a)\; k_{CC'}(\underline{d},\underline{i},a) \quad \text{a.s..} \tag{41}$$

Consequently, it follows that a.s.,

$$\lim_{n\to\infty} \frac{M^\pi_C(n)}{n} = \min_{C'\in\text{Alt}(C)} \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^K \nu_{\eta,R,C}(\underline{d},\underline{i},a)\; k_{CC'}(\underline{d},\underline{i},a). \tag{42}$$

Proof: See Appendix E. ∎

From (32), we note that the right hand side of (42) may be lower bounded by

$$\begin{aligned}
&\eta \min_{C'\in\text{Alt}(C)} \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^K \nu^{\text{unif}}_{C,R}(\underline{d},\underline{i},a)\; k_{CC'}(\underline{d},\underline{i},a) \\
&+ (1-\eta) \min_{C'\in\text{Alt}(C)} \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^K \nu^\star_{C,R}(\underline{d},\underline{i},a)\; k_{CC'}(\underline{d},\underline{i},a) \\
&= \eta \min_{C'\in\text{Alt}(C)} \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^K \nu^{\text{unif}}_{C,R}(\underline{d},\underline{i},a)\; k_{CC'}(\underline{d},\underline{i},a) \\
&\qquad + (1-\eta)\; T_R^\star(C),
\end{aligned} \tag{43}$$

which, as $\eta \downarrow 0$, converges to $T_R^\star(C)$. Using this observation, we shall show later that our policy achieves an upper bound of $1/T_R^\star(C)$ in the limit as $\eta \downarrow 0$.

D. Asymptotic Growth of Stopping Time

In this section, we demonstrate that $\tau(\pi^\star(L,\eta,R))$ grows as $L \to \infty$ (equivalently $\epsilon \downarrow 0$).

Lemma 4: Fix $\eta \in (0,1]$, $R \in \mathbb{N} \cap (K,\infty)$, and $C \in \mathcal{C}$. Under the assignment of the TPMs C,

$$\liminf_{L\to\infty} \tau(\pi^\star(L,\eta,R)) = \infty \text{ a.s..} \tag{44}$$

Proof: See Appendix F. ∎

As a consequence of Lemma 4, we get that under the assignment of the TPMs C and under $\pi = \pi^\star(L,\eta,R)$,

$$\lim_{L\to\infty} \frac{M^\pi_C(\tau(\pi))}{\tau(\pi)} = \min_{C'\in\text{Alt}(C)} \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^K \nu_{\eta,R,C}(\underline{d},\underline{i},a)\; k_{CC'}(\underline{d},\underline{i},a) \quad \text{a.s..} \tag{45}$$

E. Almost Sure Asymptotic Upper Bound on the Stopping Time

In this section, we derive an almost sure asymptotic upper bound on the stopping time of R-DCR-BAI as $L \to \infty$, and show that this upper bound is arbitrarily close to $1/T_R^\star(C)$ under the assignment of the TPMs C. In the next section, we combine the almost sure upper bound of this section with a certain uniform integrability result to claim that the expected stopping time of R-DCR-BAI satisfies the same upper bound as that derived in this section.

Lemma 5: Fix $\eta \in (0,1]$, $R \in \mathbb{N} \cap (K, \infty)$, and $C \in \mathcal{C}$. Under the assignment of the TPMs C and under the policy $\pi = \pi^*(L, \eta, R)$,

$$\begin{aligned}&\limsup_{L\to\infty} \frac{\tau(\pi)}{\log L}\\ &\leq \frac{1}{\min\limits_{C'\in \text{Alt}(C)} \sum\limits_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum\limits_{a=1}^{K} \nu_{\eta,R,C}(\underline{d},\underline{i},a)\ k_{CC'}(\underline{d},\underline{i},a)}\\ &\leq \frac{1}{\eta\, T_R^{\text{unif}}(C) + (1-\eta)\ T_R^\star(C)} \quad \text{a.s.,} \end{aligned} \tag{46}$$

where

$$T_R^{\text{unif}}(C) := \min_{C'\in \text{Alt}(C)} \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \nu_{C,R}^{\text{unif}}(\underline{d},\underline{i},a)\ k_{CC'}(\underline{d},\underline{i},a).$$

Proof: See Appendix G ∎

F. Asymptotic Upper Bound on the Expected Stopping Time

In this section, we show that the expected value of the stopping time of policy R-DCR-BAI satisfies an asymptotic upper bound that matches with the right hand side of (46) as $L \to \infty$.

Proposition 3: Fix $\eta \in (0,1]$, $R \in \mathbb{N} \cap (K, \infty)$, and $C \in \mathcal{C}$. Under the assignment of the TPMs C, the policy R-DCR-BAI satisfies

$$\limsup_{L\to\infty} \frac{\mathbb{E}_C^\pi[\tau(\pi)]}{\log L} \leq \frac{1}{\eta\, T_R^{\text{unif}}(C) + (1-\eta)\ T_R^\star(C)}. \tag{47}$$

Proof: In the proof, which we provide in Appendix H, we first show that the family $\{\tau(\pi)/\log L : L > 1\}$ is uniformly integrable. Combining (46) with the uniform integrability result yields (47). ∎

VIII. A Key Monotonicity Property and the Main Result

In this section, we establish a key monotonicity property for $T_R^\star(C)$, which is that $T_R^\star(C) \leq T_{R'}^\star(C)$ for all $R < R'$. This, combined together with the fact that $T_R^\star(C) \leq T^\star(C)$, implies that $\lim_{R\to\infty}\ T_R^\star(C)$ exists. We conclude the section by stating the main result of the paper.

A. A Key Monotonicity Property

The below result asserts that $T_R^\star(C)$ is monotonically non-decreasing as R increases.

Lemma 6: $T_R^\star(C) \leq T_{R+1}^\star(C)$ for all $R \in \mathbb{N} \cap (K, \infty)$.

Proof: Fix $R \in \mathbb{N} \cap (K, \infty)$. The key idea behind the proof is to note that (a) $\mathbb{S}_R \subset \mathbb{S}_{R+1}$, and (b) any ν that satisfies (27)-(30) with parameter R also satisfies them with parameter $R+1$. The details follow. Let $\nu_{C,R}^\star$ and $\nu_{C,R+1}^\star$ be the optimal state-action measures when the arm delays are constrained to no more than R and $R+1$ respectively. Note that both $\nu_{C,R}^\star$ and $\nu_{C,R+1}^\star$ satisfy (27)-(30) (with the corresponding parameters R and $R+1$). Further, $\nu_{C,R}^\star$ satisfies (27)-(30) with parameter $R+1$. Define $\widetilde{\nu}_{C,R+1}$ as

$$\widetilde{\nu}_{C,R+1}(\underline{d},\underline{i},a) := \begin{cases} \nu_{C,R}^\star(\underline{d},\underline{i},a) & \text{if } (\underline{d},\underline{i}) \in \mathbb{S}_R,\\ 0, & \text{otherwise.}\end{cases}$$

Clearly, $\widetilde{\nu}_{C,R+1}$ satisfies (27)-(30) with parameter $R+1$. We therefore have

$$\begin{aligned} &T_{R+1}^\star(C)\\ &= \min_{C'\in \text{Alt}(C)} \sum_{(\underline{d},\underline{i})\in\mathbb{S}_{R+1}} \sum_{a=1}^{K} \nu_{C,R+1}^\star(\underline{d},\underline{i},a)\ k_{CC'}(\underline{d},\underline{i},a)\\ &\overset{(a)}{\geq} \min_{C'\in \text{Alt}(C)} \sum_{(\underline{d},\underline{i})\in\mathbb{S}_{R+1}} \sum_{a=1}^{K} \widetilde{\nu}_{C,R+1}(\underline{d},\underline{i},a)\ k_{CC'}(\underline{d},\underline{i},a)\\ &= \min_{C'\in \text{Alt}(C)} \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \widetilde{\nu}_{C,R+1}(\underline{d},\underline{i},a)\ k_{CC'}(\underline{d},\underline{i},a)\\ &= \min_{C'\in \text{Alt}(C)} \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \nu_{C,R}^\star(\underline{d},\underline{i},a)\ k_{CC'}(\underline{d},\underline{i},a)\\ &= T_R^\star(C), \end{aligned}$$

thus establishing the desired result. ∎

Lemma 6, in conjunction with the observation that $T_R^\star(C) \leq T^\star(C)$ for all R, implies that $\lim_{R\to\infty} T_R^\star(C)$ exists and $\lim_{R\to\infty} T_R^\star(C) \leq T^\star(C)$. The question of whether this inequality is an equality seems difficult to prove, and is discussed in the next section.

B. Main Result

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

Theorem 4: Consider a multi-armed bandit with $K \geq 2$ arms in which each arm is a time homogeneous and ergodic discrete-time Markov process on the finite state space $\mathcal{S}$. Given TPMs $P_1, \ldots, P_K$ and a permutation $\sigma : \{1, \ldots, K\} \to \{1, \ldots, K\}$, let $C = (P_{\sigma(1)}, \ldots, P_{\sigma(K)})$ be the underlying assignment of the TPMs where $P_{\sigma(a)}$ denotes the TPM of arm a. The growth rate of the expected time required to find the best arm in C satisfies the lower bound

$$\liminf_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{\mathbb{E}_C^\pi[\tau(\pi)]}{\log(1/\epsilon)} \geq \frac{1}{T^\star(C)}. \tag{48}$$

Further, given any $\epsilon > 0$, the policy $\pi^\star(1/\epsilon, \eta, R) \in \Pi(\epsilon)$ for all $\eta \in (0,1]$ and $R \in \mathbb{N} \cap (K, \infty)$. Additionally,

$$\begin{aligned}&\limsup_{R\to\infty}\ \limsup_{\eta\downarrow 0}\ \limsup_{L\to\infty}\ \frac{\mathbb{E}_C^{\pi^\star(L,\eta,R)}[\tau(\pi^\star(L,\eta,R))]}{\log L}\\ &\leq \frac{1}{\lim_{R\to\infty}\ T_R^\star(C)}, \end{aligned} \tag{49}$$

thereby yielding

$$\begin{aligned}
&\frac{1}{T^\star(C)}\\
&\le \liminf_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{\mathbb{E}_C^\pi[\tau(\pi)]}{\log(1/\epsilon)}\\
&\le \limsup_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{\mathbb{E}_C^\pi[\tau(\pi)]}{\log(1/\epsilon)}\\
&\le \limsup_{R\to\infty} \limsup_{\eta\downarrow 0} \limsup_{L\to\infty} \frac{\mathbb{E}_C^{\pi^\star(L,\eta,R)}[\tau(\pi^\star(L,\eta,R))]}{\log L}\\
&\le \frac{1}{\lim_{R\to\infty} T_R^\star(C)}. \qquad (50)
\end{aligned}$$

Thus, the lower bound on the growth rate of the expected stopping time is $1/T^\star(C)$, and the upper bound is $1/(\lim_{R\to\infty} T_R^\star(C))$.

Proof: The asymptotic lower bound in (48) follows from Proposition 1. From Lemma 3, we know that for any $\epsilon > 0$, the policy $\pi^\star(1/\epsilon, \eta, R) \in \Pi(\epsilon)$ for all $\eta \in (0, 1]$ and $R \in \mathbb{N} \cap (K, \infty)$. Therefore, it follows that

$$\inf_{\pi\in\Pi(\epsilon)} \frac{\mathbb{E}_C^\pi[\tau(\pi)]}{\log(1/\epsilon)} \le \frac{\mathbb{E}^{\pi^\star(1/\epsilon,\eta,R)}[\tau(\pi^\star(1/\epsilon,\eta,R))]}{\log(1/\epsilon)}. \qquad (51)$$

Fixing η, R, and letting $\epsilon \downarrow 0$ (or equivalently, substituting $L = 1/\epsilon$ and letting $L \to \infty$) in (51), and using the upper bound in (47), we get

$$\begin{aligned}
&\limsup_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{\mathbb{E}_C^\pi[\tau(\pi)]}{\log(1/\epsilon)}\\
&\le \limsup_{L\to\infty} \frac{\mathbb{E}_C^{\pi^\star(L,\eta,R)}[\tau(\pi^\star(L,\eta,R))]}{\log L}\\
&\le \frac{1}{\eta\ T_R^{\text{unif}}(C) + (1-\eta)\ T_R^\star(C)}. \qquad (52)
\end{aligned}$$

Letting $\eta \downarrow 0$ in (52) and noting that the leftmost term in (52) does not depend on η, we get

$$\begin{aligned}
&\limsup_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{\mathbb{E}_C^\pi[\tau(\pi)]}{\log(1/\epsilon)}\\
&\le \limsup_{\eta\downarrow 0} \limsup_{L\to\infty} \frac{\mathbb{E}_C^{\pi^\star(L,\eta,R)}[\tau(\pi^\star(L,\eta,R))]}{\log L} \le \frac{1}{T_R^\star(C)}. \qquad (53)
\end{aligned}$$

Finally, letting $R \to \infty$ in (53), we arrive at (50). ∎

IX. On the Convergence of $T_R^\star(C)$ to $T^\star(C)$ as $R \to \infty$

Recall that $T^\star(C)$ is the optimal value of the infinite-dimensional LP in (20), where the supremum in (20) is over all ν satisfying (21)-(23), and $T_R^\star(C)$ is the optimal value of the finite-dimensional LP in (26) that arises when the the delay of each arm is constrained to be no more than R. From our exposition in Section VIII-A, we know that $\lim_{R\to\infty} T_R^\star(C) \le T^\star(C)$. Showing that, in general, this inequality is an equality appears to be difficult. In this section, we show that in the special case when the arm TPMs $P_1, \ldots, P_K$ have identical rows, which is akin to obtaining *i.i.d.* observations from the arms, we have $\lim_{R\to\infty} T_R^\star(C) = T^\star(C)$, thus leading to matching upper and lower bounds in this special setting.

Assume that the TPMs $P_1, \ldots, P_K$ have identical rows, and suppose that $\mu_1, \ldots, \mu_K$ are the unique stationary distributions associated with $P_1, \ldots, P_K$ respectively. Then, by the convergence result [29, Theorem 4.9] for finite-state Markov processes, each row of P_k must be equal to μ_k, $k = 1, \ldots, K$. In this special setting, the below result states that $T_R^\star(C) = T^\star(C)$ for all $R \in \mathbb{N} \cap (K, \infty)$, and therefore $\lim_{R\to\infty} T_R^\star(C) = T^\star(C)$.

Lemma 7: Suppose each row of P_k is equal to μ_k, $k = 1, \ldots, K$. In this special setting, $T^\star(C) = T_R^\star(C)$ for all $R \in \mathbb{N} \cap (K, \infty)$. Consequently, $\lim_{R\to\infty} T_R^\star(C) = T^\star(C)$.

Proof: The proof uses the key idea that for a given $C \in \mathcal{C}$ and for all $d \in \mathbb{N}$, $i \in \mathcal{S}$, and $C' \in \text{Alt}(C)$,

$$D_{\text{KL}}((P_C^a)^d(\cdot|i) \| (P_{C'}^a)^d(\cdot|i)) = D_{\text{KL}}(\mu_C^a \| \mu_{C'}^a),$$

where μ_C^a denotes the stationary distribution associated with the TPM P_C^a. The complete proof of Lemma 7 is given in Appendix I. ∎

For general arm TPMs with non-identical rows, we leave open the question of whether $\lim_{R\to\infty} T_R^\star(C)$ equals $T^\star(C)$ for future study.

X. Concluding Remarks and Discussion

1) We studied the problem of finding the best arm in a restless Markov multi-armed bandit as quickly as possible, subject to an upper bound on the error probability. For this optimal stopping problem, we showed that under the problem instance C, the expected time required to find the best arm with an error probability no more than ϵ is lower bounded by $\log(1/\epsilon)/T^\star(C)$ in the limit as $\epsilon \downarrow 0$ (converse). Here, $T^\star(C)$ is a problem-instance dependent constant that captures the hardness of the problem. We also devised a policy that, for an input parameter $R \in \mathbb{N} \cap (K, \infty)$, forcibly selects an arm which has not been selected for R consecutive time instants, and finds the best arm in at most $\log(1/\epsilon)/T_R^\star(C)$ time instants on the average as $\epsilon \downarrow 0$ (achievability).
2) We showed that $T_R^\star(C)$ is monotonically non-decreasing in R, and that $\lim_{R\to\infty} T_R^\star(C) \le T^\star(C)$. Showing that, in general, this inequality is an equality appears to be a difficult problem and remains open. Notwithstanding this, we showed that in the special case when the TPM of each arm has identical rows (which is akin to obtaining *i.i.d.* observations from each arm), the above inequality is indeed an equality.
3) The trembling hand-based policy of [7] is not practically implementable because it operates on the countable set $\mathbb{S}$ of *all* delays and last observed states which cannot be handled on a machine with finite-size memory. However, for any given $R \in \mathbb{N} \cap (K, \infty)$, our policy operates on the finite set $\mathbb{S}_R$ which can easily be stored in finite-size memory on a machine, thereby making it practically implementable.
4) In our achievability analysis, we assumed that the initial state of each arm follows a certain distribution ϕ that is

independent of the underlying assignment of the TPMs. However, this may not actually be the case. For instance, if the Markov process of each arm has evolved for a sufficiently long time and reached stationarity before the decision entity begins sampling the arms at $t = 0$, then the initial state of each arm follows the arm's stationary distribution. This will lead to a mismatch between the LLR expressions in our work (resulting from ϕ) and the actual LLR expressions (resulting from the stationary distributions). Suppose $\bar{Z}^{\pi}_{CC'}(n)$ denotes the analogue of (17) resulting from using the stationary distributions in place of ϕ. Then, fixing $C \in \mathcal{C}$, it can be shown that

$$\lim_{n\to\infty} \frac{Z^{\pi}_{CC'}(n)}{n} - \frac{\bar{Z}^{\pi}_{CC'}(n)}{n} = 0 \quad \text{a.s..}$$

That is, the asymptotic drift of $Z^{\pi}_{CC'}(n)$ is identical to that of $\bar{Z}^{\pi}_{CC'}(n)$, and therefore the assumption $X^a_0 \sim \phi$ does not affect the asymptotic analysis in any way.

5) It will be interesting to extend the results of our paper to the case when the arm TPMs $P_1, \ldots, P_K$ are not known to the decision entity beforehand. The difficulty here is that for any given $C \in \mathcal{C}$, the set of alternatives $\text{Alt}(C)$ is uncountably infinite. Also, the arm TPMs they must be estimated on-the-fly using the observations from the arms. In this case, showing that the TPM estimates converge to their true values is the key challenge. Indeed, this direction is interesting, as it looks amenable to the certainty equivalence approach, which is essentially applied in related works.
6) The function f in (3) used to define the best arm appears implicitly in the analyses of the lower and the upper bounds wherever one evaluates $\text{Alt}(C)$ for any given C. By redefining $\text{Alt}(C)$ appropriately, our analyses of the lower and the upper bounds may be extended to more general sequential hypothesis testing problems such as (a) finding the permutation σ such that the underlying assignment of the TPMs is $C = (P_{\sigma(1)}, \ldots, P_{\sigma(K)})$, (b) finding the second-best arm, (c) finding the top-M arms for $M \leq K$, etc. For instance, to analyse the problem in (a) above, we may redefine $\text{Alt}(C)$ as the set of all $C' \neq C$. Similarly, to analyse the problem in (b), assuming that the second-best arm is unique, we may redefine $\text{Alt}(C)$ as the set of all C' such that the index of the second-best arm in C' differs from that in C, and so on.
7) It will be interesting to extend our results to the more realistic setting in which the decision entity only observes $Y^a_t = f(X^a_t)$ and not the underlying state X^a_t of arm a at time t, i.e., the arms yield *hidden Markov* observations. Here, f is the same function that appears in (3). We anticipate that f will appear more explicitly in the expressions for $T^{\star}(C)$ and $T^{\star}_R(C)$, and also in the analyses of the lower and the upper bounds (instead of only appearing implicitly in the evaluation of $\text{Alt}(C)$ as in this paper). Because $\{Y^a_t : t \geq 0\}$ is not a Markov process in general, the analyses of the lower and the upper bounds in this modified setting appear to be quite challenging and worth exploring.

APPENDIX A
PROOF OF PROPOSITION 1

It suffices to prove (19) for all π such that $\mathbb{E}^{\pi}_C[\tau(\pi)] < \infty$, as (19) trivially holds when $\mathbb{E}^{\pi}_C[\tau(\pi)] = \infty$. This proof is organised as follows. In Section A-A, we derive a change-of-measure result that is the analogue of [6, Lemma 18] for the setting of restless arms (see (55)). Using the change-of-measure result together with [6, Lemma 19], we derive in Section A-B a lower bound for the expected LLR in terms of the error probability. In Section A-C, we derive an upper bound for the expected LLR in terms of the expected stopping time (see (72)). Combining the lower bound of Section A-B and the upper bound of Section A-C, and letting the error probability vanish, we arrive at the lower bound (19).

A. A Change-of-Measure Result for Restless Arms

The following change-of-measure result is the analogue of [6, Lemma 18] for the setting of restless arms. The proof technique is along the lines of the proof of [6, Lemma 18].

Lemma 8: Fix $C, C' \in \mathcal{C}$. Given a policy π with stopping time $\tau(\pi)$ such that $P^{\pi}_C(\tau(\pi) < \infty) = 1$, $P^{\pi}_{C'}(\tau(\pi) < \infty) = 1$, let

$$\mathcal{F}_{\tau(\pi)} := \{E \in \mathcal{F} : E \cap \{\tau(\pi) = t\} \in \mathcal{F}_t \text{ for all } t \geq 0\}, \tag{54}$$

where $\{\mathcal{F}_t : t \geq 0\}$ is as defined in (4). Then,

$$P^{\pi}_{C'}(E) = \mathbb{E}^{\pi}_C\big[\mathbb{I}_E \exp\big(-Z^{\pi}_{CC'}(\tau(\pi))\big)\big], \quad E \in \mathcal{F}_{\tau(\pi)}. \tag{55}$$

Proof of Lemma 8: We prove (55) by first demonstrating, through mathematical induction, that the relation

$$\mathbb{E}_{C'}[g(A_{0:t}, \bar{X}_{0:t})] = \mathbb{E}^{\pi}_C\big[g(A_{0:t}, \bar{X}_{0:t}) \exp\big(-Z^{\pi}_{CC'}(t)\big)\big] \tag{56}$$

holds for all $t \geq 0$ and for all measurable functions $g : \mathcal{A}^{t+1} \times \mathcal{S}^{t+1} \to \mathbb{R}$. Then, (55) follows from (56) by noting that for any $E \in \mathcal{F}_{\tau(\pi)}$,

$$\begin{aligned}
P^{\pi}_{C'}(E) &= \mathbb{E}^{\pi}_{C'}[\mathbb{I}_E] \\
&= \mathbb{E}^{\pi}_{C'}\left[\sum_{t\geq 0} \mathbb{I}_{E\cap\{\tau(\pi)=t\}}\right] \\
&\overset{(a)}{=} \sum_{t\geq 0} \mathbb{E}^{\pi}_{C'}\left[\mathbb{I}_{E\cap\{\tau(\pi)=t\}}\right] \\
&\overset{(b)}{=} \sum_{t\geq 0} \mathbb{E}^{\pi}_{C}\left[\mathbb{I}_{E\cap\{\tau(\pi)=t\}} \exp\big(-Z^{\pi}_{CC'}(t)\big)\right] \\
&= \sum_{t\geq 0} \mathbb{E}^{\pi}_{C}\left[\mathbb{I}_{E\cap\{\tau(\pi)=t\}} \exp\big(-Z^{\pi}_{CC'}(\tau(\pi))\big)\right] \\
&= \mathbb{E}^{\pi}_{C}\left[\mathbb{I}_E \exp\big(-Z^{\pi}_{CC'}(\tau(\pi))\big)\right],
\end{aligned} \tag{57}$$

where (a) is due to the monotone convergence theorem, and (b) above follows from (56) and the fact that $E \in \mathcal{F}_{\tau(\pi)}$ implies that $E \cap \{\tau(\pi) = t\} \in \mathcal{F}_t$ for all $t \geq 0$.

The proof of (56) for the case $t = 0$ may be obtained as follows: for any measurable $g : \mathcal{A} \times \mathcal{S} \to \mathbb{R}$,

$$\mathbb{E}^{\pi}_{C'}[g(A_0, \bar{X}_0)]$$

$$
\begin{aligned}
&= \sum_{a=1}^{K} \sum_{i\in\mathcal{S}} g(a,i)\ P^{\pi}_{C'}(A_0 = a, \bar{X}_0 = i) \\
&= \sum_{a=1}^{K} \sum_{i\in\mathcal{S}} g(a,i)\ P^{\pi}_{C'}(A_0 = a)\ P^{\pi}_{C'}(\bar{X}_0 = i | A_0 = a) \\
&\overset{(a)}{=} \sum_{a=1}^{K} \sum_{i\in\mathcal{S}} g(a,i)\ P^{\pi}_{C}(A_0 = a)\ \phi(i) \\
&= \sum_{a=1}^{K} \sum_{i\in\mathcal{S}} g(a,i)\ P^{\pi}_{C}(A_0 = a)\ P^{\pi}_{C}(X^a_0 = i | A_0 = a) \\
&= \mathbb{E}^{\pi}_{C}[g(A_0, \bar{X}_0)] \\
&\overset{(b)}{=} \mathbb{E}^{\pi}_{C}\big[g(A_0, \bar{X}_0)\ \exp\big(-Z^{\pi}_{CC'}(0)\big)\big], \qquad (58)
\end{aligned}
$$

where in writing (a), we make use of (i) the fact that $P^{\pi}_{C'}(A_0 = a) = P^{\pi}_{C}(A_0 = a)$ because the policy π selects arms without the knowledge of the underlying assignment of the TPMs, and (ii) the assumption that $X^a_0 \sim \phi$ for all $a \in \mathcal{A}$, where ϕ is a probability distribution on $\mathcal{S}$ that does not depend on the underlying assignment of the TPMs. In writing (b) above, we make use of the observation that

$$
Z^{\pi}_{CC'}(0) = \log \frac{P^{\pi}_{C}(A_0, \bar{X}_0)}{P^{\pi}_{C'}(A_0, \bar{X}_0)} = 0. \qquad (59)
$$

We now assume that (56) is true for some $t > 0$, and show that it also true for $t+1$. By the law of iterated expectations, $\mathbb{E}^{\pi}_{C'}[g(A_{0:t+1}, \bar{X}_{0:t+1})] = \mathbb{E}^{\pi}_{C'}[\mathbb{E}^{\pi}_{C'}[g(A_{0:t+1}, \bar{X}_{0:t+1})|\mathcal{F}_{t+1}]]$. Because $\mathbb{E}^{\pi}_{C'}[g(A_{0:t+1}, \bar{X}_{0:t+1})|\mathcal{F}_{t+1}]$ is a measurable function of $(A_{0:t}, \bar{X}_{0:t})$, by the induction hypothesis, we have

$$
\begin{aligned}
&\mathbb{E}^{\pi}_{C'}[g(A_{0:t+1}, \bar{X}_{0:t+1})|\mathcal{F}_{t+1}] \\
&= \mathbb{E}^{\pi}_{C}\big[\mathbb{E}^{\pi}_{C'}[g(A_{0:t+1}, \bar{X}_{0:t+1})\,\big|\,\mathcal{F}_{t+1}]\ \exp\big(-Z^{\pi}_{CC'}(t)\big)\big] \\
&= \mathbb{E}^{\pi}_{C}\big[\mathbb{E}^{\pi}_{C'}\big[g(A_{0:t+1}, \bar{X}_{0:t+1})\ \exp\big(-Z^{\pi}_{CC'}(t)\big)\ \big|\ \mathcal{F}_{t+1}\big]\big], \qquad (60)
\end{aligned}
$$

where the last line above follows by noting that $Z^{\pi}_{CC'}(t)$ is measurable with respect to $\mathcal{F}_{t+1}$. We now note that

$$
\begin{aligned}
&\mathbb{E}^{\pi}_{C'}\big[g(A_{0:t+1}, \bar{X}_{0:t+1})\ \exp\big(-Z^{\pi}_{CC'}(t)\big)\ \big|\ \mathcal{F}_{t+1}\big] \\
&= \sum_{a=1}^{K} \sum_{i\in\mathcal{S}} \Big[g(A_{0:t}, \bar{X}_{0:t}, a, i)\ P^{\pi}_{C'}(A_{t+1} = a|\mathcal{F}_{t+1}) \\
&\qquad P^{\pi}_{C'}(\bar{X}_{t+1} = i | A_{t+1} = a, \mathcal{F}_{t+1})\ \exp\big(-Z^{\pi}_{CC'}(t)\big)\Big] \\
&\overset{(a)}{=} \sum_{a=1}^{K} \sum_{i\in\mathcal{S}} \Big[g(A_{0:t}, \bar{X}_{0:t}, a, i)\ P^{\pi}_{C}(A_{t+1} = a|\mathcal{F}_{t+1}) \\
&\qquad P^{\pi}_{C'}(\bar{X}_{t+1} = i | A_{t+1} = a, \mathcal{F}_{t+1})\ \exp\big(-Z^{\pi}_{CC'}(t)\big)\Big], \qquad (61)
\end{aligned}
$$

where in writing (a) above, we make use of the fact that $P^{\pi}_{C'}(A_{t+1} = a|\mathcal{F}_{t+1}) = P^{\pi}_{C}(A_{t+1} = a|\mathcal{F}_{t+1})$ because π selects arms without the knowledge of the underlying assignment of the TPMs. Also, we note that

$$
\begin{aligned}
&P^{\pi}_{C'}(\bar{X}_{t+1} = i | A_{t+1} = a, \mathcal{F}_{t+1})\ \exp\big(-Z^{\pi}_{CC'}(t)\big) \\
&= \exp\big(-Z^{\pi}_{CC'}(t+1)\big)\ P^{\pi}_{C}(\bar{X}_{t+1} = i | A_{t+1} = a, \mathcal{F}_{t+1}). \qquad (62)
\end{aligned}
$$

Substituting (62) in (61) and simplifying, we get

$$
\begin{aligned}
&\mathbb{E}^{\pi}_{C'}\big[g(A_{0:t+1}, \bar{X}_{0:t+1})\ \exp\big(-Z^{\pi}_{CC'}(t)\big)\ \big|\ \mathcal{F}_{t+1}\big] \\
&= \sum_{a=1}^{K} \sum_{i\in\mathcal{S}} \Big[g(A_{0:t}, \bar{X}_{0:t}, a, i)\ P^{\pi}_{C}(A_{t+1} = a|\mathcal{F}_{t+1}) \\
&\qquad P^{\pi}_{C}(\bar{X}_{t+1} = i | A_{t+1} = a, \mathcal{F}_{t+1})\ \exp\big(-Z^{\pi}_{CC'}(t+1)\big)\Big] \\
&= \mathbb{E}^{\pi}_{C}\big[g(A_{0:t+1}, \bar{X}_{0:t+1})\ \exp\big(-Z^{\pi}_{CC'}(t+1)\big)\ \big|\ \mathcal{F}_{t+1}\big]. \qquad (63)
\end{aligned}
$$

Substituting (63) in (60), we get

$$
\begin{aligned}
&\mathbb{E}^{\pi}_{C'}[g(A_{0:t+1}, \bar{X}_{0:t+1})|\mathcal{F}_{t+1}] \\
&= \mathbb{E}^{\pi}_{C}\big[\mathbb{E}^{\pi}_{C}\big[g(A_{0:t+1}, \bar{X}_{0:t+1})\ \exp\big(-Z^{\pi}_{CC'}(t+1)\big)\big|\mathcal{F}_{t+1}\big]\big] \\
&= \mathbb{E}^{\pi}_{C}\big[g(A_{0:t+1}, \bar{X}_{0:t+1})\ \exp\big(-Z^{\pi}_{CC'}(t+1)\big)\big]. \qquad (64)
\end{aligned}
$$

Taking $\mathbb{E}^{\pi}_{C'}[\cdot]$ on both sides of (64), we get the desired result. ■

B. A Lower Bound for $\mathbb{E}^{\pi}_{C}[Z^{\pi}_{CC'}(\tau(\pi))]$ When $\pi \in \Pi(\epsilon)$

We note the following lower bound on the expected LLR. We omit the proof as it follows directly from the proof of [6, Lemma 19].

Lemma 9: Fix $C \in \mathcal{C}$, $C' \in \text{Alt}(C)$, and π such that $P^{\pi}_{C}(\tau(\pi) < \infty) = 1$, $P^{\pi}_{C'}(\tau(\pi) < \infty) = 1$. Then,

1) P^{π}_{C} and $P^{\pi}_{C'}$ are mutually absolutely continuous, and
2) for all $E \in \mathcal{F}_{\tau(\pi)}$ such that $P^{\pi}_{C}(E) > 0$, $P^{\pi}_{C'}(E) > 0$,

$$
\mathbb{E}^{\pi}_{C}[Z^{\pi}_{CC'}(\tau(\pi))] \geq d(P^{\pi}_{C}(E), P^{\pi}_{C'}(E)), \qquad (65)
$$

where $d(x, y)$ denotes the relative entropy between two Bernoulli distributions with parameters x and y.

Fix $\epsilon > 0$. Recall the set $\Pi(\epsilon)$ in (5). A direct consequence of Lemma 9 is that for any $\pi \in \Pi(\epsilon)$, setting $E = \{\omega \in \Omega : \theta(\tau(\pi)) = a^{\star}(C)\}$, where $a^{\star}(C)$ is the index of the best arm in C, noting that $P^{\pi}_{C}(E) \geq 1 - \epsilon$, $P^{\pi}_{C'}(E) \leq \epsilon$ for all $C' \in \text{Alt}(C)$, and using the fact that $x \mapsto d(x, y)$ is monotone increasing for $x < y$ and the $y \mapsto d(x, y)$ is monotone decreasing for any fixed x, we get

$$
\mathbb{E}^{\pi}_{C}[Z^{\pi}_{CC'}(\tau(\pi))] \geq d(\epsilon, 1-\epsilon) \qquad (66)
$$

for all $C' \in \text{Alt}(C)$. Thus, we have

$$
\min_{C'\in\text{Alt}(C)}\ \mathbb{E}^{\pi}_{C}[Z^{\pi}_{CC'}(\tau(\pi))] \geq d(\epsilon, 1-\epsilon) \quad \forall \pi \in \Pi(\epsilon).
$$

C. An Upper Bound for $\mathbb{E}^{\pi}_{C}[Z^{\pi}_{CC'}(\tau(\pi))]$ in Terms of $\mathbb{E}^{\pi}_{C}[\tau(\pi)]$

We first note the following result.

Lemma 10: Fix π and $C \in \mathcal{C}$. For all $(\underline{d}, \underline{i}) \in \mathbb{S}$, $a \in \mathcal{A}$, and $j \in \mathcal{S}$,

$$
\mathbb{E}^{\pi}_{C}\big[N(\tau(\pi), \underline{d}, \underline{i}, a, j)\big] = (P^{a}_{C})^{d_a}(j|i_a)\ \mathbb{E}^{\pi}_{C}\big[N(\tau(\pi), \underline{d}, \underline{i}, a)\big]. \qquad (67)
$$

Proof: [Proof of Lemma 10] We note that

$$\mathbb{E}_C^\pi[\mathbb{E}_C^\pi[N(\tau(\pi),\underline{d},\underline{i},a,j)|X_{a-1}^a]|\tau(\pi)] = \mathbb{E}_C^\pi\left[\mathbb{E}_C^\pi\left[\sum_{t=K}^{\tau(\pi)} 1_{\{\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a,X_t^a=j\}}\middle| X_{a-1}^a\right]\middle|\tau(\pi)\right]. \tag{68}$$

For each t in the range of the summation in (68), the conditional probability term for t may be expressed as

$$\begin{aligned} &P_C^\pi(\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a,X_t^a=j|X_{a-1}^a) \\ &= \Big(P_C^\pi(\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a|X_{a-1}^a) \\ &\qquad \cdot P_C^\pi(X_t^a=j|A_t=a,\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},X_{a-1}^a)\Big) \\ &= P_C^\pi(\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a|X_{a-1}^a)\cdot(P_C^a)^{d_a}(j|i_a). \end{aligned} \tag{69}$$

Plugging (69) back in (68) and taking $\mathbb{E}_C^\pi[\cdot]$ on both sides of (68), we arrive at (67). ∎

From (17), we note that for all $C' \in \text{Alt}(C)$,

$$\begin{aligned} &\mathbb{E}_C^\pi[Z_{CC'}^\pi(\tau(\pi))] \\ &\overset{(a)}{=} \mathbb{E}_C^\pi\left[\sum_{a=1}^K \log\frac{P_C^\pi(X_{a-1}^a)}{P_{C'}^\pi(X_{a-1}^a)}\right] \\ &+ \mathbb{E}_C^\pi\left[\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^K\sum_{j\in\mathcal{S}} N(\tau(\pi),\underline{d},\underline{i},a,j)\log\frac{(P_C^a)^{d_a}(j|i_a)}{(P_{C'}^a)^{d_a}(j|i_a)}\right] \\ &= \mathbb{E}_C^\pi\left[\sum_{a=1}^K \log\frac{P_C^\pi(X_{a-1}^a)}{P_{C'}^\pi(X_{a-1}^a)}\right] \\ &+ \sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^K\sum_{j\in\mathcal{S}} \mathbb{E}_C^\pi\big[N(\tau(\pi),\underline{d},\underline{i},a,j)\big]\log\frac{(P_C^a)^{d_a}(j|i_a)}{(P_{C'}^a)^{d_a}(j|i_a)} \\ &\overset{(b)}{=} \mathbb{E}_C^\pi\left[\sum_{a=1}^K \log\frac{P_C^\pi(X_{a-1}^a)}{P_{C'}^\pi(X_{a-1}^a)}\right] \\ &+ \sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^K\sum_{j\in\mathcal{S}}\Bigg[\mathbb{E}_C^\pi\big[N(\tau(\pi),\underline{d},\underline{i},a)\big] \\ &\qquad\qquad \cdot (P_C^a)^{d_a}(j|i_a)\ \log\frac{(P_C^a)^{d_a}(j|i_a)}{(P_{C'}^a)^{d_a}(j|i_a)}\Bigg] \\ &= \mathbb{E}_C^\pi\left[\sum_{a=1}^K \log\frac{P_C^\pi(X_{a-1}^a)}{P_{C'}^\pi(X_{a-1}^a)}\right] \\ &+ \sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^K \mathbb{E}_C^\pi\big[N(\tau(\pi),\underline{d},\underline{i},a)\big]\ k_{CC'}(\underline{d},\underline{i},a). \end{aligned} \tag{70}$$

In the above chain of equations, (a) follows from the dominated convergence theorem (noting that each row of $(P_C^a)^d$ is mutually absolutely continuous with respect to the corresponding row of $(P_{C'}^a)^d$ for all $d \geq 1$), and (b) follows from Lemma 10. Continuing with (70), we have

$$\begin{aligned} &\mathbb{E}_C^\pi[Z_{CC'}^\pi(\tau(\pi))] \leq \mathbb{E}_C^\pi\left[\sum_{a=1}^K \log\frac{1}{P_{C'}^\pi(X_{a-1}^a)}\right] \\ &+ \sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^K\Big[\mathbb{E}_C^\pi\big[N(\tau(\pi),\underline{d},\underline{i},a)\big] \\ &\qquad\qquad D_{\text{KL}}((P_C^a)^{d_a}(\cdot|i_a)\|(P_{C'}^a)^{d_a}(\cdot|i_a))\Big] \\ &\leq \mathbb{E}_C^\pi\left[\sum_{a=1}^K \log\frac{1}{P_{C'}^\pi(X_{a-1}^a)}\right] \\ &+ (\mathbb{E}_C^\pi[\tau(\pi)-K+1])\cdot \\ &\qquad\Bigg\{\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^K \frac{\mathbb{E}_C^\pi\big[N(\tau(\pi),\underline{d},\underline{i},a)\big]}{\mathbb{E}_C^\pi[\tau(\pi)-K+1]}\ k_{CC'}(\underline{d},\underline{i},a)\Bigg\}, \end{aligned} \tag{71}$$

for all $C' \in \text{Alt}(C)$, from which it follows that

$$\begin{aligned} &\min_{C'\in\text{Alt}(C)}\ \mathbb{E}_C^\pi[Z_{CC'}^\pi(\tau(\pi))] \\ &\leq \min_{C'\in\text{Alt}(C)}\ \mathbb{E}_C^\pi\left[\sum_{a=1}^K \log\frac{1}{P_{C'}^\pi(X_{a-1}^a)}\right] \\ &+ (\mathbb{E}_C^\pi[\tau(\pi)-K+1])\cdot \\ &\Bigg\{\min_{C'\in\text{Alt}(C)}\ \sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^K \frac{\mathbb{E}_C^\pi\big[N(\tau(\pi),\underline{d},\underline{i},a)\big]}{\mathbb{E}_C^\pi[\tau(\pi)-K+1]} \\ &\qquad\qquad D_{\text{KL}}((P_C^a)^{d_a}(\cdot|i_a)\|(P_{C'}^a)^{d_a}(\cdot|i_a))\Bigg\} \\ &\leq \min_{C'\in\text{Alt}(C)}\ \mathbb{E}_C^\pi\left[\sum_{a=1}^K \log\frac{1}{P_{C'}^\pi(X_{a-1}^a)}\right] \\ &\qquad + (\mathbb{E}_C^\pi[\tau(\pi)-K+1])\cdot T^\star(C), \end{aligned} \tag{72}$$

where the supremum in (72) is over all ν that satisfy (21)-(22). The expression within braces in (72) is $T^\star(C)$.

D. The Final Steps

Combining the results of Sections A-B and A-C, we get

$$\begin{aligned} d(\epsilon,1-\epsilon) \leq \min_{C'\in\text{Alt}(C)}\ &\mathbb{E}_C^\pi\left[\sum_{a=1}^K \log\frac{1}{P_{C'}^\pi(X_{a-1}^a)}\right] \\ &+ (\mathbb{E}_C^\pi[\tau(\pi)-K+1])\ T^\star(C). \end{aligned} \tag{73}$$

Noting that (a) $d(\epsilon,1-\epsilon)/\log(1/\epsilon) \to 1$ as $\epsilon \downarrow 0$, and (b) the first term on the right hand side of (73) is bounded from above, by dividing (73) throughout by $d(\epsilon,1-\epsilon)$ and letting $\epsilon \downarrow 0$, we arrive at the lower bound (19).

APPENDIX B
PROOF OF LEMMA 1

Fix an assignment of the TPMs $C \in \mathcal{C}$. In this proof, we establish that $\{(\underline{d}(t),\underline{i}(t)) : t \geq K\}$ irreducible, aperiodic, positive recurrent, and therefore ergodic under the policy π_R^{unif}.

Proof: [Proof of Irreducibility] Fix $(\underline{d},\underline{i}),(\underline{d}',\underline{i}') \in \mathbb{S}_R$, and suppose that $(\underline{d}(T_0),\underline{i}(T_0)) = (\underline{d},\underline{i})$. We now demonstrate that

there exists an integer N (possibly depending on $(\underline{d}, \underline{i})$ and $(\underline{d}', \underline{i}')$) such that

$$P_C^{\pi_R^{\text{unif}}}(\underline{d}(T_0+N)=\underline{d}', \underline{i}(T_0+N)=\underline{i}' | \underline{d}(T_0)=\underline{d}, \underline{i}(T_0)=\underline{i})$$

is strictly positive. Assume without loss of generality that $\underline{d}'$, the vector of arm delays in the destination state $(\underline{d}', \underline{i}')$, is such that $d_1' > d_2' > \cdots > d_K' = 1$. Noting that $P_1, \ldots, P_K$ are TPMs on the finite set $\mathcal{S}$, we use [29, Proposition 1.7] for finite state Markov processes to deduce that there exist integers $M_1, \ldots, M_K$ such that for all $m \geq M := \max\{M_1, \ldots, M_K\}$,

$$P_1^m(j|i) > 0, \ldots, P_K^m(j|i) > 0 \quad \text{for all } i, j \in \mathcal{S}. \tag{74}$$

Order the components of $\underline{d}$, the vector of arm delays in the starting state $(\underline{d}, \underline{i})$, in decreasing order. Under π_R^{unif}, consider the sequence of arm selections and observations as follows: for a total of M time instants, from $t = T_0$ to $t = T_0 + M - 1$, select the arms in a round robin fashion in the decreasing order of their component values in $\underline{d}$. At time $t = T_0 + M$, select arm 1 and observe the state i_1' on it. Thereafter, select arms $2, \ldots, K$ in a round robin fashion in the decreasing order of their component values in $\underline{d}$ until time $t = T_0 + M + d_1' - d_2' - 1$. At time $t = T_0 + M + d_1' - d_2'$, select arm 2 and observe the state i_2' on it. Continue the round robin sampling on arms $3, \ldots, K$ till time $t = T_0 + M + d_1' - d_3' - 1$. At time $t = T_0 + M + d_1' - d_3'$, select arm 3 and observe the state i_3' on it. Continue this process till arm K is selected at time $t = T_0 + M + d_1' - 1$ and the state i_K' is observed on it.

Clearly, because of the round robin selection procedure, the delay of each arm at any time is no more than K. Also, the above sequence of arm selections and observations leads to the state $(\underline{d}', \underline{i}')$ at time $t = T_0 + M + d_1'$. Thus, the probability of starting from the state $(\underline{d}, \underline{i})$ and reaching the state $(\underline{d}', \underline{i}')$ may be lower bounded by the probability that the above sequence of actions and observations occur under π_R^{unif}, which in turn may be lower bounded by

$$\left(\frac{1}{K}\right)^{M+d_1'} \cdot \left[\prod_{a=1}^{K} (P_C^a)^{M+d_a+d_1'-d_a'}(i_a'|i_a)\right]. \tag{75}$$

Noting that $M \leq M + d_1 + d_1' - d_a' \leq M + 2R - 1$, let

$$\bar{\varepsilon} := \min_{M \leq m \leq M+2R-1} \Big\{ (P_C^a)^m(j|i) : \ i, j \in \mathcal{S}, \ a \in \mathcal{A} \Big\}. \tag{76}$$

It is clear that $\bar{\varepsilon} > 0$, and (75) may further be lower bounded by

$$\left(\frac{1}{K}\right)^{M+d_1'} \bar{\varepsilon}^K > 0. \tag{77}$$

Thus, we see that the Markov process $\{(\underline{d}(t), \underline{i}(t) : t \geq K)\}$ is in the state $(\underline{d}', \underline{i}')$ after $N = M + d_1'$ time instants with a strictly positive probability. This establishes irreducibility. ■

Proof of Aperiodicity: It suffices to show that for each $(\underline{d}, \underline{i}) \in \mathbb{S}_R$, there exists N (possibly depending on $(\underline{d}, \underline{i})$) such that the probability of the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ starting from the state $(\underline{d}, \underline{i})$ at some time $t = T_0$ and returning to the state $(\underline{d}, \underline{i})$ after N time instants and also after $N + 1$ time instants is strictly positive. This follows directly from the proof of irreducibility presented above by setting $(\underline{d}', \underline{i}') = (\underline{d}, \underline{i})$ and $N = M + d_1$, where M is such that (74) holds for all $m \geq M$. ■

Proof of Positive Recurrence: This follows from the facts that (a) $\mathbb{S}_R$ is finite, (b) $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is irreducible under π_R^{unif}, and (c) an irreducible Markov process evolving on a finite state space is positive recurrent. ■

APPENDIX C
PROOF OF LEMMA 2

This proof is organised as follows. First, we show in Section C-A that for all $(\underline{d}, \underline{i}) \in \mathbb{S}_R$,

$$\liminf_{n\to\infty} \frac{N(n, \underline{d}, \underline{i})}{n} > 0 \quad \text{a.s..} \tag{78}$$

Next, we show in Section C-B that a.s.,

$$\liminf_{n\to\infty} \ \frac{N(n, \underline{d}, \underline{i}, a)}{n} \begin{cases} > 0, & \text{if } (\underline{d}, \underline{i}) \in \mathbb{S}_{R,a} \\ & \text{or } (\underline{d}, \underline{i}) \notin \bigcup_{a'=1}^{K} \ \mathbb{S}_{R,a'}, \\ = 0, & \text{if } \exists a' \neq a : (\underline{d}, \underline{i}) \in \mathbb{S}_{R,a'}. \end{cases} \tag{79}$$

Using the above results, we establish (39) in Section C-C.

A. Limiting Drift of $N(n, \underline{d}, \underline{i})$

Let M be sufficiently large so that (74) holds for all $m \geq M$. Fix an arbitrary $(\underline{d}, \underline{i}) \in \mathbb{S}_R$, and assume without loss of generality that $\underline{d}$ is such that $d_1 > d_2 > \cdots > d_K = 1$. Let $p_C(\underline{d}, \underline{i})$ denote the probability of the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ starting in the state $(\underline{d}, \underline{i})$ and returning back to the state $(\underline{d}, \underline{i})$ under the assignment of the TPMs C. Following the exposition in Appendix B, it can be shown that $p_C(\underline{d}, \underline{i}) > 0$. Now, the term $N(n, \underline{d}, \underline{i})$ may be lower bounded a.s. by the number of visits to the state $(\underline{d}, \underline{i})$ measured only at times $t = K + M + d_1, K + 2(M + d_1), K + 3(M + d_1)$ and so on until time $t = n$. At each of these time instants, the probability that the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is in the state $(\underline{d}, \underline{i})$ under the assignment of the TPMs C is equal to $p_C(\underline{d}, \underline{i})$. Thus, we have

$$N(n, \underline{d}, \underline{i}) \geq \text{Bin}\left(\frac{n-K+1}{M+d_1}, \ p_C(\underline{d}, \underline{i})\right) \quad \text{a.s.}, \tag{80}$$

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

$$\begin{aligned} &\liminf_{n\to\infty} \frac{N(n, \underline{d}, \underline{i})}{n} \\ &\geq \liminf_{n\to\infty} \frac{\text{Bin}\left(\frac{n-K+1}{M+d_1}, \ \ p_C(\underline{d}, \underline{i})\right)}{n} \\ &= \liminf_{n\to\infty} \frac{\text{Bin}\left(\frac{n-K+1}{M+d_1}, \ p_C(\underline{d}, \underline{i})\right)}{\frac{n-K+1}{M+d_1}} \cdot \frac{n-K+1}{n} \cdot \frac{1}{M+d_1} \\ &\overset{(a)}{=} \frac{p_C(\underline{d}, \underline{i})}{M+d_1} \\ &> 0, \end{aligned} \tag{81}$$

where (a) above is due to the strong law of large numbers. This establishes (78).

B. Limiting Drift of $N(n, \underline{d}, \underline{i}, a)$

If $(\underline{d}, \underline{i}) \in \mathbb{S}_{R,a}$, then $N(n, \underline{d}, \underline{i}, a) = N(n, \underline{d}, \underline{i})$, and consequently

$$\liminf_{n\to\infty} \frac{N(n, \underline{d}, \underline{i}, a)}{n} = \liminf_{n\to\infty} \frac{N(n, \underline{d}, \underline{i})}{n} > 0 \quad \text{a.s..}$$

If $(\underline{d}, \underline{i}) \in \mathbb{S}_{R,a'}$ for some $a' \neq a$, then $N(n, \underline{d}, \underline{i}, a) = 0$ for all $n \geq K$. Thus, it remains to show (79) for $(\underline{d}, \underline{i}) \notin \bigcup_{a'=1}^{K} \mathbb{S}_{R,a'}$. Fix one such arbitrary $(\underline{d}, \underline{i})$ and define

$$S(n, \underline{d}, \underline{i}, a) := \sum_{t=K}^{n} \Big[\mathbb{I}_{\{A_t = a,\, \underline{d}(t) = \underline{d},\, \underline{i}(t) = \underline{i}\}} - P_C^{\pi}(A_t = a,\, \underline{d}(t) = \underline{d},\, \underline{i}(t) = \underline{i} | A_{0:t-1}, \bar{X}_{0:t-1}) \Big]. \tag{82}$$

For each $t \geq K$, because $|\mathbb{I}_{\{A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}\}} - P_C^{\pi}(A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | A_{0:t-1}, \bar{X}_{0:t-1})| \leq 2$ a.s., and $\mathbb{E}_C^{\pi}[\mathbb{I}_{\{A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}\}} - P_C^{\pi}(A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | A_{0:t-1}, \bar{X}_{0:t-1}) | A_{0:t-1}, \bar{X}_{0:t-1}] = 0$ a.s., the collection $\{\mathbb{I}_{\{A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}\}} - P_C^{\pi}(A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | A_{0:t-1}, \bar{X}_{0:t-1})\}_{t \geq K}$ is a bounded martingale difference sequence. Using the concentration result [30, Theorem 1.2A] for bounded martingale difference sequences, and subsequently applying the Borel–Cantelli lemma, we get that

$$\frac{S(n, \underline{d}, \underline{i}, a)}{n} \longrightarrow 0 \quad \text{as } n \to \infty, \quad \text{a.s..} \tag{83}$$

This implies that for every choice of $\varepsilon > 0$, there exists $N_\varepsilon = N_\varepsilon(\underline{d}, \underline{i}, a)$ sufficiently large such that

$$\frac{N(n, \underline{d}, \underline{i}, a)}{n} \geq \frac{1}{n} \sum_{t=K}^{n} P_C^{\pi}(A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | A_{0:t-1}, \bar{X}_{0:t-1}) - \varepsilon \quad \forall\, n \geq N_\varepsilon, \text{ a.s..} \tag{84}$$

Now, for each $t \geq K$, under $\pi = \pi^\star(L, \eta, R)$,

$$\begin{aligned}
&P_C^{\pi}(A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | A_{0:t-1}, \bar{X}_{0:t-1}) \\
&= P_C^{\pi}(A_t = a | \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, A_{0:t-1}, \bar{X}_{0:t-1}) \\
&\qquad \cdot P_C^{\pi}(\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | A_{0:t-1}, \bar{X}_{0:t-1}) \\
&= \lambda_{\eta,R,\bar{C}(t-1)}(a | \underline{d}, \underline{i})\, P_C^{\pi}(\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | A_{0:t-1}, \bar{X}_{0:t-1}) \\
&\overset{(a)}{=} \left\{ \frac{\eta\, \nu^{\text{unif}}_{\bar{C}(t-1),R}(\underline{d}, \underline{i}, a) + (1-\eta)\, \nu^{\star}_{\bar{C}(t-1),R}(\underline{d}, \underline{i}, a)}{\eta\, \mu^{\text{unif}}_{\bar{C}(t-1),R}(\underline{d}, \underline{i}) + (1-\eta) \sum_{a'=1}^{K} \nu^{\star}_{\bar{C}(t-1),R}(\underline{d}, \underline{i}, a')} \cdot \mathbb{I}_{\{\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}\}} \right\} \\
&\geq \frac{\eta}{K} \cdot \mu^{\text{unif}}_{\bar{C}(t-1),R}(\underline{d}, \underline{i}) \cdot \mathbb{I}_{\{(\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}\}} \\
&\geq \frac{\eta}{K} \cdot \mu_R^{\min} \cdot \mathbb{I}_{\{(\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}\}},
\end{aligned} \tag{85}$$

where in writing (a) above, we use the fact that $(\underline{d}(t), \underline{i}(t))$ is measurable with respect to the history $(A_{0:t-1}, \bar{X}_{0:t-1})$, and $\mu_R^{\min}$ in (85) is as defined in (36).

Plugging (85) in (84), we get

$$\frac{N(n, \underline{d}, \underline{i}, a)}{n} \geq \frac{\eta}{K} \cdot \mu_R^{\min} \cdot \frac{N(n, \underline{d}, \underline{i})}{n} - \varepsilon \quad \forall\, n \geq N_\varepsilon, \quad \text{a.s..} \tag{86}$$

Using (81) in (86), we get that

$$\frac{N(n, \underline{d}, \underline{i}, a)}{n - K + 1} \geq \frac{\eta}{K} \cdot \mu_R^{\min} \cdot \frac{p_C(\underline{d}, \underline{i})}{2(M + d_1)} - \varepsilon \tag{87}$$

for all n large, a.s.. Setting $\varepsilon = \frac{\eta}{2K} \cdot \mu_R^{\min} \cdot \frac{p_C(\underline{d}, \underline{i})}{2(M+d_1)}$ establishes (79).

C. Completing the Proof of Lemma 2

Whenever $\liminf_{n\to\infty} N(n, \underline{d}, \underline{i}, a)/n > 0$ a.s., it follows from (79) and the ergodic theorem that under the TPMs assignment C,

$$\frac{N(n, \underline{d}, \underline{i}, a, j)}{N(n, \underline{d}, \underline{i}, a)} \longrightarrow (P_C^a)^{d_a}(j | i_a) \quad \text{as } n \to \infty, \quad \text{a.s..} \tag{88}$$

Under the constraint that the delay of each arm is at most R,

$$\begin{aligned}
\frac{Z^{\pi}_{CC'}(n)}{n} &= \frac{1}{n} \sum_{a=1}^{K} \log \frac{P_C^{\pi}(X^a_{a-1})}{P_{C'}^{\pi}(X^a_{a-1})} \\
&+ \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}_R} \sum_{a=1}^{K} \sum_{j \in \mathcal{S}} \frac{N(n, \underline{d}, \underline{i}, a, j)}{n} \log \frac{(P_C^a)^{d_a}(j | i_a)}{(P_{C'}^a)^{d_a}(j | i_a)}.
\end{aligned} \tag{89}$$

The first term in (89) may be lower bounded as follows: assuming that $X_0^a \sim \phi$ for all $a \in \mathcal{A}$, where ϕ is a probability distribution on $\mathcal{S}$ that is independent of the underlying TPMs C and puts a strictly positive mass on each state in $\mathcal{S}$, it follows that

$$\begin{aligned}
&\frac{1}{n} \sum_{a=1}^{K} \log \frac{P_C^{\pi}(X^a_{a-1})}{P_{C'}^{\pi}(X^a_{a-1})} \\
&\geq \frac{1}{n} \sum_{a=1}^{K} \log P_C^{\pi}(X^a_{a-1}) \\
&= \frac{1}{n} \sum_{j \in \mathcal{S}} \mathbb{I}_{\{X^a_{a-1} = j\}} \sum_{a=1}^{K} \log \left(\sum_{i \in \mathcal{S}} \phi(i)\, (P_C^a)^{a-1}(j | i) \right).
\end{aligned} \tag{90}$$

Because the right hand side of (90) converges to 0 as $n \to \infty$, given any $\varepsilon > 0$, there exists $N_1 = N_1(\varepsilon)$ such that

$$\frac{1}{n} \sum_{a=1}^{K} \log \frac{P_C^{\pi}(X^a_{a-1})}{P_{C'}^{\pi}(X^a_{a-1})} \geq -\varepsilon \quad \text{for all } n \geq N_1, \quad \text{a.s.} \tag{91}$$

The second term in (89) may be expressed as

$$\begin{aligned}
&\sum_{(\underline{d}, \underline{i}) \notin \bigcup_{a'=1}^{K} \mathbb{S}_{R,a'}} \sum_{a=1}^{K} \frac{N(n, \underline{d}, \underline{i}, a, j)}{n} \log \frac{(P_C^a)^{d_a}(j | i_a)}{(P_{C'}^a)^{d_a}(j | i_a)} \\
&+ \sum_{a=1}^{K} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}_{R,a}} \frac{N(n, \underline{d}, \underline{i}, a, j)}{n} \log \frac{(P_C^a)^{d_a}(j | i_a)}{(P_{C'}^a)^{d_a}(j | i_a)}.
\end{aligned} \tag{92}$$

Using the convergence in (88) and noting that $\mathbb{S}_R \times \mathcal{A}$ is finite, we get that there exists $N_2 = N_2(\varepsilon)$ such that for

all $n \geq N_2$, (92) is a.s. lower bounded by

$$\sum_{(\underline{d},\underline{i}) \notin \bigcup_{a'=1}^{K} \mathbb{S}_{R,a'}} \sum_{a=1}^{K} \frac{N(n,\underline{d},\underline{i},a)}{n}\, k_{CC'}(\underline{d},\underline{i},a) + \sum_{a=1}^{K} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}_{R,a}} \frac{N(n,\underline{d},\underline{i},a)}{n}\, k_{CC'}(\underline{d},\underline{i},a) - \varepsilon. \tag{93}$$

Combining (91) and (92), we get that

$$\frac{Z^{\pi}_{CC'}(n)}{n} \geq -2\varepsilon + \sum_{(\underline{d},\underline{i}) \notin \bigcup_{a'=1}^{K} \mathbb{S}_{R,a'}} \sum_{a=1}^{K} \frac{N(n,\underline{d},\underline{i},a)}{n}\, k_{CC'}(\underline{d},\underline{i},a) + \sum_{a=1}^{K} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}_{R,a}} \frac{N(n,\underline{d},\underline{i},a)}{n}\, k_{CC'}(\underline{d},\underline{i},a) \tag{94}$$

for all $n \geq \max\{N_1, N_2\}$, a.s.. Using the results in (78) and (79), we see that the limit infimum of the last two terms in (94) is strictly positive, a.s.. Because $\varepsilon > 0$ is arbitrary, the desired result follows.

Appendix D
Proof of Lemma 3

The policy $\pi = \pi^{\star}(L, \eta, R)$ commits an error if one of the following events is true:

1) The policy does not stop in finite time.
2) The policy stops in finite time and declares an incorrect best arm index.

The event in item 1 above has zero probability, thanks to Lemma 2. Thus, the probability of error of policy $\pi^{\star}(L, \eta, R)$ may be evaluated as follows: for $C \in \mathcal{C}$, recall that $a^{\star}(C)$ is the index of the best arm in C. Then, under the assignment of the TPMs C, the error probability of $\pi^{\star}(L, \eta, R)$ is given by

$$\begin{aligned} &P^{\pi}_{C}(\theta(\tau(\pi)) \neq a) \\ &= P^{\pi}_{C}\Big(\exists\, n, a' \neq a : \tau(\pi) = n,\ \theta(n) = a'\Big) \\ &= P^{\pi}_{C}\Big(\exists\, n, C' \in \text{Alt}(C) : \tau(\pi) = n,\ \theta(n) = a^{\star}(C')\Big). \end{aligned} \tag{95}$$

Let $\mathcal{R}_a(n) := \{\omega \in \Omega : \tau(\pi)(\omega) = n,\, \theta(n)(\omega) = a\}$, $a \in \mathcal{A}$, denote the set of all sample paths for which the policy stops at time n and declares a as the index of the best arm. Clearly, $\{\mathcal{R}_a(n) : a \in \mathcal{A},\ n \geq 0\}$ is a collection of mutually disjoint sets. Therefore, we have

$$\begin{aligned} &P^{\pi}_{C}(\theta(\tau(\pi)) \neq a) \\ &= P^{\pi}_{C}\left(\bigcup_{a' \neq a} \bigcup_{n=0}^{\infty} \mathcal{R}_{a'}(n)\right) \\ &= \sum_{a' \neq a} \sum_{n=0}^{\infty} P^{\pi}_{C}(\tau(\pi) = n,\ \theta(n) = a') \\ &= \sum_{C' \in \text{Alt}(C)} \sum_{n=0}^{\infty} P^{\pi}_{C}(\tau(\pi) = n,\ \theta(n) = a^{\star}(C')) \\ &= \sum_{C' \in \text{Alt}(C)} \sum_{n=0}^{\infty} \int_{\mathcal{R}_{a^{\star}(C')}(n)} dP^{\pi}_{C}(\omega) \\ &= \sum_{C' \in \text{Alt}(C)} \sum_{n=0}^{\infty} \int_{\mathcal{R}_{a^{\star}(C')}(n)} \Big\{ \exp(Z^{\pi}_{C}(n,\omega)) \\ &\qquad\qquad d(A_{0:n}(\omega), \bar{X}_{0:n}(\omega)) \Big\} \\ &= \sum_{C' \in \text{Alt}(C)} \sum_{n=0}^{\infty} \int_{\mathcal{R}_{a^{\star}(C')}(n)} \Big\{ \exp(-Z^{\pi}_{C'C}(n,\omega)) \\ &\qquad\qquad \cdot \exp(Z^{\pi}_{C'}(n,\omega))\ d(A_{0:n}(\omega), \bar{X}_{0:n}(\omega)) \Big\} \\ &\overset{(a)}{\leq} \sum_{C' \in \text{Alt}(C)} \sum_{n=0}^{\infty} \int_{\mathcal{R}_{a^{\star}(C')}(n)} \frac{1}{L(K-1)(K-1)!}\, dP^{\pi}_{C'}(\omega) \\ &= \sum_{C' \in \text{Alt}(C)} \frac{1}{L(K-1)(K-1)!}\, P^{\pi}_{C'}\left(\bigcup_{n=0}^{\infty} \mathcal{R}_{a^{\star}(C')}(n)\right) \\ &\leq \frac{1}{L}, \end{aligned} \tag{96}$$

where (a) above follows by noting that for any $C' \in \text{Alt}(C)$, the condition $M^{\pi}_{C'}(n) \geq \log(L(K-1)(K-1)!)$ holds at the stopping time $\tau(\pi) = n$ on the set $R_{a^{\star}(C')}$. In particular, this implies that $Z^{\pi}_{C'C}(n) \geq \log(L(K-1)(K-1)!)$. Setting $L = 1/\epsilon$ in (96) yields the desired result.

Appendix E
Proof of Proposition 2

We note that for all $C' \in \text{Alt}(C)$, under the policy $\pi = \pi^{\star}_{\text{NS}}(L, \eta, R)$, a.s.,

$$\begin{aligned} \limsup_{n\to\infty} M^{\pi}_{C'}(n) &= \limsup_{n\to\infty} \min_{C'' \in \text{Alt}(C')} Z^{\pi}_{C'C''}(n) \\ &\leq \limsup_{n\to\infty} Z^{\pi}_{C'C}(n) \\ &= \limsup_{n\to\infty} -Z^{\pi}_{CC'}(n) \\ &= -\liminf_{n\to\infty} Z^{\pi}_{CC'}(n) \\ &\leq -\liminf_{n\to\infty} M^{\pi}_{C}(n) \\ &< 0, \end{aligned} \tag{97}$$

where the last line above is due to Lemma 2. Furthermore, for any $\bar{C} \neq C$ such that $a^{\star}(\bar{C}) = a^{\star}(C)$,[6] following the exposition in Appendix C-C with C' replaced by $\bar{C}$, we get that $\liminf_{n\to\infty} Z^{\pi}_{C\bar{C}}(n)/n > 0$ a.s., and therefore $\liminf_{n\to\infty} Z^{\pi}_{C\bar{C}}(n) > 0$ a.s.. This implies that $\liminf_{n\to\infty} Z^{\pi}_{CC'}(n) - Z^{\pi}_{\bar{C}C'}(n) > 0$ a.s. for all $C' \in \text{Alt}(C)$, which in turn implies that $\liminf_{n\to\infty} Z^{\pi}_{CC'}(n) > \limsup_{n\to\infty} Z^{\pi}_{\bar{C}C'}(n)$ a.s. for all $C' \in \text{Alt}(C)$. Noting that $\text{Alt}(C) = \text{Alt}(\bar{C})$, it follows that

$$\liminf_{n\to\infty} M^{\pi}_{C}(n) > \limsup_{n\to\infty} M^{\pi}_{\bar{C}}(n) \quad \text{a.s.} \tag{98}$$

for all $\bar{C} \neq C$ such that $a^{\star}(\bar{C}) = a^{\star}(C)$. Combining (98) and (97), we get that under $\pi = \pi^{\star}_{\text{NS}}(L, \eta, R)$,

$$\bar{C}(n) = C \quad \text{for all } n \text{ large}, \quad \text{a.s.}, \tag{99}$$

[6]Recall that $a^{\star}(C)$ denotes the index of the best arm in C.

when the underlying assignment of the TPMs is C, from which we may deduce that under $\pi = \pi^{\star}_{\text{NS}}(L,\eta,R)$,

$$\begin{aligned}&\lim_{n\to\infty} P^{\pi}\Big(A_n = a \big| A_{0:n-1}, \{(\underline{d}(s),\underline{i}(s)) : K \le s \le n-1\}, \\ &\qquad\qquad \underline{d}(n) = \underline{d}, \underline{i}(n) = \underline{i}\Big) \\ &= \lim_{n\to\infty} \lambda_{\eta,R,\bar{C}(n)}(a|\underline{d},\underline{i}) \\ &= \lambda_{\eta,R,C}(a|\underline{d},\underline{i}) \quad \text{a.s..} \end{aligned} \tag{100}$$

This shows that $\pi^{\star}_{\text{NS}}(L,\eta,R)$ *eventually* makes the process $\{(\underline{d}(t),\underline{i}(t)) : t \ge K\}$ ergodic with $\nu_{\eta,R,C}$ as its ergodic state-action occupancy measure. As a consequence, for all $(\underline{d},\underline{i}) \in \mathbb{S}_R$ and $a \in \mathcal{A}$, it follows that

$$\lim_{n\to\infty} \frac{N(n,\underline{d},\underline{i},a)}{n} = \nu_{\eta,R,C}(\underline{d},\underline{i},a) \text{ a.s..} \tag{101}$$

Therefore, under $\pi = \pi^{\star}_{\text{NS}}(L,\eta,R)$, we have

$$\begin{aligned}&\lim_{n\to\infty} \frac{Z^{\pi}_{CC'}(n)}{n} \\ &= \lim_{n\to\infty} \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},a,j)}{n} \log \frac{(P^a_C)^{d_a}(j|i_a)}{(P^a_{C'})^{d_a}(j|i_a)} \\ &= \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \sum_{j\in\mathcal{S}} \lim_{n\to\infty} \frac{N(n,\underline{d},\underline{i},a,j)}{n} \log \frac{(P^a_C)^{d_a}(j|i_a)}{(P^a_{C'})^{d_a}(j|i_a)} \\ &\overset{(a)}{=} \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \sum_{j\in\mathcal{S}} \Big[\nu_{\eta,R,C}(\underline{d},\underline{i},a)\cdot(P^a_C)^{d_a}(j|i_a) \\ &\qquad\qquad \cdot \log \frac{(P^a_C)^{d_a}(j|i_a)}{(P^a_{C'})^{d_a}(j|i_a)}\Big] \\ &= \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \nu_{\eta,R,C}(\underline{d},\underline{i},a)\, k_{CC'}(\underline{d},\underline{i},a) \quad \text{a.s.,} \end{aligned} \tag{102}$$

where (a) follows from (88). Eq. (42) is then immediate from (102).

Appendix F
Proof of Lemma 4

Because $\pi = \pi^{\star}(L,\eta,R)$ selects arm 1 at time $t=0$, arm 2 at time $t=1$, etc., and arm K at time $t = K-1$, in order to prove the lemma, it suffices to prove that for all $m \ge K$,

$$\lim_{L\to\infty} P^{\pi}_C(\tau(\pi) \le m) = 0. \tag{103}$$

Fix an arbitrary $m \ge K$. Then,

$$\begin{aligned}&\limsup_{L\to\infty} P^{\pi}_C(\tau(\pi) \le m) \\ &= \limsup_{L\to\infty} P^{\pi}_C\Big(\exists\ K \le n \le m,\ C' \in \mathcal{C} : \\ &\qquad\qquad M^{\pi}_{C'}(n) \ge \log(L(K-1)(K-1)!)\Big) \\ &\le \limsup_{L\to\infty} \sum_{C'\in\mathcal{C}} \sum_{n=K}^{m} P^{\pi}_C(M^{\pi}_{C'}(n) \ge \log(L(K-1)(K-1)!)) \\ &\le \limsup_{L\to\infty} \frac{\sum_{C'\in\mathcal{C}} \sum_{n=K}^{m} \mathbb{E}^{\pi}_C[M^{\pi}_{C'}(n)]}{\log(L(K-1)(K-1)!)} \end{aligned} \tag{104}$$

where the first and the second inequalities above follow from the union bound and Markov's inequality respectively. We now show that for each $n \in \{K,\dots,m\}$, the expectation term inside the summation in (104) is finite. This will then imply that the limit supremum on the right-hand side of (104) is equal to 0, thus proving the desired result.

Note that

$$M^{\pi}_{C'}(n) = \min_{\tilde{C}\in \text{Alt}(C')} Z^{\pi}_{C'\tilde{C}}(n) \le Z^{\pi}_{C'\tilde{C}}(n) \quad \text{for all } \tilde{C} \in \text{Alt}(C'). \tag{105}$$

Fix an arbitrary $\tilde{C} \in \text{Alt}(C')$ and recall that

$$\begin{aligned}Z^{\pi}_{C'\tilde{C}}(n) &= \sum_{a=1}^{K} \log \frac{P^{\pi}_{C'}(X^a_{a-1})}{P^{\pi}_{\tilde{C}}(X^a_{a-1})} \\ &\quad + \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \sum_{j\in\mathcal{S}} N(n,\underline{d},\underline{i},a,j)\ \log \frac{(P^a_{C'})^{d_a}(j|i_a)}{(P^a_{\tilde{C}})^{d_a}(j|i_a)}. \end{aligned} \tag{106}$$

Because each row of $(P^a_{C'})^d$ is mutually absolutely continuous with the corresponding row of $(P^a_{\tilde{C}})^d$ for all $d \ge 1$, we may upper bound the second term in (106) as

$$\begin{aligned}&\sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \sum_{j\in\mathcal{S}} N(n,\underline{d},\underline{i},a,j)\ \log \frac{(P^a_{C'})^{d_a}(j|i_a)}{(P^a_{\tilde{C}})^{d_a}(j|i_a)} \\ &\le A \cdot \left(\sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \sum_{j\in\mathcal{S}} N(n,\underline{d},\underline{i},a,j)\right) \\ &= A\cdot(n-K+1) \quad \text{a.s.,} \end{aligned} \tag{107}$$

where

$$A = \max_{\substack{d\in\mathbb{N},\\ i,j\in\mathcal{S},\\ a,a'\in\mathcal{A}}} \left\{\log \frac{P^d_a(j|i)}{P^d_{a'}(j|i)} : P^d_a(j|i) \ne 0, P^d_{a'}(j|i) \ne 0\right\} < \infty. \tag{108}$$

Furthermore, suppose that $X^a_0 \sim \phi$ for all $a \in \mathcal{A}$, where ϕ is a probability distribution on $\mathcal{S}$ that is independent of a and the underlying assignment of the TPMs C. Without loss of generality, let $\phi(i) > 0$ for all $i \in \mathcal{S}$. Then, the first term in (106) may be upper bounded as

$$\begin{aligned}&\sum_{a=1}^{K} \log \frac{P^{\pi}_{C'}(X^a_{a-1})}{P^{\pi}_{\tilde{C}}(X^a_{a-1})} \\ &= \log \frac{P^{\pi}_{C'}(X^1_0)}{P^{\pi}_{\tilde{C}}(X^1_0)} + \sum_{a=2}^{K} \log \frac{P^{\pi}_{C'}(X^a_{a-1})}{P^{\pi}_{\tilde{C}}(X^a_{a-1})} \\ &= \sum_{a=2}^{K} \sum_{j\in\mathcal{S}} \mathbb{I}_{\{X^a_{a-1}=j\}}\ \log \frac{P^{\pi}_{C'}(X^a_{a-1}=j)}{P^{\pi}_{\tilde{C}}(X^a_{a-1}=j)} \\ &= \sum_{a=2}^{K} \sum_{j\in\mathcal{S}} \mathbb{I}_{\{X^a_{a-1}=j\}}\ \log \frac{\sum_{i\in\mathcal{S}}\ \phi(i)\cdot(P^a_{C'})^{a-1}(j|i)}{\sum_{i\in\mathcal{S}}\ \phi(i)\cdot(P^a_{\tilde{C}})^{a-1}(j|i)} \\ &\overset{(a)}{\le} \sum_{a=2}^{K} \sum_{j\in\mathcal{S}} \mathbb{I}_{\{X^a_{a-1}=j\}} \Big[\sum_{i\in\mathcal{S}}\ \frac{\phi(i)\,(P^a_{C'})^{a-1}(j|i)}{\sum_{i'\in\mathcal{S}}\phi(i')\,(P^a_{C'})^{a-1}(j|i')} \\ &\qquad\qquad \cdot \log \frac{(P^a_{C'})^{a-1}(j|i)}{(P^a_{\tilde{C}})^{a-1}(j|i)}\Big] \\ &\le A\,(K-1) \quad \text{a.s.,} \end{aligned} \tag{109}$$

where (a) above follows from the log-sum inequality [31, Theorem 2.7.1]. Combining (107) and (109), we get

$$Z^{\pi}_{C'\tilde{C}}(n) \leq A \cdot n \quad \text{a.s.}, \tag{110}$$

from which it follows that $\mathbb{E}^{\pi}_{C}[Z^{\pi}_{C'\tilde{C}}(n)] \leq A \cdot n$.

Appendix G
Proof of Lemma 5

By the definition of $\tau(\pi)$, we know that under the assignment of the TPMs C,

$$M^{\pi}_{C}(\tau(\pi)-1) < \log(L(K-1)(K-1)!).$$

Therefore, a.s.,

$$\begin{aligned}
1 &= \limsup_{L\to\infty} \frac{\log(L(K-1)(K-1)!)}{\log L} \\
&\geq \limsup_{L\to\infty} \frac{M^{\pi}_{C}(\tau(\pi)-1)}{\log L} \\
&= \limsup_{L\to\infty} \frac{M^{\pi}_{C}(\tau(\pi)-1)}{\tau(\pi)-1} \cdot \frac{\tau(\pi)-1}{\log L} \\
&\geq \left(\limsup_{L\to\infty} \frac{\tau(\pi)-1}{\log L}\right) \cdot \\
&\quad \left(\min_{C'\in \text{Alt}(C)} \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \nu_{\eta,R,C}(\underline{d},\underline{i},a)\ k_{CC'}(\underline{d},\underline{i},a)\right),
\end{aligned} \tag{111}$$

where the last line above is due to (45) and the fact that the increment $M^{\pi}_{C}(n) - M^{\pi}_{C}(n-1)$ is bounded for all $n \geq K$. We then note that

$$\begin{aligned}
&\min_{C'\in \text{Alt}(C)} \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \nu_{\eta,R,C}(\underline{d},\underline{i},a)\ k_{CC'}(\underline{d},\underline{i},a) \\
&\geq \eta \min_{C'\in \text{Alt}(C)} \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \nu^{\text{unif}}_{C,R}(\underline{d},\underline{i},a)\ k_{CC'}(\underline{d},\underline{i},a) \\
&+ (1-\eta) \min_{C'\in \text{Alt}(C)} \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \nu^{\star}_{C,R}(\underline{d},\underline{i},a)\ k_{CC'}(\underline{d},\underline{i},a) \\
&\geq \eta\ T^{\text{unif}}_{R}(C) + (1-\eta)\ T^{\star}_{R}(C),
\end{aligned} \tag{112}$$

thus establishing the desired result.

Appendix H
Proof of Proposition 3

We prove here that the family $\{\tau(\pi)/\log L : L > 1\}$, where $\pi = \pi^{\star}(L,\eta,R)$, is uniformly integrable. Towards this, we show that

$$\limsup_{L\to\infty}\ \mathbb{E}^{\pi}_{C}\left[\left(\frac{\tau(\pi)}{\log L}\right)^2\right] < \infty. \tag{113}$$

Then, [32, Lemma 3, pp. 227] implies the desired uniform integrability result from (113) (by using $G(t) = t^2$ in [32, Lemma 3, pp. 227]). Let

$$\psi(L) := \frac{\log(L(K-1)(K-1)!)}{\log L}, \tag{114}$$

and let $\pi^{\star}_{C} = \pi^{\star}_{C}(L,\eta,R)$ denote the version of the policy R-DCR-BAI that stops only when the event

$$M^{\pi^{\star}_{C}}_{C}(n) \geq \log(L(K-1)(K-1)!)$$

occurs. Clearly, $\tau(\pi^{\star}_{C}) \geq \tau(\pi)$ a.s.. Then,

$$\begin{aligned}
&\limsup_{L\to\infty}\ \mathbb{E}^{\pi}_{C}\left[\left(\frac{\tau(\pi)}{\log L}\right)^2\right] \\
&= \limsup_{L\to\infty} \int_0^{\infty} P^{\pi}_{C}\left(\left(\frac{\tau(\pi)}{\log L}\right)^2 > x\right) dx \\
&= \limsup_{L\to\infty} \int_0^{\infty} P^{\pi}_{C}\Big(\tau(\pi) > (\sqrt{x})(\log L)\Big)\, dx \\
&\leq \limsup_{L\to\infty} \int_0^{\infty} P^{\pi}_{C}\Big(\tau(\pi^{\star}_{C}) > (\sqrt{x})(\log L)\Big)\, dx \\
&\overset{(a)}{\leq} \limsup_{L\to\infty} \left\{\psi(L) + \int_{\psi(L)}^{\infty} P^{\pi}_{C}\left(\tau(\pi^{\star}_{C}) > (\sqrt{x})(\log L)\right)\, dx\right\} \\
&\leq \limsup_{L\to\infty}\ \psi(L) \\
&+ \limsup_{L\to\infty} \sum_{n\geq\sqrt{\psi(L)}\log L}^{\infty} \frac{2n+1}{(\log L)^2}\ P^{\pi}_{C}(\tau(\pi^{\star}_{C}) > n) \\
&\leq 1 + \limsup_{L\to\infty} \sum_{n\geq\sqrt{\psi(L)}\log L}^{\infty} \Bigg[\frac{2n+1}{(\log L)^2} \\
&\qquad \cdot P^{\pi}_{C}(M^{\pi}_{C}(n) < \log(L(K-1)(K-1)!))\Bigg],
\end{aligned} \tag{115}$$

where (a) above follows by upper bounding the probability term by 1 for all $x \leq \psi(L)$. Below, we show that $P^{\pi}_{C}(M^{\pi}_{C}(n) < \log(L(K-1)(K-1)!))$ is $O(1/n^3)$, which implies that the infinite sum in (115) is finite. This will then prove that the right-hand side of (115) is finite. Note that

$$\begin{aligned}
&P^{\pi}_{C}(M^{\pi}_{C}(n) < \log(L(K-1)(K-1)!)) \\
&= P^{\pi}_{C}\left(\min_{C'\in \text{Alt}(C)} Z^{\pi}_{CC'}(n) < \log(L(K-1)(K-1)!)\right) \\
&\leq \sum_{C'\in \text{Alt}(C)} P^{\pi}_{C}\left(Z^{\pi}_{CC'}(n) < \log(L(K-1)(K-1)!)\right),
\end{aligned} \tag{116}$$

where the last line above follows from the union bound.

We now show that each term inside the summation in (116) is $O(1/n^3)$. Fix $C' \in \text{Alt}(C)$ and observe that

$$\begin{aligned}
&P^{\pi}_{C}\left(Z^{\pi}_{CC'}(n) < \log(L(K-1)(K-1)!)\right) \\
&= P^{\pi}_{C}\left(\frac{Z^{\pi}_{CC'}(n)}{n} < \frac{\log(L(K-1)(K-1)!)}{n}\right) \\
&= P^{\pi}_{C}\Bigg(\frac{1}{n}\sum_{a=1}^{K} \log\frac{P^{\pi}_{C}(X^{a}_{a-1})}{P^{\pi}_{C'}(X^{a}_{a-1})} \\
&+ \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R} \sum_{a=1}^{K} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},a,j)}{n}\ \log\frac{(P^{a}_{C})^{d_a}(j|i_a)}{(P^{a}_{C'})^{d_a}(j|i_a)} \\
&\qquad < \frac{\log(L(K-1)(K-1)!)}{n}\Bigg)
\end{aligned}$$

$$\leq P_C^\pi\left(\frac{1}{n}\sum_{a=1}^{K}\log\frac{P_C^\pi(X_{a-1}^a)}{P_{C'}^\pi(X_{a-1}^a)}<-\varepsilon\right) \tag{117}$$

$$+P_C^\pi\Bigg(\sum_{(\underline{d},\underline{i})\in\mathbb{S}_R}\sum_{a=1}^{K}\sum_{j\in\mathcal{S}}\frac{N(n,\underline{d},\underline{i},a,j)}{n}\log\frac{(P_C^a)^{d_a}(j|i_a)}{(P_{C'}^a)^{d_a}(j|i_a)} - \sum_{(\underline{d},\underline{i})\in\mathbb{S}_R}\sum_{a=1}^{K}\frac{N(n,\underline{d},\underline{i},a)}{n}k_{CC'}(\underline{d},\underline{i},a)<-\varepsilon\Bigg) \tag{118}$$

$$+P_C^\pi\Bigg(\sum_{(\underline{d},\underline{i})\in\mathbb{S}_R}\sum_{a=1}^{K}\frac{N(n,\underline{d},\underline{i},a)}{n}k_{CC'}(\underline{d},\underline{i},a)-2\varepsilon<\frac{\log(L(K-1)(K-1)!)}{n}\Bigg) \tag{119}$$

for all $\varepsilon>0$. Fixing ε, we handle (117)-(119) individually. We also show how to choose ε (later in (129)).

1) Handling (117): This term is equal to 0 for all sufficiently large values of n because the left hand side inside the probability term converges a.s. to 0, whereas the right hand side is negative.
2) Handling (118): This term may be expressed as

$$P_C^\pi\Bigg(\sum_{(\underline{d},\underline{i})\in\mathbb{S}_R}\sum_{a=1}^{K}\sum_{j\in\mathcal{S}}\left[\frac{N(n,\underline{d},\underline{i},a,j)}{n}-\frac{N(n,\underline{d},\underline{i},a)}{n}(P_C^a)^{d_a}(j|i_a)\right]\log\frac{(P_C^a)^{d_a}(j|i_a)}{(P_{C'}^a)^{d_a}(j|i_a)}<-\varepsilon\Bigg). \tag{120}$$

It is easy to check that

$$M_n:=\sum_{(\underline{d},\underline{i})\in\mathbb{S}_R}\sum_{a=1}^{K}\sum_{j\in\mathcal{S}}\left[N(n,\underline{d},\underline{i},a,j)-N(n,\underline{d},\underline{i},a)\;(P_C^a)^{d_a}(j|i_a)\right]\log\frac{(P_C^a)^{d_a}(j|i_a)}{(P_{C'}^a)^{d_a}(j|i_a)} \tag{121}$$

is a bounded martingale whose quadratic variation process

$$\begin{aligned}
&\langle M_n\rangle\\
&:=\sum_{t=K}^{n}\mathbb{E}_C^\pi\big[M_t^2|\mathcal{F}_t\big]\\
&\leq\sum_{t=K}^{n}\mathbb{E}_C^\pi\Bigg[\sum_{(\underline{d},\underline{i})\in\mathbb{S}_R}\sum_{a=1}^{K}\sum_{j\in\mathcal{S}}\Bigg\{\mathbb{I}_{\{\underline{d}(t)=\underline{d},\;\underline{i}(t)=\underline{i},\;A_t=a\}}\\
&\qquad\cdot\left(\mathbb{I}_{\{\bar{X}_t=j\}}-(P_C^a)^{d_a}(j|i_a)\right)^2\\
&\qquad\cdot\left(\log\frac{(P_C^a)^{d_a}(j|i_a)}{(P_{C'}^a)^{d_a}(j|i_a)}\right)^2\Bigg\}\Bigg|\mathcal{F}_t\Bigg]\\
&\leq 4\;A^2\\
&\cdot\sum_{t=K}^{n}\mathbb{E}_C^\pi\Bigg[\sum_{(\underline{d},\underline{i})\in\mathbb{S}_R}\sum_{a=1}^{K}\sum_{j\in\mathcal{S}}\mathbb{I}_{\{\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a\}}\Bigg|\mathcal{F}_t\Bigg]\\
&\leq n\;(4A^2|\mathcal{S}|)\quad\text{a.s.},
\end{aligned} \tag{122}$$

where A above is as defined in (108). We then have

$$\begin{aligned}
P_C^\pi\left(M_n<-n\varepsilon\right)&\leq P_C^\pi\left(|M_n|>n\varepsilon\right)\\
&\leq P^\pi\left(\sup_{K\leq t\leq n}|M_t|>n\varepsilon\right)\\
&\overset{(a)}{\leq}\frac{\mathbb{E}_C^\pi\left[\left(\sup_{K\leq t\leq n}|M_t|\right)^6\right]}{n^6\;\varepsilon^6}\\
&\overset{(b)}{\leq}\frac{B}{n^6\;\varepsilon^6}\;\mathbb{E}_C^\pi[|\langle M_n\rangle|^3]\\
&\overset{(c)}{\leq}\frac{B}{n^6\;\varepsilon^6}\cdot n^3\cdot\left(4A^2|\mathcal{S}|\right)^3\\
&=\frac{A'}{n^3},
\end{aligned} \tag{123}$$

where (a) above is due to Markov's inequality, (b) is due to Burkholder's inequality [33, pp. 414], and (c) follows from (122). We have thus shown that (118) is $O(1/n^3)$.

3) Handling (119): Observe that (119) may be upper bounded by

$$P_C^\pi\Bigg(\frac{N(n,\underline{d},\underline{i},a)}{n}k_{CC'}(\underline{d},\underline{i},a)-2\varepsilon<\frac{\log(L(K-1)(K-1)!)}{n}\Bigg) \tag{124}$$

for all $(\underline{d},\underline{i})\in\mathbb{S}_R$ and $a\in\mathcal{A}$. Fix $\underline{d}=\underline{d}^\star=(K,K-1,\ldots,1)$, $\underline{i}=\underline{i}^\star\in\mathcal{S}^K$, and $a\in\mathcal{A}$. Using the convergence in (101), we get that for every $\epsilon'>0$, under the policy $\pi_{\mathrm{NS}}^\star(L,\eta,R)$,

$$\frac{N(n,\underline{d}^\star,\underline{i}^\star,a)}{n}\geq\nu_{\eta,R,C}(\underline{d}^\star,\underline{i}^\star,a)(1-\epsilon') \tag{125}$$

for all n large, a.s.. Leveraging this, we define

$$E_n:=\left\{\omega\in\Omega:\frac{N(n,\underline{d}^\star,\underline{i}^\star,a,\omega)}{n}\geq\nu_{\eta,R,C}(\underline{d}^\star,\underline{i}^\star,a)(1-\epsilon')\right\}, \tag{126}$$

and write (124) as the sum of two terms, one of which is

$$\begin{aligned}
&P_C^\pi\Bigg(E_n\bigcap\left\{\frac{N(n,\underline{d}^\star,\underline{i}^\star,a)}{n}k_{CC'}(\underline{d}^\star,\underline{i}^\star,a)-2\varepsilon<\frac{\log(L(K-1)(K-1)!)}{n}\right\}\Bigg)\\
&\leq P_C^\pi\Bigg((1-\epsilon')\;\nu_{\eta,R,C}(\underline{d}^\star,\underline{i}^\star,a)\;k_{CC'}(\underline{d}^\star,\underline{i}^\star,a)-2\varepsilon<\frac{\log(L(K-1)(K-1)!)}{n}\Bigg),
\end{aligned} \tag{127}$$

and the other is

$$\begin{aligned}
&P_C^\pi\Bigg(E_n^c\bigcap\left\{\frac{N(n,\underline{d}^\star,\underline{i}^\star,a)}{n}k_{CC'}(\underline{d}^\star,\underline{i}^\star,a)-2\varepsilon<\frac{\log(L(K-1)(K-1)!)}{n}\right\}\Bigg)\\
&\leq P_C^\pi\left(\frac{N(n,\underline{d}^\star,\underline{i}^\star,a)}{n}<\nu_{\eta,R,C}(\underline{d}^\star,\underline{i}^\star,a)(1-\epsilon')\right).
\end{aligned} \tag{128}$$

We shall see how to choose ϵ' (later in (135)). Choosing ε such that

$$(1-\epsilon')\ \nu_{\eta,R,C}(\underline{d}^\star,\underline{i}^\star,a)\ k_{CC'}(\underline{d}^\star,\underline{i}^\star,a)-2\varepsilon>0, \tag{129}$$

we see that the left hand side of the probability term in (127) is strictly positive, whereas the right hand side goes to 0 as $n\to\infty$. Thus, for all sufficiently large values of n, (127) equals 0.

It now remains to show that (128) is $O(1/n^3)$.

Showing that (128) *is* $O(1/n^3)$*:*

Let M be a large, positive integer such that (74) holds for all $m\geq M$. Along the lines of the proof of irreducibility presented in Appendix B, it can be shown that for all $(\underline{d},\underline{i})\in\mathbb{S}_R$ and T_0, the probability of transitioning from $(\underline{d}(T_0),\underline{i}(T_0))=(\underline{d},\underline{i})$ to $(\underline{d}(T_0{+}N),\underline{i}(T_0{+}N))=(\underline{d}^\star,\underline{i}^\star)$ is lower bounded by

$$\rho:=\left(\frac{\eta}{K}\ \mu_R^{\min}\right)^{M+K}\ \bar{\varepsilon}^K>0 \tag{130}$$

for $N=M+K$, where $\mu_R^{\min}$ is as defined in (36), and $\bar{\varepsilon}$ is as defined in (76). Eq. (130) states that under the policy R-DCR-BAI, the probability of starting from any $(\underline{d},\underline{i})\in\mathbb{S}_R$ and reaching the state $(\underline{d}^\star,\underline{i}^\star)$ after $N=M+K$ time instants may be lower bounded uniformly over all starting states. In the literature on controlled Markov processes, such a phenomenon is referred to as Doeblin's minorisation condition [34, Eq. (5)]. Therefore, for all $n\geq M+K$,

$$P_C^\pi(\underline{d}(n)=\underline{d}^\star,\underline{i}(n)=\underline{i}^\star)\geq\rho. \tag{131}$$

We now note that for all $n\geq M+K$,

$$\begin{aligned}N(n,\underline{d}^\star,\underline{i}^\star,a)&=\sum_{t=K}^{n}\mathbb{I}_{\{\underline{d}(t)=\underline{d}^\star,\ \underline{i}(t)=\underline{i}^\star,\ A_t=a\}}\\&\geq\sum_{t=M+K}^{n}\mathbb{I}_{\{\underline{d}(t)=\underline{d}^\star,\ \underline{i}(t)=\underline{i}^\star,\ A_t=a\}}\quad\text{a.s..}\end{aligned} \tag{132}$$

Denoting the right hand side of (132) by $N'(n,\underline{d}^\star,\underline{i}^\star,a)$, we note that

$$\begin{aligned}&\mathbb{E}_C^\pi[N'(n,\underline{d}^\star,\underline{i}^\star,a)]\\&=\sum_{t=M+K}^{n}P_C^\pi(\underline{d}(t)=\underline{d}^\star,\ \underline{i}(t)=\underline{i}^\star,\ A_t=a)\\&\geq\sum_{t=M+K}^{n}\rho\cdot P_C^\pi(A_t=a|\underline{d}(t)=\underline{d}^\star,\ \underline{i}(t)=\underline{i}^\star)\\&=\sum_{t=M+K}^{n}\rho\cdot\lambda_{\eta,R,\bar{C}(t-1)}(a|\underline{d}^\star,\underline{i}^\star)\\&\geq(n-M-K+1)\cdot\rho\cdot\left(\frac{\eta}{K}\ \mu_R^{\min}\right),\end{aligned} \tag{133}$$

where $\mu_R^{\min}$ is as defined in (36). Using (133) in (128), we arrive at the series of inequalities leading up to (134), shown at the bottom of the page, for all $n\geq M+K$. Noting that $\frac{n-M-K+1}{n}\geq\frac{1}{2}$ for all n sufficiently large, we choose ϵ' such that

$$\nu_{\eta,R,C}(\underline{d}^\star,\underline{i}^\star,a)(1-\epsilon')-\frac{\rho\,\eta\,\mu_R^{\min}}{2K}<0.$$

For instance, it suffices to set

$$\epsilon'=1-\frac{1}{2}\ \frac{\rho\,\eta\,\mu_R^{\min}}{2K\ \nu_{\eta,R,C}(\underline{d}^\star,\underline{i}^\star,a)}. \tag{135}$$

For this choice of ϵ', it follows that the probability term in (134) may be bounded above exponentially by using concentration inequalities for sub-gaussian random variables [35, p. 25]; here, $N'(n,\underline{d}^\star,\underline{i}^\star,a)$ is a sum of (not necessarily independent) indicator random variables, each of which is sub-gaussian with variance factor $1/4$. This implies that $N'(n,\underline{d}^\star,\underline{i}^\star,a)$ is also sub-gaussian [8, Lemma 17]. Therefore, for all sufficiently large n, the probability term in (134) is $O(1/n^3)$.

APPENDIX I
PROOF OF LEMMA 7

Suppose that $P_k(\cdot|i)=\mu_k(\cdot)$ for all $k=1,\ldots,K$ and $i\in\mathcal{S}$. Fixing $C\in\mathcal{C}$, it follows that for

$$\begin{aligned}&P_C^\pi\Big(N(n,\underline{d}^\star,\underline{i}^\star,a)<n\,\nu_{\eta,R,C}(\underline{d}^\star,\underline{i}^\star,a)(1-\epsilon')\Big)\\&\leq P_C^\pi\Big(N'(n,\underline{d}^\star,\underline{i}^\star,a)<n\,\nu_{\eta,R,C}(\underline{d}^\star,\underline{i}^\star,a)(1-\epsilon')\Big)\\&=P_C^\pi\Big(N'(n,\underline{d}^\star,\underline{i}^\star,a)-\mathbb{E}_C^\pi[N'(n,\underline{d}^\star,\underline{i}^\star,a)]<n\,\nu_{\eta,R,C}(\underline{d}^\star,\underline{i}^\star,a)(1-\epsilon')-\mathbb{E}_C^\pi[N'(n,\underline{d}^\star,\underline{i}^\star,a)]\Big)\\&\leq P_C^\pi\Big(N'(n,\underline{d}^\star,\underline{i}^\star,a)-\mathbb{E}_C^\pi[N'(n,\underline{d}^\star,\underline{i}^\star,a)]<n\,\nu_{\eta,R,C}(\underline{d}^\star,\underline{i}^\star,a)(1-\epsilon')\\&\qquad\qquad-(n-M-K+1)\cdot\rho\cdot\frac{\eta}{K}\cdot\mu_R^{\min}\Big)\\&\leq P_C^\pi\Big(N'(n,\underline{d}^\star,\underline{i}^\star,a)-\mathbb{E}_C^\pi[N'(n,\underline{d}^\star,\underline{i}^\star,a)]<n\Big\{\nu_{\eta,R,C}(\underline{d}^\star,\underline{i}^\star,a)(1-\epsilon')\\&\qquad\qquad-\left(\frac{n-M-K+1}{n}\right)\cdot\frac{\rho\,\eta\,\mu_R^{\min}}{K}\Big\}\Big).\end{aligned} \tag{134}$$

$$
\begin{aligned}
T_R^\star(C) &= \sup_{\nu} \min_{C' \in \text{Alt}(C)} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}_R} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a)\ D_{\text{KL}}((P_C^a)^{d_a}(\cdot|i_a) \| (P_{C'}^a)^{d_a}(\cdot|i_a)) \\
&= \sup_{\nu} \min_{C' \in \text{Alt}(C)} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}_R} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a)\ D_{\text{KL}}(\mu_C^a \| \mu_{C'}^a) \\
&= \sup_{\nu} \min_{C' \in \text{Alt}(C)} \left\{ \sum_{(\underline{d},\underline{i}) \in \mathbb{S}_1} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a)\ D_{\text{KL}}(\mu_C^a \| \mu_{C'}^a) + \sum_{(\underline{d},\underline{i}) \in \mathbb{S}_2} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a)\ D_{\text{KL}}(\mu_C^a \| \mu_{C'}^a) \right\} \\
&= \sup_{\nu} \min_{C' \in \text{Alt}(C)} \left\{ \sum_{a=1}^{K} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}_{R,a}} \nu(\underline{d},\underline{i},a)\ D_{\text{KL}}(\mu_C^a \| \mu_{C'}^a) + \sum_{(\underline{d},\underline{i}) \in \mathbb{S}_2} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a)\ D_{\text{KL}}(\mu_C^a \| \mu_{C'}^a) \right\} \\
&= \sup_{\nu} \min_{C' \in \text{Alt}(C)} \left\{ \sum_{a=1}^{K} D_{\text{KL}}(\mu_C^a \| \mu_{C'}^a) \left(\sum_{(\underline{d},\underline{i}) \in \mathbb{S}_{R,a}} \nu(\underline{d},\underline{i},a) + \sum_{(\underline{d},\underline{i}) \in \mathbb{S}_2} \nu(\underline{d},\underline{i},a) \right) \right\} \\
&\overset{(a)}{=} \sup_{\nu} \min_{C' \in \text{Alt}(C)} \left\{ \sum_{a=1}^{K} D_{\text{KL}}(\mu_C^a \| \mu_{C'}^a) \left(\sum_{(\underline{d},\underline{i}) \in \mathbb{S}_{R,a}} \nu(\underline{d},\underline{i},a) + \sum_{a' \neq a} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}_{R,a'}} \nu(\underline{d},\underline{i},a) + \sum_{(\underline{d},\underline{i}) \in \mathbb{S}_2} \nu(\underline{d},\underline{i},a) \right) \right\} \\
&= \sup_{\nu} \min_{C' \in \text{Alt}(C)} \left\{ \sum_{a=1}^{K} D_{\text{KL}}(\mu_C^a \| \mu_{C'}^a) \left(\sum_{(\underline{d},\underline{i}) \in \mathbb{S}_R} \nu(\underline{d},\underline{i},a) \right) \right\} \\
&= \sup_{\kappa} \min_{C' \in \text{Alt}(C)} \sum_{a=1}^{K} \kappa(a)\ D_{\text{KL}}(\mu_C^a \| \mu_{C'}^a), \qquad (138)
\end{aligned}
$$

all $d \in \mathbb{N}$, $i \in \mathcal{S}$, and $C' \in \text{Alt}(C)$,

$$D_{\text{KL}}((P_C^a)^d(\cdot|i) \| (P_{C'}^a)^d(\cdot|i)) = D_{\text{KL}}(\mu_C^a \| \mu_{C'}^a), \qquad (136)$$

where μ_C^a denotes the stationary distribution associated with the TPM P_C^a. As a consequence of (136), we have

$$
\begin{aligned}
T^\star(C) &= \sup_{\nu} \min_{C' \in \text{Alt}(C)} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a)\ k_{CC'}(\underline{d},\underline{i},a) \\
&= \sup_{\nu} \min_{C' \in \text{Alt}(C)} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a)\ D_{\text{KL}}(\mu_C^a \| \mu_{C'}^a) \\
&= \sup_{\kappa} \min_{C' \in \text{Alt}(C)} \sum_{a=1}^{K} \kappa(a)\ D_{\text{KL}}(\mu_C^a \| \mu_{C'}^a), \qquad (137)
\end{aligned}
$$

where $\kappa(a) := \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \nu(\underline{d},\underline{i},a)$ for all $a \in \mathcal{A}$, and the supremum in (137) is over all κ which are probability distributions on the set of arms $\mathcal{A}$.

For a fixed $R \in \mathbb{N} \cap (K, \infty)$, suppose that

$$\mathbb{S}_1 = \bigcup_{a=1}^{K} \mathbb{S}_{R,a}, \quad \mathbb{S}_2 = \mathbb{S}_R \setminus \mathbb{S}_1.$$

Then, $T_R^\star(C)$ may be simplified as shown in the series of equalities leading up to (138), in which (a) follows from the observation that any ν participating in the supremum meets the R-max-delay constraint in (25), and therefore satisfies $\nu(\underline{d},\underline{i},a) = 0$ for all $(\underline{d},\underline{i}) \in \mathbb{S}_{R,a'}$, $a' \neq a$. From (137) and (138), shown at the top of the page, we see that $T_R^\star(C) = T^\star(C)$ for all R, thus proving that $\lim_{R \to \infty} T_R^\star(C) = T^\star(C)$ in the special case when each of the arm TPMs have identical rows.

REFERENCES

[1] P. N. Karthik, K. S. Reddy, and V. Y. F. Tan, "Best restless Markov arm identification," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2022, pp. 648–653.

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

[3] A. E. Albert, "The sequential design of experiments for infinitely many states of nature," *Ann. Math. Statist.*, vol. 32, no. 3, pp. 774–799, Sep. 1961.

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

[5] V. Moulos, "Optimal best Markovian arm identification with fixed confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.

[6] E. Kaufmann, O. Cappé, and A. Garivier, "On the complexity of best-arm identification in multi-armed bandit models," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1–42, 2016.

[7] P. N. Karthik and R. Sundaresan, "Detecting an odd restless Markov arm with a trembling hand," *IEEE Trans. Inf. Theory*, vol. 67, no. 8, pp. 5230–5258, Aug. 2021.

[8] P. N. Karthik and R. Sundaresan, "Learning to detect an odd restless Markov arm with a trembling hand," 2021, *arXiv:2105.03603*.

[9] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, Mar. 1985.

[10] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—Part II: Markovian rewards," *IEEE Trans. Autom. Control*, vol. AC-32, no. 11, pp. 977–982, Nov. 1987.

[11] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5588–5611, Aug. 2012.

[12] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Found. Trends Mach. Learn.*, vol. 5, no. 1, pp. 1–122, 2012.

[13] S. Mannor and J. N. Tsitsiklis, "The sample complexity of exploration in the multi-armed bandit problem," *J. Mach. Learn. Res.*, vol. 5, no. 6, pp. 623–648, 2004.

[14] E. Even-Dar, S. Mannor, and Y. Mansour, "PAC bounds for multi-armed bandit and Markov decision processes," in *Proc. Int. Conf. Comput. Learn. Theory*. New York, NY, USA: Springer, 2002, pp. 255–270.

[15] Z. Karnin, T. Koren, and O. Somekh, "Almost optimal exploration in multi-armed bandits," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1238–1246.
[16] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck, "lil'UCB: An optimal exploration algorithm for multi-armed bandits," in *Proc. Conf. Learn. Theory*, 2014, pp. 423–439.
[17] J.-Y. Audibert, S. Bubeck, and R. Munos, "Best arm identification in multi-armed bandits," in *Proc. Conf. Learn. Theory*, 2010, pp. 41–53.
[18] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone, "PAC subset selection in stochastic multi-armed bandits," in *Proc. Int. Conf. Mach. Learn.*, vol. 12, 2012, pp. 655–662.
[19] E. Kaufmann and S. Kalyanakrishnan, "Information complexity in bandit subset selection," in *Proc. Conf. Learn. Theory*, 2013, pp. 228–251.
[20] S. Gupta, G. Joshi, and O. Yağan, "Best-arm identification in correlated multi-armed bandits," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 2, pp. 549–563, May 2021.
[21] Z. Zhong, W. C. Cheung, and V. Tan, "Best arm identification for cascading bandits in the fixed confidence setting," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 11481–11491.
[22] O. Dekel, J. Ding, T. Koren, and Y. Peres, "Bandits with switching costs: $T^{2/3}$ regret," in *Proc. ACM Symp. Theory Comput.*, 2014, pp. 459–467.
[23] Z. Zhong, W. C. Cheung, and V. Tan, "Probabilistic sequential shrinking: A best arm identification algorithm for stochastic bandits with corruptions," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12772–12781.
[24] P. N. Karthik and R. Sundaresan, "Learning to detect an odd Markov arm," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4324–4348, Jul. 2020.
[25] A. Deshmukh, S. Bhashyam, and V. V. Veeravalli, "Controlled sensing for composite multihypothesis testing with application to anomaly detection," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018, pp. 2109–2113.
[26] A. Deshmukh, V. V. Veeravalli, and S. Bhashyam, "Sequential controlled sensing for composite multihypothesis testing," *Sequential Anal.*, vol. 40, no. 2, pp. 259–289, Apr. 2021.
[27] G. R. Prabhu, S. Bhashyam, A. Gopalan, and R. Sundaresan, "Sequential multi-hypothesis testing in multi-armed bandit problems: An approach for asymptotic optimality," *IEEE Trans. Inf. Theory*, vol. 68, no. 7, pp. 4790–4817, Jul. 2022.
[28] V. S. Borkar, "Control of Markov chains with long-run average cost criterion," in *Stochastic Differential Systems, Stochastic Control Theory and Applications*. New York, NY, USA: Springer, 1988, pp. 57–77.
[29] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times*, vol. 107. Providence, RI, USA: American Mathematical Society, 2017.
[30] V. H. De La Peña, "A general class of exponential inequalities for Martingales and ratios," *Ann. Probab.*, vol. 27, no. 1, pp. 537–564, 1999.
[31] M. Thomas and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2006.
[32] A. N. Shiryaev, *Probability-1*, vol. 95. New York, NY, USA: Springer, 2016.
[33] Y. S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales*. New York, NY, USA: Springer, 2012.
[34] I. Kontoyiannis, L. A. Lastras-Montaño, and S. P. Meyn, "Relative entropy and exponential deviation bounds for general Markov chains," in *Proc. Int. Symp. Inf. Theory*, 2005, pp. 1563–1567.
[35] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. London, U.K.: Oxford Univ. Press, 2013.

P. N. Karthik (Member, IEEE) received the Bachelor of Engineering degree in electronics and communications from the Rashtreeya Vidyalaya College of Engineering, Bengaluru, and the dual Master of Science (Engineering) and Ph.D. degrees from the Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru, in 2022. He is currently a Research Fellow at the Institute of Data Science, National University of Singapore. His research lies broadly at the intersection of information theory and statistical learning theory, and includes topics such as Markov decision problems, multi-armed bandits, stochastic adaptive control, and federated learning.

Kota Srinivas Reddy (Member, IEEE) received the M.Tech. degree from the Department of Electronics and Electrical Communication Engineering (E&ECE), Indian Institute of Technology Kharagpur (IIT Kharagpur), in 2014, and the Ph.D. degree from the Department of Electrical Engineering (EE), Indian Institute of Technology Bombay (IIT Bombay), in 2021. He is currently a Research Fellow at the Department of Electrical and Computer Engineering (ECE), National University of Singapore (NUS). His research interests include information theory, statistical inference, error-correcting codes, and networks. He received the IIT Bombay Naik and Rastogi Awards for Excellence in Ph.D. Research in 2021 and the INSPIRE Faculty Fellowship in 2022.

Vincent Y. F. Tan (Senior Member, IEEE) was born in Singapore, in 1981. He received the B.A. and M.Eng. degrees in electrical and information sciences from Cambridge University in 2005 and the Ph.D. degree in electrical engineering and computer science (EECS) from the Massachusetts Institute of Technology (MIT) in 2011.

He is currently an Associate Professor at the Department of Mathematics and the Department of Electrical and Computer Engineering, National University of Singapore (NUS). His research interests include network information theory, machine learning, and statistical signal processing. He is a member of the IEEE Information Theory Society Board of Governors. He received the MIT EECS Jin-Au Kong Outstanding Doctoral Thesis Prize in 2011, the NUS Young Investigator Award in 2014, the Singapore National Research Foundation (NRF) Fellowship (Class of 2018), and the NUS Young Researcher Award in 2019. He is currently serving as a Senior Area Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and an Associate Editor of Machine Learning and Statistics for the IEEE TRANSACTIONS ON INFORMATION THEORY. He was also an IEEE Information Theory Society Distinguished Lecturer for 2018/2019.