Sequential Controlled Sensing to Detect an Anomalous Process

A Thesis

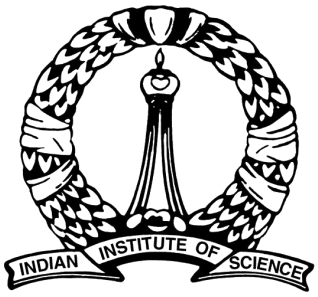
Submitted for the Degree of

Doctor of Philosophy

in the **Faculty of Engineering**

by

Karthik Periyapattana Narayanaprasad



Electrical Communication Engineering
Indian Institute of Science
Bangalore – 560 012 (INDIA)

November, 2021

To my parents.

Acknowledgements

If there are two people to whom the beginning of this journey of PhD can be traced back to, they are Prof. Chandra R. Murthy of the Department of ECE, IISc, and Dr. K. G. Nagananda, a close colleague and a doting mentor of mine. While Prof. Chandra introduced me to the world of academic research and sowed in me the seeds that would later germinate and embrace the vast ocean that research is, Nanda taught me the art of searching for pearls hidden deep beneath this ocean. While one taught me how to fly long distances looking for a flower, the other taught me the art of sucking the ambrosia from the flower without robbing the flower of its form and fragrance. If not for these noble souls, I would not have had the courage to pursue PhD. I am immensely thankful to both of them.

In my search for pearls in the ocean of research, I found one in my adviser, Prof. Rajesh Sundaresan. From heartily welcoming me into his lab space to making me feel at home, from patiently making me understand my paltry background in mathematics to guiding me gracefully towards speaking fluently in a language that he naturally spoke, from giving me the space and time to do things at my own pace to encouraging me to do things in my own way, from being an astute adviser who recognises the many infirmities of his students to thinking many steps ahead towards turning them into strengths, Prof. Rajesh has been with me through the thick and thin of my PhD journey. In his absence, he has reminded me that he is the sturdy earth beneath my feet along this journey, always present and ever ready to take on weight. For all the important lessons in research and in life that Prof. Rajesh has taught me, I shall remain forever indebted to him. I hope to brighten the lives of many students in much the same way that Prof. Rajesh did of mine, perhaps the least I can do to carry forward his legacy.

I am thankful to Prof. Utpal Mukherji, Prof. Navin Kashyap, Prof. Himanshu Tyagi, and Prof. Parimal Parag for serving as lighthouses along the journey. Their words of encouragement during every step of my PhD have been invaluable. From providing me with multiple opportunities to work alongside them in various capacities, be as a teaching assistant or otherwise, to readily agreeing to serve as my professional referees, they have been a true source of support. In particular, I would like to thank Prof. Navin and Prof. Himanshu for serving on the annual

progress review committee and providing a platform for me to share with them the difficulties I faced along the journey without the fear of being judged.

I am immensely thankful to all the professors of the Department of ECE and the Department of Mathematics at IISc for the superior quality lectures that I was fortunate to listen to. I am also thankful to the professors and students of TIFR Mumbai, IIT Bombay, IIT Madras, ISI Bangalore, ISI Delhi, and ISI Kolkata for the opportunities given to me to attend many of their seminars, workshops and lectures, and present my work at these places on several occasions.

An integral part of my PhD journey, one that will stay the closest to my heart, is the people with whom I associated myself while at IISc. Kishan, Chatan, Nihesh, Sarath, Krishna, Prakash, Hemanth, Akhil, Thiru, Nidhin, Bharath, Surabhi, Vinay, Garima, Chinmaya, Hari, Lakshmi Priya, Prathamesh, Sahasranand, Lekshmi, Bala, Sarvendranath, Vaishali, Rooji, Tarun – these names will stay etched in my heart for the rest of my life. There are many more names that I have not mentioned here, courtesy paucity of space.

This work was supported in part by the Science and Engineering Research Board (SERB) of the Department of Science and Technology, in part by the Ministry of Human Resource Development, Government of India, in part by the Centre for Networked Intelligence (CNI), IISc, and in part by the Robert Bosch Center for Cyber-Physical Systems (RBCCPS), IISc. I am thankful to RBCCPS for providing me the monetary support to travel to France to present my work at the 2019 IEEE International Symposium on Information Theory (ISIT). I would like to thank Lalitha Bai and Mr. Srinivasamurthy for extending support to me in matters concerning travel and disbursement of salary on a regular basis.

This journey could not have been completed without the blessings of my parents and my gurus. I am thankful to Providence for gifting me a loving sister who cared for our parents in my absence. I am indebted to the many relatives whose love, care and words of encouragement kept me going along the journey. I only regret not being able to give them a satisfactory reply every time they asked me how much longer it would take me to finish my PhD studies. I am glad that they will not have to ask me anymore!

Abstract

In this thesis, we study the problem of identifying an anomalous arm in a multi-armed bandit as quickly as possible, subject to an upper bound on the error probability. Also known as odd arm identification, this problem falls within the class of optimal stopping problems in decision theory and can be embedded within the framework of active sequential hypothesis testing. Prior works on odd arm identification dealt with independent and identically distributed observations from each arm. We provide the first known extension to the case of Markov observations from each arm. Our analysis and results are in the asymptotic regime of vanishing error probability.

We associate with each arm an ergodic discrete time Markov process that evolves on a common, finite state space. The notion of anomaly is that the transition probability matrix (TPM) of the Markov process of one of the arms (the *odd arm*) is some P_1, and that of each non-odd arm is a different P_2, $P_2 \neq P_1$. A learning agent whose goal it is to find the odd arm samples the arms sequentially, one at any given time, and observes the state of the Markov process of the sampled arm. The Markov processes of the unobserved arms may either remain frozen at their last observed states until their next sampling instant (*rested arms*) or continue to undergo state evolution (*restless arms*). The TPMs P_1 and P_2 may be known to the learning agent beforehand or unknown. We analyse the following cases: (a) rested arms with TPMs unknown, (b) restless arms with TPMs known, and (c) restless arms with TPMs unknown. For each of these cases, we first present an asymptotic lower bound on the growth rate of the expected time required to find the odd arm, and subsequently devise a sequential arm selection policy which we show achieves the lower bound and is therefore asymptotically optimal.

A key ingredient in our analysis of the setting of rested arms is the observation that for the Markov process of each arm, the long term fraction of entries into a state is equal to the long term fraction of exits from the state (global balance). When the arms are restless, it is necessary for the learning agent to keep track of the time since each arm was last sampled (arm's *delay*) and the state of each arm when it was last sampled (arm's *last observed state*). We show that the arm delays and the last observed states form a controlled Markov process which is ergodic under any stationary arm selection policy that picks each arm with a strictly positive

probability. Our approach of considering the delays and the last observed states of all the arms jointly offers a global perspective of the arms and serves as a 'lift' from the local perspective of dealing with the delay and the last observed state of each arm separately, one that is suggested by the prior works. Lastly, when the TPMs are unknown and have to be estimated along the way, it is important to ensure that the estimates converge almost surely to their true values asymptotically, i.e., the system is *identifiable*. We show identifiability follows from the ergodic theorem in the rested case, and provide sufficient conditions for it in the restless case.

Contents

List of Figures

Publications Based on this Thesis

1. P. N. Karthik, R. Sundaresan, *Learning to Detect an Odd Restless Markov Arm with a Trembling Hand*, submitted.

2. P. N. Karthik, R. Sundaresan, *Detecting an Odd Restless Markov Arm with a Trembling Hand*, IEEE Transactions on Information Theory, July 2021.

3. P. N. Karthik, R. Sundaresan, *Learning to Detect an Odd Markov Arm*, IEEE Transactions on Information Theory, July 2020, vol. 66, no. 7, pp. 4324 – 4348.

4. P. N. Karthik, R. Sundaresan, *Learning to Detect an Odd Restless Markov Arm*, proceedings of the 2021 IEEE International Symposium on Information Theory (ISIT), virtual conference.

5. P. N. Karthik, R. Sundaresan, *Detecting an Odd Restless Markov Arm with a Trembling Hand*, proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT), virtual conference.

6. P. N. Karthik, R. Sundaresan, *Learning to Detect an Odd Markov Arm*, proceedings of the 2019 IEEE International Symposium on Information Theory (ISIT), Paris, France, July 2019.

Chapter 1

Introduction

Suppose that a decision entity wishes to select a subset of one or more alternatives among a fixed number of alternatives, each of which yields a random reward from an unknown distribution upon being selected. In the face of uncertainty, the decision entity must select the alternatives sequentially in a way so as to maximise its overall reward accumulated over a finite time horizon of T time instants or an infinite time horizon. Such examples of decision making under uncertainty are popularly modeled using *multi-armed bandits* in which an arm is analogous to an altenative. Defining *regret* of an arms selection strategy (policy) as the difference between the total expected reward obtained under the policy and the total reward obtained under a policy which knows the reward distributions, the goal of the decision entity is, equivalently, to repeatedly sample the arms so as to minimise the finite or the infinite time horizon regret. Typically, the successive rewards from any given arm are assumed either to be independent and identically distributed (iid) or to have a first order dependence (Markov), and independent of the rewards of the other arms.

While one line of works on multi-armed bandits is based on the theme of maximising rewards (or equivalently minimising regret) as described above, another line of works focuses on the theme of *optimal stopping* where the goal is to test for the validity of one or more hypotheses as quickly as possible and stop further sampling of the arms. For instance, the problem of best arm identification [1] in which the goal is to find the index of the arm that yields the largest mean reward as quickly as possible and stop, is an example of an optimal stopping problem in multi-armed bandits. Another instance of an optimal stopping problem in multi-armed bandits, one that constitutes the main topic of this thesis, is the problem of *odd arm identification* in which one of the arms in the multi-armed bandit is anomalous, and the goal is to find the index of the anomalous arm (or *odd arm*) as quickly as possible. It is interesting to note here that the arm sampling policies that are optimal in the context of reward maximisation (or regret

minimisation) may not necessarily be so for optimal stopping problems; we refer the reader to [2] for a discussion on this.

In this thesis, we study of the problem of odd arm identification in the setting when the successive observations from each arm form a Markov process. In this context, anomaly simply means that the transition probability matrix (TPM) of the odd arm is P_1, and that of each non-odd arm is P_2, where $P_2 \neq P_1$. Throughout the thesis, we study the close interplay between the following important quantities that dictate the performance of any arm sampling policy: (a) the time to stoppage, i.e., the time to identify the index of the odd arm, and (b) the error probability at stoppage. For the purpose of carrying out analysis, it is customary to fix one of these quantities and study the behaviour of the other in terms of the quantity fixed. In this thesis, we fix the error probability and characterise the asymptotic behaviour of the time to stoppage as a function of the error probability, where the asymptotics is as the error probability vanishes.

1.1 Motivation

Our motivation to study the problem of odd arm identification in the setting of Markov observations from the arms comes from the desire to extend, to more general settings, the decision theoretic formulation of a certain visual search experiment conducted by Sripati and Olson [3] and analysed in Vaidhiyan et al. [4, 5]. In this experiment, human subjects were shown a number of images at once, with one *oddball* image in a sea of *distracter* images. The goal of the experiment was to understand the relationship between (a) the average time taken by the human subject to identify the oddball image, and (b) the dissimilarity between the oddball and distracter images as perceived by the human subject. Vaidhiyan et al. modelled visual search for locating an oddball image in a sea of distracter images, as quickly as possible, as an odd arm identification problem with Poisson observations. The Poisson observations stemmed from the Poisson point process model for the neuron firings when the human subject focuses on a particular image, the analogue of pulling an arm. They showed that dissimilarity in neural responses to the oddball and the distracter images predicted the time taken by human subjects in detecting the location of the oddball image. The analysis was extended to the case when the parameters of the process were unknown, but had to be learnt during search, in [6].

The oddball and distracter images in the experiments analysed in Vaidhiyan et al. [4, 5, 6] and in Sripati and Olson [3] were static images. Similar experiments, but with dynamic drifting-dots images (movies), as in Krueger et al. [7], were conducted by Vaidhiyan et al. to see how evidence is accumulated in slow perceptual decision making. In these experiments, the dots executed Brownian motions with a fixed drift at each location. Moreover, the drifts were

identical in the distracter locations and were different from the drift in the oddball location. Subjects had to identify the oddball (anomalous) movie location as quickly as possible. A proper analysis of this visual search, along the lines of [4], [5] and [6], requires an understanding of the so-called *restless* odd Markov arm problem where[1] each arm yields Markov observations, one of the arms is anomalous, and the states of the unobserved arms continue to change (restless arms). Indeed, in the aforementioned drifting-dots experiment, the positions of the dots (state) will have changed when the subject returns to observe a particular location after a decision to look at another location. An analysis of the restless odd Markov arm problem forms the main subject of this thesis.

1.2 Prior works on Odd Arm Identification

The problem of odd arm identification has been studied in the literature for the case of independent and identically distributed (iid) observations from each arm. The works of Vaidhiyan et al. [4, 5, 6] study the case of iid, indeed Poisson, observations from each arm. Prabhu et al. [8] extend the analysis of Vaidhiyan et al. to the case of iid observations belonging to a generic exponential family. The works [4], [5], [6] and [8] can be embedded within the classical works of Chernoff [9] and Albert [10], and provide a general framework for the analysis of lower bounds on expected number of samples required for identifying the index of the odd arm. In addition, they also provide explicit policies that achieve these lower bounds in the asymptotic regime as error probability vanishes. We refer the reader to also [1, 11, 12, 13, 14, 15, 16, 17] for other related works on iid observations. While the aforementioned works deal with iid arms, the novelty in this paper is that we consider Markov arms. To the best of our knowledge, we believe that our work is the first to consider Markov arms in the context of odd arm identification.

1.3 Problem Setup and Objective

Formally, our problem setup is as follows. We consider a multi-armed bandit with $K \geq 3$[2] arms. We associate with each arm a time homogeneous and ergodic discrete time Markov process evolving on a common, finite state space. We assume that the TPM of the odd arm is P_1 and that of each non-odd arm is $P_2 \neq P_1$. Given an error probability threshold $\epsilon > 0$, a learning agent wishes to identify the index of the odd arm as quickly as possible, subject to its error probability not exceeding ϵ. The learning agent may or may not possess the knowledge

[1] The correspondence between the oddball movie experiment and the restless odd Markov arm problem is as follows: an arm corresponds to a movie; an observation from an arm corresponds to the positions of the dots in a movie frame; the Markov relationship between the successive observations from an arm corresponds to the dependence between the positions of the dots in the successive frames of a movie.

[2] This ensures that the notion of an "odd" arm makes sense.

of the TPMs of the arms, and samples the arms sequentially, one at any given time, until it is sufficiently confident of which arm is the odd arm. While the learning agent observes only one arm at any given time, the unobserved arms continue to evolve (restless arms).

Suppose π is an arm selection *policy* of the learning agent whose time to find the odd arm is $\tau(\pi)$. Given $\epsilon > 0$, let $\Pi(\epsilon)$ denote the collection of all policies whose error probability is at most ϵ. Let $C = (h, P_1, P_2)$ denote an arms configuration in which h is the index of the odd arm, P_1 is the TPM of arm h and P_2 is the TPM of each non-odd arm $a \neq h$. Writing $E^{\pi}[\cdot|C]$ to denote the expectation computed under the policy π and under the arms configuration C, an examination of the prior works reveals that

$$\inf_{\pi \in \Pi(\epsilon)} E^{\pi}[\tau(\pi)|C] = O\left(\log \frac{1}{\epsilon}\right);$$

the constant multiplying $\log(1/\epsilon)$ in the above equation is, in general, a function of the arms configuration C. Our objective is to characterise

$$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^{\pi}[\tau(\pi)|C]}{\log(1/\epsilon)}. \tag{1.1}$$

That is, our interest is to capture mathematically the asymptotic growth rate of the expected time required to find the odd arm subject to an upper bound on the error probability, where the asymptotics is as the error probability vanishes.

1.4 Our Contributions

Although our end goal is to study the restless odd Markov arm problem, the continued evolution of the arms in the restless setting presents many challenges in the analysis. Therefore, as a key first step towards an understanding of the restless setting, we analyse the simpler setting of *rested* arms in which the unobserved arms do not evolve and remain frozen at their previously observed states. For the settings of rested arms and restless arms, we provide answers to (1.1) by first deriving a problem-instance dependent (arms configuration dependent) lower bound for (1.1). We show that the lower bound is of the form $1/\alpha(h, P_1, P_2)$, where the constant $\alpha(h, P_1, P_2)$ captures the hardness of the problem: the closer the TPMs P_1 and P_2 are (in an appropriately defined sense), the smaller the value of α is, and therefore the larger the lower bound, thereby implying longer times to find the odd arm. The exact form of $\alpha(h, P_1, P_2)$ depends on whether or not the TPMs P_1 and P_2 are known to the learning agent beforehand. Complementing the lower bound results in each of the above cases (rested/restless, TPMs unknown/known), we devise a sequential arm selection policy and demonstrate that the

expected time for the policy to find the odd arm satisfies an asymptotic upper bound that is equal to $1/\alpha(h, P_1, P_2)$. Thus, our answer to (1.1) is $1/\alpha(h, P_1, P_2)$, where $\alpha(h, P_1, P_2)$ depends on whether the arms are rested/restless, whether the TPMs P_1 and P_2 are known or unknown, etc. From the symmetry of the problem, it can be deduced that $\alpha(h, P_1, P_2)$ does not depend on h, the index of the odd arm.

1.5 Organisation of the Thesis

This thesis is organised as follows. In Chapter 2, we analyse the setting of rested arms when the TPMs P_1 and P_2 are not known beforehand[1]. A key component in our analysis of the lower and the upper bounds for this case is the identification of the fact that for the Markov process of each arm, the long term fraction of entries into a state is equal to the long term fraction of exits from the state (global balance). Both these quantities are in turn equal to the probability of observing the state under the arm's stationary distribution.

In Chapter 3, we derive the results for the setting of restless arms assuming that the TPMs P_1 and P_2 are known beforehand. The restless nature of the arms makes it necessary for the learning agent to keep track of the time since each arm was last sampled (arm's delay) and the state of each arm when it was last sampled (arm's last observed state). We introduce the notion of "trembling hand" commonly observed in visual search experiments, in which at any given time the learning agent intends to sample a certain arm but the actual arm sampled differs from the intended arm with a small probability. We show that the arm delays and the last observed states form a controlled Markov process which, under the trembling hand model, is ergodic under any stationary arm selection policy. Our approach of considering the delays and the last observed states of all the arms jointly offers a global perspective of the arms and serves as a 'lift' from the local perspective of dealing with the delay and the last observed state of each arm separately (as done in the prior works). In the absence of the trembling hand, we discuss the difficulties associated with showing that the lower and the upper bounds match.

In Chapter 4, we extend the results of Chapter 3 to the case when the TPMs P_1 and P_2 are not known beforehand and must be learnt along the way. Here, we estimate the TPMs using maximum likelihood (ML) estimation. The key challenge here is in showing that the ML estimates of the TPMs converge to their true values, i.e., the system is *identifiable*. We prove identifiability under a continuous selection assumption and a certain regularity assumption on the TPMs. Our achievability (upper bound) analysis relies crucially on resolving the identifiability problem. For a simpler policy that estimates the TPMs by repeated sampling of each

[1]The case of known TPMs in the setting of rested arms is a straighforward extension of the case of iid observations from the arms, and is therefore omitted.

arm before switching to another arm, we highlight the difficulties in achieving the lower bound and meeting the desired error probability.

We conclude the thesis in Chapter 5 and propose some future directions for exploration.

Chapter 2

Rested Arms

2.1 Preamble

In this chapter, we analyse the setting of rested arms with TPMs unknown. That is, given a multi-armed bandit with $K \geq 3$ arms and an arms configuration $C = (h, P_1, P_2)$ in which h is the index of the odd arm, the TPM of arm h is P_1, and the TPM of each of the remaining arms is $P_2 \neq P_1$, we wish to characterise

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

for the setting of rested arms when P_1 and P_2 are not known beforehand. Our analysis of the setting of rested arms forms a key first step towards analysing the more difficult setting of restless arms. The rested case has its own interesting features. For example, as we shall see later in this chapter, the asymptotically optimal arm selection strategy does not explicitly depend on the last observed states of the arms. This, at first glance, is surprising.

2.1.1 Prior Works on Rested Arms

One of the earliest works to consider the setting of rested arms is that of Gittins' [18] in which it is assumed that each arm yields a random 'reward' when selected, and that the successive rewards from any given arm constitute a Markov process. In this reward setting, the central problem is one of maximising the infinite horizon average discounted reward. For this problem, Gittins proposed and demonstrated the optimality of a simple index-based policy that, at each time, involves constructing an index for every arm based on the knowledge of the transition laws of the arms and selecting an arm with the largest index. Agarwal et al. [19] consider a similar setting as Gittins' in which each arm yields Markov rewards. However, unlike in [18],

the authors of [19] do not assume the knowledge of the transition laws of the arms. Instead, they assume that the transition law of each arm is parameterised by an unknown parameter belonging to a known, finite, parameter space. Define 'regret' as the difference between the infinite time horizon expected sum of rewards generated by any policy and that generated by a policy which knows the parameters of the arms. The goal of the authors of [19] is to design policies whose regret, in the asymptotic limit as time $n \to \infty$, is $o(n^\alpha)$ for every $\alpha > 0$. For this problem, the authors of [19] provide a lower bound in which the long-term regret grows asymptotically as $\log n$ times a multiplicative constant that captures the hardness of the problem. Furthermore, they propose a policy and demonstrate that it achieves the lower bound in the limit as $n \to \infty$.

While the aforementioned works deal with reward maximisation, and the associated regret minimisation in the unknown parameters setting, our problem is one of *optimal stopping*. We note that the lower bound in [19], although quantifying the asymptotic growth rate of regret, does not reflect the quickness of learning the underlying parameters of the arms. That is, the results in [19] do not shed light on the minimum number of arm selections that are needed, on the average, in order to learn the parameters of the arms up to a desired level of accuracy. We answer this question when one of the arms is anomalous and the asymptotics is as the error probability vanishes. In doing so, we treat the state of any selected arm as merely a Markov observation from the arm and not as a reward, since our objective is one of optimal stopping and not of regret minimisation. We note here that policies which are optimal in the context of the problem studied here may not necessarily be optimal in the context of regret minimisation, and vice-versa; see Bubeck et al. [2] for a discussion on this. Finally, the unknown parameters of our problem are the transition laws of the odd arm and the non-odd arm Markov processes, and the index of the odd arm, thus making our parameter set a continuum, unlike the finite parameter set considered in [19].

Finally, a recent and independent work of Moulos [20] studies a closely related problem of *best arm identification* in rested multi-armed bandits, where the goal is to find the arm with the largest stationary mean. The results presented in [20] are in terms of an asymptotic and a non-asymptotic lower bound, where the asymptotics is as the error probability vanishes, and a policy for best arm identification whose asymptotic upper bound is four times larger than the asymptotic lower bound. In this chapter, we present the first known asymptotic lower bound for the problem of odd arm identification, and an asymptotically optimal policy that meets the lower bound. This is in contrast with the gap between the upper bound and the lower bound in [20] for best arm identification. We anticipate that a policy similar to ours should close the gap between the upper and lower bounds in [20].

2.1.2 An Overview of Our Contributions

Below, we highlight our contributions and bring out the challenges that we need to overcome in the analysis of the setting of rested arms.

1. In Section 2.3, we derive an asymptotic lower bound on the growth rate of the expected number of arm selections required by any policy that the learning agent may use to identify the index of the odd arm. Here, the asymptotics is as the error probability vanishes. Similar to the lower bounds appearing in [8] and [19], our lower bound has a problem-instance (or arms configuration) dependent constant that quantifies the effort required by any policy to identify the true index of the odd arm by guarding itself against identifying the nearest, incorrect alternative. This constant is a function of the transition probability matrices of the odd arm and the non-odd arms.

2. We characterise the growth rate of the expected number of arm selections required by any policy as a function of the maximum acceptable error probability, and show that in the regime of vanishingly small error probabilitys, this growth rate is logarithmic in the inverse of the error probability. The analysis of the lower bounds in [6] and [8] uses the familiar data processing inequality presented in the work of Kaufmann et al. [1] that is based on Wald's identity [21] for iid processes. However, the Markov setting in our problem does not permit the use of Wald's identity. Therefore, we derive results for our Markov setting generalising those appearing in [1], and subsequently use these generalisations to arrive at the lower bound. See Section 2.3 for the details.

3. In the analysis of the lower bound, we bring out the key idea that any two successive selections of an arm result in the learning agent observing a transition from the state corresponding to the arm's first selection to the state corresponding to the arm's second selection. As a consequence, for each state in the state space, the empirical proportion of times an arm occupies the state prior to a transition is equal, in the long run, to the empirical proportion of times the arm occupies the state after a transition. We then replace these common proportions by the probability of the arm occupying this state under its stationary distribution. Such a replacement by stationary probabilities is possible mainly due to the rested nature of the arms, and may not be possible in more general settings such as when the arms are restless.

4. In Section 2.4, we propose a sequential arm selection scheme that takes as inputs two parameters, one of which may be chosen appropriately to meet the acceptable error prob-

ability, while the other may be tuned to ensure that the performance of our scheme comes arbitrarily close to the lower bound, thus making our scheme near-optimal.

We now contrast the near-optimality of our scheme with the near-optimality of the scheme proposed by Vaidhiyan et al. in [6], and highlight a key simplification that our scheme entails. The scheme of Vaidhiyan et al. is built around the important fact that each arm is sampled at a non-trivial, strictly positive and optimal rate that is bounded away from zero, as given by the lower bound, thereby allowing for exploration of the arms in an optimal manner. This stemmed from their specific Poisson observations. However, the lower bound presented in Section 2.3 may not have this property in the context of Markov observations. Therefore, recognising the requirement of sampling the arms at a non-trivial rate for good performance of our scheme, in this chapter, we use the idea of "forced exploration" proposed by Albert in [10]. In particular, we propose a simplified way of sampling the arms by considering a mixture of uniform sampling and the optimal sampling given by the lower bound in Section 2.3. We do this by introducing an appropriately tuneable parameter that controls the probability of switching between uniform sampling and optimal sampling, the latter being given by the lower bound. While this ensures that our policy samples each arm with a strictly positive probability at each step, it also gives us the flexibility to select an appropriate value for this parameter so that the upper bound on the performance of our scheme may be made arbitrarily close to our lower bound. We refer the reader to Section 2.4 for the details.

2.1.3 Chapter Organisation

The rest of this chapter is organised as follows. In Section 2.2, we set up the notations that will be used throughout the chapter. In Section 2.3, we present a lower bound on the performance of any policy. In Section 2.4, we present a sequential arm selection policy and demonstrate its near optimality. We present the main result of this chapter in Section 2.5, combining the results of Sections 2.3 and 2.4. In Section 2.6, we provide some simulation to support the theoretical development. We present the proofs of all the results in Section 2.7, and summarise the key points of this chapter in Section 2.8.

2.2 Notations

In this section, we set up the notations that will be used throughout the chapter. Let $K \geq 3$ denote the number of arms, and let $\mathcal{A} = \{1, 2, \ldots, K\}$ denote the set of arms. We associate with each arm an irreducible, aperiodic, time homogeneous discrete-time Markov process on a finite state space $\mathcal{S}$, where the Markov process of each arm is independent of the Markov

processes of the other arms. We denote by $|\mathcal{S}|$ the cardinality of $\mathcal{S}$. Without loss of generality, we take $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$. Hereinafter, we use the phrase 'Markov process of arm a' to refer to the Markov process associated with arm $a \in \mathcal{A}$.

At each discrete time instant, one out of the K arms is selected and its state is observed. We let A_n denote the arm selected at time n, and let $\bar{X}_n$ denote the state of arm A_n, where $n \in \{0, 1, 2, \ldots\}$. We treat A_0 as the zeroth arm selection and $\bar{X}_0$ as the zeroth observation. Selection of an arm at time n is based on the history $(\bar{X}^{n-1}, A^{n-1})$ of past observations and arms selected; here, $\bar{X}^k$ (resp. A^k) is a shorthand notation for the sequence $\bar{X}_0, \ldots, \bar{X}_k$ (resp. $A_0, \ldots, A_k$). We shall refer to such a sequence of arm selections and observations as a policy, which we generically denote by π. For each $a \in \mathcal{A}$, we denote the Markov process of arm a by the collection $(X_k^a)_{k \geq 0}$ of random variables. Further, we denote by $N_a(n)$ the number of times arm a is selected by a policy up to (and including) time n, i.e.,

$$N_a(n) = \sum_{t=0}^{n} 1_{\{A_t = a\}}. \tag{2.1}$$

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

$$\bar{X}_n = X^{A_n}_{N_{A_n}(n)-1}. \tag{2.2}$$

We consider a scenario in which the Markov process of one of the arms (hereinafter referred to as the odd arm) follows a probability transition matrix $P_1 = (P_1(j|i))_{i,j \in \mathcal{S}}$, while those of rest of the arms follow a probability transition matrix $P_2 = (P_2(j|i))_{i,j \in \mathcal{S}}$, where $P_2 \neq P_1$; here, $P(j|i)$ denotes the entry in the ith row and jth column of the matrix P. Further, we let μ_1 and μ_2 denote the unique stationary distributions of P_1 and P_2 respectively. We denote by ν the common distribution for the initial state of each Markov process. In other words, for arm $a \in \mathcal{A}$, we have $X_0^a \sim \nu$, and this is the same distribution for all arms. We operate in a setting where the probability transition matrices and their associated stationary distributions are unknown to the learning agent.

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

$$N_a(n, i) = \sum_{m=1}^{N_a(n)-1} 1_{\{X^a_{m-1} = i\}}. \tag{2.3}$$

Similarly, for each $i, j \in \mathcal{S}$, we denote by $N_a(n, i, j)$ the number of times up to (and including)

time n the Markov process of arm a is observed to make a transition from state i to state j, i.e.,

$$N_a(n,i,j) = \sum_{m=1}^{N_a(n)-1} 1_{\{X^a_{m-1}=i,\, X^a_m=j\}}. \tag{2.4}$$

Clearly, then, the following hold:

1. For each $a \in \mathcal{A}$ and $i \in \mathcal{S}$,
$$\sum_{j\in\mathcal{S}} N_a(n,i,j) = N_a(n,i). \tag{2.5a}$$
2. For each $a \in \mathcal{A}$,
$$\sum_{i\in\mathcal{S}} N_a(n,i) = N_a(n) - 1. \tag{2.5b}$$
3. For each n,
$$\sum_{a\in\mathcal{A}} N_a(n) = n + 1. \tag{2.5c}$$

We note here that the upper index of the summation in (2.3) is $N_a(n)-1$, and not $N_a(n)$, since the last observed transition on arm a would be from the state $X^a_{N_a(n)-2}$ to the state $X^a_{N_a(n)-1}$. This is further reflected by the summation in (2.5b).

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

$$\mathcal{C} = \{C = (h, P_1, P_2) : h \in \mathcal{A},\ P_1 \text{ and } P_2 \text{ are transition probability matrices on } \mathcal{S},\ P_2 \neq P_1\}$$

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

$$Z^a_h(n) = \begin{cases} 0, & N_a(n) = 0, \\ \log \nu(X^a_0), & N_a(n) = 1, \\ \log \nu(X^a_0) + \sum_{m=1}^{N_a(n)-1} \log P^a_h(X^a_m | X^a_{m-1}), & N_a(n) \geq 2, \end{cases} \tag{2.6}$$

where $P^a_h(j|i)$ denotes the conditional probability under hypothesis H_h of observing state j on

arm a given that state i was observed on arm a at the previous sampling instant, and is given by

$$P_h^a(j|i) = \begin{cases} P_1(j|i), & a = h, \\ P_2(j|i), & a \neq h. \end{cases} \tag{2.7}$$

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

$$Z_h(n) = \sum_{a=1}^{K} Z_h^a(n), \tag{2.8}$$

where $Z_h^a(n)$ is as given in (2.6). On similar lines, for any two configurations $C = (h, P_1, P_2)$ and $C' = (h', P_1', P_2')$, where $P_2' \neq P_1'$ and $h' \neq h$, for each $a \in \mathcal{A}$, we define the log-likelihood process $(Z_{hh'}^a(n))_{n\geq 0}$ of configuration C with respect to configuration C' for arm a as

$$\begin{aligned} Z_{hh'}^a(n) &= Z_h^a(n) - Z_{h'}^a(n) \\ &= \begin{cases} 0, & N_a(n) = 0, 1, \\ \sum\limits_{m=1}^{N_a(n)-1} \log \dfrac{P_h^a(X_m^a|X_{m-1}^a)}{P_{h'}^a(X_m^a|X_{m-1}^a)}, & N_a(n) \geq 2. \end{cases} \end{aligned} \tag{2.9}$$

We note that in the above equation, for P_h^a, we should use (2.7), and for $P_{h'}^a$, we shall use, for all $a \in \mathcal{A}$ and $i, j \in \mathcal{S}$,

$$P_{h'}^a(j|i) = \begin{cases} P_1'(j|i), & a = h', \\ P_2'(j|i), & a \neq h'. \end{cases} \tag{2.10}$$

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

$$Z_{hh'}(n) = \sum_{a=1}^{K} Z_{hh'}^a(n), \tag{2.11}$$

which includes all arm selections and observations.

The observation process $(\bar{X}_n)_{n\geq 0}$ and the arm selection process $(A_n)_{n\geq 0}$ are assumed to be defined on a common probability space $(\Omega, \mathcal{F}, P)$. We define the filtration $(\mathcal{F}_n)_{n\geq 0}$ as

$$\mathcal{F}_n = \sigma(A^n, \bar{X}^n), \quad n \geq 0. \tag{2.12}$$

We use the convention that the zeroth arm selection A_0 is measurable with respect to the sigma algebra $\{\phi, \Omega\}$, whereas for all $n \geq 1$, the nth arm selection A_n is $\mathcal{F}_{n-1}$-measurable. For any stopping time τ with respect to the filtration in (2.12), we denote by $\mathcal{F}_\tau$ the σ-algebra

$$\mathcal{F}_\tau = \{E \in \mathcal{F} : E \cap \{\tau = n\} \in \mathcal{F}_n \text{ for all } n \geq 0\}. \tag{2.13}$$

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

1. Select an arm, based on the history of past observations and arms selected, according to a fixed distribution λ independent of the underlying configuration of the arms, i.e., for each $n \geq 1$,
$$P(A_n = a | A^{n-1}, \bar{X}^{n-1}) = \lambda(a). \tag{2.14}$$
2. Stop selecting arms, and declare the index $I(\pi)$ as the odd arm.

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

$$\Pi(\epsilon) = \left\{ \pi : P^\pi(I(\pi) \neq h | C) \leq \epsilon \ \forall \ C = (h, P_1, P_2), \text{ where } h \in \mathcal{A} \text{ and } P_2 \neq P_1 \right\}. \tag{2.15}$$

For a policy π, we denote its stopping time by $\tau(\pi)$. Further, we write $E^\pi[\cdot|C]$ and $P^\pi(\cdot|C)$ to denote expectations and probabilities given that the underlying configuration of the arms is C. In this chapter, we characterise the behaviour of $E^\pi[\tau(\pi)|C]$ for any policy $\pi \in \Pi(\epsilon)$, as ϵ approaches zero. We re-emphasise that π cannot depend on the knowledge of P_1 or P_2, but could attempt to learn these along the way.

Remark 1. *Fix an odd arm index h, and consider the simpler case when P_1, P_2 are known, $P_2 \neq P_1$. Let $\Pi(\epsilon|P_1, P_2)$ denote the set of all policies whose probability of error at stoppage is within ϵ. From the definition of $\Pi(\epsilon)$ in (2.15), it follows that*

$$\Pi(\epsilon) = \bigcap_{P_1, P_2 : P_2 \neq P_1} \Pi(\epsilon | P_1, P_2). \tag{2.16}$$

That is, policies in $\Pi(\epsilon)$ work for any P_1, P_2, with $P_2 \neq P_1$. It is not a priori clear whether the set $\Pi(\epsilon)$ is nonempty. That it is nonempty for the case of iid observations was established in

[9]. In this chapter, we show that $\Pi(\epsilon)$ *is nonempty even for the setting of rested and Markov arms.*

Remark 2. *The distribution* λ *appearing in* (2.14) *may, in general, be a function of time index* n*.*

In the next section, we provide a configuration dependent lower bound on $E^{\pi}[\tau(\pi)|C]$ for any policy $\pi \in \Pi(\epsilon)$. In Section 2.4, we propose a sequential arm selection policy that achieves the lower bound asymptotically as the probability of error vanishes. We present the proofs in Section 2.7.

2.3 Converse: Lower Bound

For any two transition probability matrices P and Q of dimension $|\mathcal{S}| \times |\mathcal{S}|$, and a probability distribution μ on $\mathcal{S}$, define $D(P\|Q|\mu)$ as

$$D(P\|Q|\mu) := \sum_{i\in\mathcal{S}} \mu(i) \sum_{j\in\mathcal{S}} P(j|i) \log \frac{P(j|i)}{Q(j|i)}, \tag{2.17}$$

with the convention $0 \log 0 = 0 \log \frac{0}{0} = 0$. The quantity in (2.17) is known as *conditional informational divergence*, and the notation used above for representing the same is standard in the literature. See, for instance, Csiszár and Körner [22, Eq. (2.4)]. The following proposition gives an asymptotic lower bound on the growth rate of the expected stopping time of any policy $\pi \in \Pi(\epsilon)$ as $\epsilon \downarrow 0$.

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

$$\liminf_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{E^{\pi}[\tau(\pi)|C]}{\log(1/\epsilon)} \geq \frac{1}{D^*(h, P_1, P_2)}, \tag{2.18}$$

where $D^*(h, P_1, P_2)$ *is an arms configuration-dependent constant given by*

$$D^*(h, P_1, P_2) = \max_{0\leq\lambda_1\leq 1} \left\{ \lambda_1 \, D(P_1\|P|\mu_1) + (1-\lambda_1)\frac{(K-2)}{(K-1)} D(P_2\|P|\mu_2) \right\}. \tag{2.19}$$

In (2.19)*,* P *is a probability transition matrix whose* (i,j)*th entry is given by*

$$P(j|i) = \frac{\lambda_1\mu_1(i)P_1(j|i) + (1-\lambda_1)\frac{(K-2)}{(K-1)}\mu_2(i)P_2(j|i)}{\lambda_1\mu_1(i) + (1-\lambda_1)\frac{(K-2)}{(K-1)}\mu_2(i)}. \tag{2.20}$$

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

$$d(\epsilon, 1-\epsilon) := \epsilon \log \frac{\epsilon}{1-\epsilon} + (1-\epsilon) \log \frac{1-\epsilon}{\epsilon}. \tag{2.21}$$

Next, we express the expected sum of log-likelihoods up to the stopping time $\tau(\pi)$ in terms of the expected value of the stopping time. It is in obtaining such an expression that works such as [1], [6] and [8] that are based on iid observations use Wald's identity, which greatly simplifies their analysis of the lower bound. The Markov observations of our setting does not permit the use of Wald's identity. Therefore, we first obtain a generalisation of [1, Lemma 18], a change of measure based argument, for the setting of Markov observations, and subsequently use this generalisation to obtain the desired relation.

We then show that for any arm $a \in \mathcal{A}$, the long run frequency of observing the arm occupying state $i \in \mathcal{S}$ prior to a transition is equal to that of arm a occupying the state i after a transition, and note that this common frequency is the stationary probability of observing the arm in the state i. This explains the appearance of the unique stationary distributions μ_1 and μ_2 of the odd arm and the non-odd arms respectively in the expression (2.19). We wish to emphasise that this step in the proof is possible due to the rested nature of the arms.

Finally, combining the above steps and using $d(\epsilon, 1-\epsilon)/\log \frac{1}{\epsilon} \to 1$ as $\epsilon \downarrow 0$, we arrive at the lower bound in (2.18). The details may be found in Section 2.7.1.

Remark 3. *The right-hand side of* (2.19) *is a function only of the transition probability matrices P_1 and P_2, and does not depend on h, the index of the odd arm. This is due to symmetry in the structure of arms, and we deduce that $D^*(h, P_1, P_2)$ does not depend on h. However, we include h for the sake of consistency with the notation $C = (h, P_1, P_2)$ used to denote arm configurations. Further, it reminds us that D^* may depend on all the parameters of the underlying configuration in more general composite hypothesis testing settings.*

Going further, we let $\lambda^* \in [0, 1]$ denote the value of λ that achieves the maximum in (2.19). We then define $\lambda_{opt}(h, P_1, P_2) = (\lambda_{opt}(h, P_1, P_2)(a))_{a \in \mathcal{A}}$ as the probability distribution on $\mathcal{A}$ given by

$$\lambda_{opt}(h, P_1, P_2)(a) := \begin{cases} \lambda^*, & a = h, \\ \frac{1-\lambda^*}{K-1}, & a \neq h. \end{cases} \tag{2.22}$$

In the next section, we construct a policy that, at each time step, chooses arms with probabilities that match with those in (2.22) in the long run, in an attempt to reach the lower bound. While it is not a priori clear that this yields an asymptotically optimal policy, we show that this is indeed the case.

2.4 Achievability

In this section, we propose a scheme that achieves the lower bound of Section 2.3 asymptotically as the probability of error vanishes. Our policy is a modification of the policy proposed by Prabhu et al. [8] for the case of K iid processes. We denote our policy by $\pi^\star(L, \delta)$, where $L > 1$ and $\delta \in (0, 1)$ are the parameters of the policy.

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

2.4.1 The Modified GLR Test Statistic

We revisit (2.8), and suppose that each arm is selected once in the first K time slots. Note that this does not affect the asymptotic performance. Then, under configuration $C = (h, P_1, P_2)$, the log-likelihood process $Z_h(n)$ may be expressed for any $n \geq K$ as

$$Z_h(n) = \sum_{a=1}^{K} \log \nu(X_0^a) + \sum_{i,j \in \mathcal{S}} N_h(n, i, j) \log P_1(j|i) + \sum_{i,j \in \mathcal{S}} \left(\sum_{a \neq h} N_a(n, i, j) \right) \log P_2(j|i), \quad (2.23)$$

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

$$f(A^n, \bar{X}^n | C) = \prod_{a=1}^{K} \nu(X_0^a) \prod_{i,j \in \mathcal{S}} (P_1(j|i))^{N_h(n,i,j)} \cdot \prod_{i,j \in \mathcal{S}} (P_2(j|i))^{\sum_{a \neq h} N_a(n,i,j)}. \quad (2.24)$$

We now introduce an artificial prior on the space of all transition probability matrices for the state space $\mathcal{S}$. Our choice of the prior is motivated by the requirement of having an appropriate conjugate prior for the likelihood in (2.24). We therefore construct the Dirichlet distribution-based prior, noting that it meets our requirement. Let $\mathrm{Dir}(1, \ldots, 1)$ denote the Dirichlet distribution with $|\mathcal{S}|$ parameters $\alpha_1, \ldots, \alpha_{|\mathcal{S}|}$, where $\alpha_j = 1$ for all $j \in \mathcal{S}$. Then, denoting by $\mathscr{P}(\mathcal{S})$ the space of all transition probability matrices of size $|\mathcal{S}| \times |\mathcal{S}|$, we specify a prior on $\mathscr{P}(\mathcal{S})$ using the above Dirichlet distribution as follows: for any $P = (P(j|i))_{i,j \in \mathcal{S}} \in$

$\mathscr{P}(\mathcal{S})$, $P(\cdot|i)$ is chosen according to the above Dirichlet distribution, independently of $P(\cdot|j)$ for all $j \neq i$. Further, for any two matrices $P, Q \in \mathscr{P}(\mathcal{S})$, the rows of P are independent of those of Q. Then, it follows that under this prior, the joint density at $(P_1,\ P_2)$ for $P_1, P_2 \in \mathscr{P}(\mathcal{S})$ is

$$\begin{aligned}\mathscr{D}(P_1, P_2) &:= \prod_{i\in\mathcal{S}} \frac{\prod_{j\in\mathcal{S}} (P_1(j|i))^{\alpha_j - 1}}{B(1\ldots,1)} \prod_{i\in\mathcal{S}} \frac{\prod_{j\in\mathcal{S}} (P_2(j|i))^{\alpha_j - 1}}{B(1\ldots,1)} \\ &= \frac{1}{B(1,\ldots,1)^{2|\mathcal{S}|}},\end{aligned} \tag{2.25}$$

where $B(1,\ldots,1)$ denotes the normalisation factor for the distribution $\mathrm{Dir}(1,\ldots,1)$, and the second line above follows by substituting $\alpha_j = 1$, $j \in \mathcal{S}$.

By a minor abuse of notation, we denote by $f(A^n, \bar{X}^n|H_h)$ the average of the likelihood in (2.24) computed with respect to the prior in (2.25). From the property that the Dirichlet distribution is the appropriate conjugate prior for the observation process,

$$f(A^n, \bar{X}^n|H_h) = \prod_{a=1}^{K} \nu(X_0^a) \prod_{i\in\mathcal{S}} \frac{B((N_h(n,i,j)+1)_{j\in\mathcal{S}})}{B(1,\ldots,1)} \prod_{i\in\mathcal{S}} \frac{B\left(\left(\sum_{a\neq h} N_a(n,i,j) + 1\right)_{j\in\mathcal{S}}\right)}{B(1,\ldots,1)}, \tag{2.26}$$

where in the above expression, $B((N_h(n,i,j)+1)_{j\in\mathcal{S}})$ denotes the normalisation factor for a Dirichlet distribution with parameters $(N_h(n,i,j)+1)_{j\in\mathcal{S}}$. It can be shown that $f(A^n, \bar{X}^n|H_h)$ is also the expected value of the likelihood in (2.24) computed with respect to the prior in (2.25), i.e.,

$$f(A^n, \bar{X}^n|H_h) = \prod_{a=1}^{K} \nu(X_0^a) \prod_{i\in\mathcal{S}} E\left[\prod_{j\in\mathcal{S}} X_{ij}^{N_h(n,i,j)} \cdot Y_{ij}^{\sum_{a\neq h} N_a(n,i,j)}\right] \tag{2.27}$$

where in the above set of equations, the random vectors $(X_{ij})_{i,j\in\mathcal{S}}$ and $(Y_{ij})_{i,j\in\mathcal{S}}$ are independent with independent components, and jointly distributed according to (2.25), and the expectation is also with respect to this joint density.

Let $\hat{P}^n_{h,1}$ and $\hat{P}^n_{h,2}$ denote the maximum likelihood estimates of probability transition matrices P_1 and P_2 respectively under the hypothesis H_h. Taking partial derivatives of the right-hand side of (2.24) with respect to $P_1(j|i)$ and $P_2(j|i)$ for each $i, j \in \mathcal{S}$, and setting each of these

derivatives to zero, we get

$$\hat{P}^n_{h,1}(j|i) = \frac{N_h(n,i,j)}{N_h(n,i)}, \quad \hat{P}^n_{h,2}(j|i) = \frac{\sum\limits_{a\neq h} N_a(n,i,j)}{\sum\limits_{a\neq h} N_a(n,i)}. \tag{2.28}$$

Plugging the estimates in (2.28) back into (2.24), we get the maximum likelihood of all observations and actions under hypothesis H_h:

$$\begin{aligned}\hat{f}(A^n, \bar{X}^n|H_h) &\coloneqq \max_{C=(h,\cdot,\cdot)} f(A^n, \bar{X}^n|C) \\ &= \prod_{a=1}^{K} \nu(X_0^a) \prod_{i,j\in\mathcal{S}} \left\{ \left(\frac{N_h(n,i,j)}{N_h(n,i)}\right)^{N_h(n,i,j)} \left(\frac{\sum\limits_{a\neq h} N_a(n,i,j)}{\sum\limits_{a\neq h} N_a(n,i)}\right)^{\sum\limits_{a\neq h} N_a(n,i,j)} \right\}. \end{aligned} \tag{2.29}$$

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

$$\begin{aligned} M_{hh'}(n) &= \log \frac{f(A^n, \bar{X}^n|H_h)}{\hat{f}(A^n, \bar{X}^n|H_{h'})} \\ &= T_1 + T_2(n) + T_3(n) + T_4(n) + T_5(n), \end{aligned} \tag{2.30}$$

where the terms appearing in (2.30) are as follows.

1. The term T_1 is given by
$$T_1 = 2|\mathcal{S}| \log\left(\frac{1}{B(1,\ldots,1)}\right). \tag{2.31}$$

2. The term $T_2(n)$ is given by
$$T_2(n) = \sum_{i\in\mathcal{S}} \log B((N_h(n,i,j)+1)_{j\in\mathcal{S}}). \tag{2.32}$$

3. The term $T_3(n)$ is given by
$$T_3(n) = \sum_{i\in\mathcal{S}} \log B\left(\left(\sum_{a\neq h} N_a(n,i,j) + 1\right)_{j\in\mathcal{S}}\right). \tag{2.33}$$

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

$$T_4(n) = -\sum_{i,j\in\mathcal{S}} N_{h'}(n,i,j)\log\frac{N_{h'}(n,i,j)}{N_{h'}(n,i)}. \tag{2.34}$$

5. The term $T_5(n)$ is given by

$$T_5(n) = -\sum_{i,j\in\mathcal{S}}\sum_{a\neq h'} N_a(n,i,j)\log\frac{\sum\limits_{a\neq h'} N_a(n,i,j)}{\sum\limits_{a\neq h'} N_a(n,i)}. \tag{2.35}$$

Note that ν, the distribution of the initial state of any arm, is irrelevant since it appears in both (2.26) and (2.29), and thus cancels out in writing (2.30).

Remark 4. *We wish to mention here that the expression on the right-hand side of (2.24) for $f(A^n, \bar{X}^n|C)$ represents the likelihood of all observations up to time n "conditioned on" the actions A^n up to time n. In other words, a more precise expression for $f(A^n, \bar{X}^n|C)$ is*

$$f(A^n,\bar{X}^n|C) = \left[\prod_{t=0}^{n} P_h(A_t|A^{t-1},\bar{X}^{t-1})\right]\prod_{a=1}^{K}\nu(X_0^a)\prod_{i,j\in\mathcal{S}}(P_1(j|i))^{N_h(n,i,j)}\cdot\prod_{i,j\in\mathcal{S}}(P_2(j|i))^{\sum\limits_{a\neq h} N_a(n,i,j)}, \tag{2.36}$$

where $P_h(A_t|A^{t-1},\bar{X}^{t-1})$ represents the probability of selecting the arm A_t at time t when the true hypothesis is H_h (i.e., when h is the index of the odd arm), with the convention that at time $t = 0$, this term represents $P_h(A_0)$. Note that for any policy, this must be independent of the true hypothesis H_h, and is thus the same for any two hypotheses H_h and $H_{h'}$. As a consequence of this, the first term within square brackets on the right-hand side of (2.36) appears in both the numerator and the denominator terms of the modified GLR statistic of (2.30), and thus cancels out. Hence, we omit writing this term in the expressions of (2.24), (2.26) and (2.29).

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

With the above ingredients in place, we now describe our policy based on the modified GLR test statistic of (2.30). Let

$$M_h(n) \coloneqq \min_{h'\neq h} M_{hh'}(n) \tag{2.37}$$

denote the modified GLR of hypothesis H_h, $h \in \mathcal{A}$, with respect to its nearest alternative.

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

Fix parameters $L > 1$ and $\delta \in (0,1)$. Let $(B_n)_{n\geq 1}$ be a sequence of iid Bernoulli(δ) random variables such that B_{n+1} is independent of the sequence $(A^n, \bar{X}^n)$ for all $n \in \{0,1,2,\ldots\}$. Let $A_0 = 1$, $A_1 = 2$ and so on until $A_{K-1} = K$. For $n \geq K-1$, follow the below mentioned steps until stoppage.

1. Compute $\theta(n) \in \arg\max_{h\in\mathcal{A}} M_h(n)$. Resolve ties, if any, uniformly at random.
2. If $M_{\theta(n)}(n) < \log((K-1)L$, then sample the next arm A_{n+1} based on the history $(A^n, \bar{X}^n)$ as per the following rule:
   (a) If $B_{n+1} = 1$, sample A_{n+1} uniformly at random.
   (b) If $B_{n+1} = 0$, sample A_{n+1} according to $\lambda_{opt}(\theta(n), \hat{P}^n_{\theta(n),1}, \hat{P}^n_{\theta(n),2})$.

   Update $n \leftarrow n+1$ and go back to step 1.
3. If $M_{\theta(n)}(n) \geq \log((K-1)L)$, then stop and declare $\theta(n)$ as the odd arm index.

In the above policy, $\theta(n)$ provides the best estimate of the odd arm at time n. If the modified GLR test statistic of arm $\theta(n)$ is sufficiently larger than that of its nearest incorrect alternative ($\geq \log((K-1)L)$), then this indicates that the policy is confident that $\theta(n)$ is the odd arm. At this stage, the policy stops taking further samples and declares $\theta(n)$ as the index of the odd arm. If not, the policy continues to obtain further samples.

We refer to the rule in item (2) above as *forced exploration* with parameter δ. A similar rule also appears in [10]. Based on the description in items 2(a) and 2(b) above, it follows that for each $a \in \mathcal{A}$,

$$\begin{aligned} P(A_{n+1} = a | A^n, \bar{X}^n) &= \frac{\delta}{K} + (1-\delta)\,\lambda_{opt}(\theta(n), \hat{P}^n_{\theta(n),1}, \hat{P}^n_{\theta(n),2})(a) \\ &\geq \frac{\delta}{K} > 0. \end{aligned} \tag{2.38}$$

As we will see, the strictly positive lower bound in (2.38) will ensure that the policy selects each arm at a non-trivial frequency so as to allow for sufficient exploration of all the arms. Also, we will show that the parameters L and δ may be selected so that our policy achieves a desired target error probability, while also ensuring that the normalised expected stopping time of the policy is arbitrarily close to the lower bound in (2.18).

Remark 5. *Evaluating the distribution* $\lambda_{opt}(\theta(n), \hat{P}^n_{\theta(n),1}, \hat{P}^n_{\theta(n),2})$ *in step 2(a) of the policy involves solving the maximisation problem in* (2.19) *with the probability transition matrices* P_1 *and* P_2 *replaced by their corresponding ML estimates* $\hat{P}^n_{\theta(n),1}$ *and* $\hat{P}^n_{\theta(n),2}$ *respectively at each time* $n \geq K-1$ *until stoppage. In the event when any of the rows of the estimated matrices has all its entries as zero, we substitute the corresponding zero row by a row with a single '1' in one of the* $|\mathbb{S}|$ *positions picked uniformly at random. As we shall demonstrate shortly, the ML estimates converge to their respective true values as more observations are accumulated. Therefore, such a substitution operation (or any modification thereof that replaces the all-zero rows by an arbitrary probability vector) needs to be carried out only for finitely many time slots, and does not affect the asymptotic performance of the policy.*

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

In this subsection, we show that the expected number of samples required by policy $\pi^\star(L, \delta)$ to find the index of the odd arm can be made arbitrarily close to that in (2.18) in the regime of vanishing error probabilities. We show that this can be achieved by choosing the parameters L and δ carefully. We organise this subsection as follows:

1. First, we show that when the odd arm index is h, the modified GLR test statistic $M_h(n)$ has a strictly positive drift under our policy. Subsequently, we show that our policy stops in finite time almost surely.

2. For any fixed target error probability $\epsilon > 0$, we show that by setting $L = 1/\epsilon$, our policy belongs to the family $\Pi(\epsilon)$, i.e., its probability of error at stoppage is within ϵ.

3. We obtain an upper bound on the expected stopping time of our policy, and demonstrate that this upper bound may be made arbitrarily close to the lower bound in (2.18) by choosing an appropriate value of $\delta \in (0, 1)$.

2.4.3.1 Strictly Positive Drift of the Modified GLR Test Statistic

Consider a version of the policy $\pi^\star(L, \delta)$ that never stops, i.e., a version that never checks for the condition in item (3) in the description of the policy. Call this version of the policy the *non-stopping version of* $\pi^\star(L, \delta)$. The main result on the strictly positive drift of the modified GLR test statistic is as described in the following proposition.

Proposition 2. *Fix parameters* $L > 1$ *and* $\delta \in (0, 1)$. *For all* $h, h' \in \mathcal{A}$ *such that* $h' \neq h$, *under the non-stopping version of* $\pi^\star(L, \delta)$,

$$\liminf_{n\to\infty} \frac{M_{hh'}(n)}{n} > 0 \quad \textit{almost surely.} \tag{2.39}$$

The proof of Proposition 2 is based on the key idea that forced exploration with parameter $\delta \in (0,1)$ (items 2(a) and 2(b) of policy $\pi^\star(L,\delta)$) results in sampling each arm with a strictly positive rate that grows linearly. It is in showing an analogue of Proposition 2 for iid Poisson observations that the authors of [6] use their result of [6, Proposition 3] on guaranteed exploration at a strictly positive rate. Since it is not clear if the analogue of [6, Proposition 3] holds in general, we use the idea in [10] of forced exploration. We present the details in Section 2.7.2. We refer the reader to [11] on how to make do with forced exploration at a sublinear rate.

As an immediate consequence of the above proposition, we note that almost surely,

$$\liminf_{n\to\infty} M_h(n) = \liminf_{n\to\infty} \min_{h'\neq h} M_{hh'}(n) > 0. \tag{2.40}$$

The result in (2.40) has the following implication. For $h, h' \in \mathcal{A}$ such that $h' \neq h$, let $\mathsf{GLR}_{hh'}(n)$ denote the classical GLR test statistic of hypothesis $\mathcal{H}_h$ with respect to $\mathcal{H}_{h'}$ at time n. Clearly, $M_{hh'}(n) \leq \mathsf{GLR}_{hh'}(n)$ almost surely for all n. Further, $\mathsf{GLR}_{hh'}(n) = -\mathsf{GLR}_{h'h}(n)$. We then have

$$\begin{aligned}
\limsup_{n\to\infty} M_{h'}(n) &= \limsup_{n\to\infty} \min_{a\neq h'} M_{h'a}(n) \\
&\leq \limsup_{n\to\infty} M_{h'h}(n) \\
&\leq \limsup_{n\to\infty} \mathsf{GLR}_{h'h}(n) \\
&= \limsup_{n\to\infty} -\mathsf{GLR}_{hh'}(n) \\
&\leq -\liminf_{n\to\infty} M_{hh'}(n) \\
&\leq -\liminf_{n\to\infty} M_h(n) \\
&< 0 \quad \text{almost surely.}
\end{aligned} \tag{2.41}$$

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

$$\theta(n) = h \quad \forall\ n \text{ sufficiently large.} \tag{2.42}$$

Let $\pi_h^\star(L,\delta)$ denote a version of the policy $\pi^\star(L,\delta)$ that stops only at declaration h. It then follows that the stopping times of policies $\pi^\star(L,\delta)$ and $\pi_h^\star(L,\delta)$ are almost surely related as $\tau(\pi_h^\star(L,\delta)) \geq \tau(\pi^\star(L,\delta))$, as a consequence of which we have the following set of almost sure inequalities:

$$\tau(\pi^\star(L,\delta)) \leq \tau(\pi_h^\star(L,\delta))$$

$$
\begin{aligned}
&= \inf\{n \geq 1 : M_h(n) \geq \log((K-1)L)\} \\
&\leq \inf\left\{n \geq 1 : M_{hh'}(n') \geq \log((K-1)L) \text{ for all } n' \geq n \text{ and for all } h' \neq h\right\} \\
&< \infty, \qquad (2.43)
\end{aligned}
$$

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

2.4.3.2 Error Probability of Policy $\pi^\star(L, \delta)$

We now show that given a target error probability $\epsilon > 0$, by setting $L = 1/\epsilon$, the policy $\pi^\star(L, \delta)$ belongs to the family $\Pi(\epsilon)$ for all $\delta \in (0, 1)$. This is formalised in the proposition below.

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

The proof uses Proposition 2 and the fact that policy $\pi^\star(L, \delta)$ stops in finite time almost surely. Further, the average in the numerator of the modified GLR test statistic, in place of the maximum in the classical GLR test statistic, plays a role. For the details, see Section 2.7.3.

2.4.3.3 Upper Bound on the Expected Stopping Time of Policy $\pi^\star(L, \delta)$

We conclude this section by presenting an upper bound on the expected stopping time of the policy $\pi^\star(L, \delta)$. We show that this upper bound may be made arbitrarily close to the lower bound in (2.18) by tuning δ appropriately.

As a first step, we show that under the non-stopping version of the policy $\pi^\star(L, \delta)$, when $C = (h, P_1, P_2)$ is the underlying arms configuration, the modified GLR process has an asymptotic drift that is close to $D^*(h, P_1, P_2)$ that appears in the lower bound (2.18).

Proposition 4. *Fix parameters $L \geq 1$ and $\delta \in (0, 1)$. Then, for all $h, h' \in \mathcal{A}$ such that $h' \neq h$, under the non-stopping version of $\pi^\star(L, \delta)$ and under the arms configuration $C = (h, P_1, P_2)$,*

$$
\lim_{n \to \infty} \frac{M_{hh'}(n)}{n} = D^*_\delta(h, P_1, P_2) \quad \textit{almost surely}, \qquad (2.44)
$$

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

$$
D^*_\delta(h, P_1, P_2) = \lambda^*_\delta \, D(P_1 || P_\delta | \mu_1) + (1 - \lambda^*_\delta) \frac{(K-2)}{(K-1)} D(P_2 || P_\delta | \mu_2), \qquad (2.45)
$$

with $\lambda^*_\delta = \frac{\delta}{K} + (1 - \delta)\lambda^*$ and for each $i, j \in \mathcal{S}$, $P_\delta(j|i)$ is as in (2.20) with λ_1 replaced by λ^*_δ.

We note that the policy $\pi^*(L, \delta)$ works with only the estimated transition probability matrices $\hat{P}^n_{\theta(n),1}$ and $\hat{P}^n_{\theta(n),2}$. To show (2.44), we must therefore ensure that the estimates approach

the true values and a property akin to continuity holds, that is, taking actions based on $\hat{P}^n_{\theta(n),1}$ and $\hat{P}^n_{\theta(n),2}$, which are only approximately close to P_1 and P_2, adds only $o(1)$ to the drift $D^*_\delta(h, P_1, P_2)$. This is the notion of *certainty equivalence* in control theory. The details of the proof may be found in Section 2.7.4.

We now state the main result of this section.

Proposition 5. *Fix $\delta \in (0,1)$. Under the policy $\pi = \pi^\star(L,\delta)$ and under the arms configuration $C = (h, P_1, P_2)$,*

$$\limsup_{L\to\infty} \frac{E^\pi[\tau(\pi)|C]}{\log L} \leq \frac{1}{D^*_\delta(h, P_1, P_2)}. \tag{2.46}$$

The proof uses Proposition 4 and involves showing that (a) the stopping time $\tau(\pi^\star(L,\delta))$ satisfies an asymptotic almost sure upper bound that matches with the right-hand side of (2.46), and (b) the family $\{\tau(\pi^\star(L,\delta))/\log L : L > 1\}$ is uniformly integrable. The almost sure convergence together with uniform integrability yields the relation (2.46). The details may be found in Section 2.7.5.

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

$$\lim_{\delta\downarrow 0} D^*_\delta(h, P_1, P_2) = D^*(h, P_1, P_2), \tag{2.47}$$

where $D^*(h, P_1, P_2)$ on the right-hand side of (2.47) is the same the constant that appears in the lower bound of (2.18). Thus, we note that δ may be tuned to make $D^*_\delta(h, P_1, P_2)$ as close as desired to $D^*(h, P_1, P_2)$, hence establishing the near-optimality of the policy $\pi^\star(L,\delta)$.

2.5 The Main Result

We now present the main result of this chapter, combining the lower and the upper bounds stated in Section 2.3 and Section 2.4 respectively.

Theorem 1. *Consider $K \geq 3$ independent Markov processes on a common finite state space that are irreducible, aperiodic and time homogeneous. Suppose that $C = (h, P_1, P_2)$ is the underlying configuration of the arms, where h is the odd arm index, and $P_2 \neq P_1$. Let $(\epsilon_n)_{n\geq 1}$ denote a sequence of error probability values with the property that $\epsilon_n \to 0$ as $n \to \infty$. Then, for each n and $\delta \in (0,1)$, the policy $\pi^\star(L_n,\delta)$ with $L_n = 1/\epsilon_n$ belongs to the family $\Pi(\epsilon_n)$. Furthermore, we have*

$$\liminf_{n\to\infty} \inf_{\pi\in\Pi(\epsilon_n)} \frac{E[\tau(\pi)|C]}{\log L_n} = \lim_{\delta\downarrow 0} \lim_{n\to\infty} \frac{E[\tau(\pi^\star(L_n,\delta))|C]}{\log L_n} = \frac{1}{D^*(h, P_1, P_2)}. \tag{2.48}$$

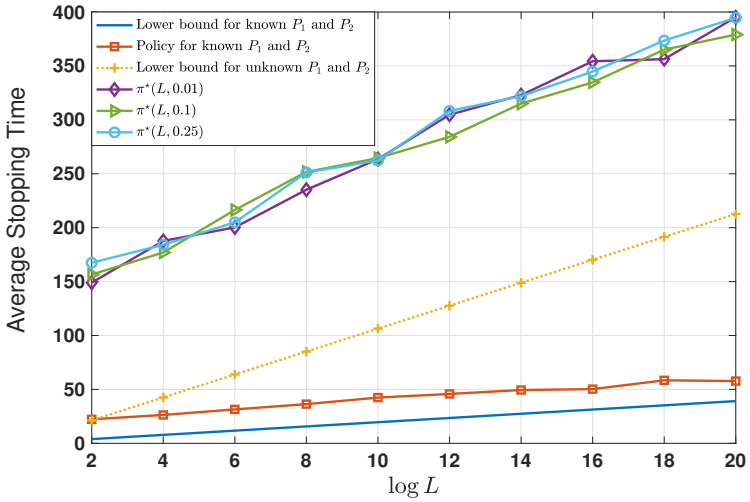


Figure 2.1: Plots of average stopping time of policy $\pi^\star(L, \delta)$, as function of $\log L$, for $\delta = 0.01, 0.1, 0.25$.

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

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

2.6 Simulation Results

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

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

Fig. 2.1 depicts the average stopping time of policy $\pi^\star(L, \delta)$ as a function of $\log L$, averaged over 100 rounds of iterations, for $\delta = 0.01, 0.1, 0.25$. For the aforementioned values of P_1 and P_2, numerical evaluation yields $D^*(h, P_1, P_2) \simeq 0.094$, thus resulting in a lower bound of $1/D^*(h, P_1, P_2) \simeq 10.635$. Since (2.18) is a statement about the slope of the growth rate of

average stopping time of policy $\pi^\star(L,\delta)$ as a function of $\log L$, the top 3 plots in the figure respect the lower bound in (2.18), with the slopes in these plots only marginally higher than that given by the lower bound. Theory predicts that as $\delta \downarrow 0$ and $L \to \infty$, the slopes will approach the lower bound. Also included in the figure are the plots of (a) the lower bound for the case when P_1 and P_2 are known, and (b) a policy similar to that of $\pi^\star(L,\delta)$ that uses the knowledge of P_1 and P_2 to identify the index of the odd arm. Such a policy clearly takes lesser time than $\pi^\star(L,\delta)$ to identify the index of the odd arm. The figure shows that the performance of this policy also matches in slope to that given by its lower bound for large values of L.

2.7 Proofs

2.7.1 Proof of Proposition 1

We first present below three lemmas that will be used in proving the proposition. The first of these, given below, is an analogue of the change of measure argument of Kaufmann et al. [1, Lemma 18] for the case of Markov observations from each arm. Recall that

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

where for each n, $\mathcal{F}_n$ is as defined in (2.12). Further, for any $h' \neq h$, define $Z_{hh'}(\tau) := Z_h(\tau) - Z_{h'}(\tau)$, where $Z_h(\tau) = \sum_{a=1}^K Z_h^a(\tau)$.

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

$$P^\pi(E|C') = E^\pi[\mathbb{I}_E \ \exp(-Z_{hh'}(\tau))|C]. \tag{2.49}$$

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

$$E_{h'}[g(A^n, \bar{X}^n)] = E_h[g(A^n, \bar{X}^n) \ \exp(-Z_{hh'}(n))]. \tag{2.50}$$

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

$$\begin{aligned} P_{h'}(E) &= E_{h'}[\mathbb{I}_E] \\ &\overset{(a)}{=} \sum_{n=0}^{\infty} E_{h'}[\mathbb{I}_E \mathbb{I}_{\{\tau=n\}}] \\ &\overset{(b)}{=} \sum_{n=0}^{\infty} E_h[\mathbb{I}_E \mathbb{I}_{\{\tau=n\}} \exp(-Z_{hh'}(n))] \\ &= E_h[\mathbb{I}_E \exp(-Z_{hh'}(\tau))], \end{aligned} \tag{2.51}$$

hence proving the lemma. In the above set of equations, (a) is due to monotone convergence theorem, and (b) follows from the application of (2.50) to the function $g(A^n, \bar{X}^n) = \mathbb{I}_E \cdot \mathbb{I}_{\{\tau=n\}}$ by noting that $E \in \mathcal{F}_\tau$, and therefore $E \cap \{\tau = n\} \in \mathcal{F}_n$ for all n.

We now proceed to prove (2.50) by induction on n. From (2.11) and (2.9), it follows that $Z_{hh'}(0) = 0$. Then, for any measurable function $g : \mathcal{A}^{n+1} \times \mathcal{S}^{n+1} \to \mathbb{R}$, the proof of (2.50) for $n = 0$ follows from the following set of equations:

$$\begin{aligned} E_{h'}[g(A_0, \bar{X}_0)] &= \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} P_{h'}(A_0 = a) \cdot P_{h'}(\bar{X}_0 = i | A_0 = a) \cdot g(a, i) \\ &= \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} P_{h'}(A_0 = a) \cdot P_{h'}(X_0^a = i) \cdot g(a, i) \\ &= \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} P_{h'}(A_0 = a) \cdot \nu(i) \cdot g(a, i) \\ &\overset{(a)}{=} \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} P_h(A_0 = a) \cdot P_h(X_0^a = i) \cdot g(a, i) \\ &= E_h[g(A_0, \bar{X}_0)] \\ &= E_h[g(A_0, \bar{X}_0) \exp(-Z_{hh'}(0))], \end{aligned} \tag{2.52}$$

where in writing (a), we use

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

We now assume that (2.50) holds for some positive integer n, and show that it also holds for $n+1$. By the law of iterated expectations, we have

$$E_{h'}[g(A^{n+1}, \bar{X}^{n+1})] = E_{h'}\left[E_{h'}\left[g(A^{n+1}, \bar{X}^{n+1})|A^n, \bar{X}^n\right]\right]. \tag{2.53}$$

Since the inner conditional expectation term on the right-hand side of (2.53) is a measurable function of $(A^n, \bar{X}^n)$, using the induction hypothesis, we get

$$\begin{aligned}
&E_{h'}[g(A^{n+1}, \bar{X}^{n+1})] \\
&= E_h\left[E_{h'}\left[g(A^{n+1}, \bar{X}^{n+1})|A^n, \bar{X}^n\right] \exp(-Z_{hh'}(n))\right] \\
&= \sum_{a^n \in \mathcal{A}^n} \sum_{\bar{x}^n \in \mathcal{S}^{n+1}} P_h(A^n = a^n, \bar{X}^n = \bar{x}^n) \cdot \exp(-z_{hh'}(n)) \cdot E_{h'}[g(A^{n+1}, \bar{X}^{n+1})|A^n = a^n, \bar{X}^n = \bar{x}^n],
\end{aligned} \tag{2.54}$$

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

$$\begin{aligned}
&E_{h'}[g(A^{n+1}, \bar{X}^{n+1})|A^n = a^n, \bar{X}^n = \bar{x}^n] \\
&= \sum_{a'=1}^{K} \sum_{j \in \mathcal{S}} g(a^n, a', \bar{x}^n, j) \cdot P_{h'}(A_{n+1} = a'|A^n = a^n, \bar{X}^n = \bar{x}^n) \cdot P_{h'}^{a'}(X^{a'}_{N_{a'}(n)} = j|X^{a'}_{N_{a'}(n)-1}) \\
&= \sum_{a'=1}^{K} \sum_{j \in \mathcal{S}} g(a^n, a', \bar{x}^n, j) \cdot P_h(A_{n+1} = a'|A^n = a^n, \bar{X}^n = \bar{x}^n) \cdot P_{h'}^{a'}(X^{a'}_{N_{a'}(n)} = j|X^{a'}_{N_{a'}(n)-1}),
\end{aligned} \tag{2.55}$$

where in writing the last line above, we use the fact that the probability of selecting an arm at any time, based on the history of past arm selections and observations, is independent of the underlying configuration of the arms, and is thus the same under hypotheses H_h and $H_{h'}$. We now write (2.55) as

$$\begin{aligned}
&E_{h'}[g(A^{n+1}, \bar{X}^{n+1})|A^n = a^n, \bar{X}^n = \bar{x}^n] \\
&= \sum_{a'=1}^{K} \sum_{j \in \mathcal{S}} \Bigg\{ g(a^n, a', \bar{x}^n, j) \cdot P_h(A_{n+1} = a'|A^n = a^n, \bar{X}^n = \bar{x}^n) \\
&\qquad \cdot \frac{P_{h'}^{a'}(X^{a'}_{N_{a'}(n)-1} = j|X^{a'}_{N_{a'}(n)-1})}{P_h^{a'}(X^{a'}_{N_{a'}(n)} = j|X^{a'}_{N_{a'}(n)-1})} \cdot P_h^{a'}(X^{a'}_{N_{a'}(n)} = j|X^{a'}_{N_{a'}(n)-1}) \Bigg\}.
\end{aligned} \tag{2.56}$$

Plugging back (2.56) in (2.54), and using

$$z_{hh'}(n+1) = z_{hh'}(n) + \log \frac{P_h^{a'}(X^{a'}_{N_{a'}(n)} = j | X^{a'}_{N_{a'}(n)-1})}{P_{h'}^{a'}(X^{a}_{N_{a'}(n)} = j | X^{a'}_{N_{a'}(n)-1})}, \tag{2.57}$$

we get

$$\begin{aligned}
&E_{h'}[g(A^{n+1}, \bar{X}^{n+1})] \\
&= \sum_{a^n \in \mathcal{A}^n} \sum_{\bar{x}^n \in \mathcal{S}^{n+1}} \sum_{a'=1}^{K} \sum_{j \in \mathcal{S}} \Big\{ g(a^n, a', \bar{x}^n, j) \cdot \exp(-z_{hh'}(n+1)) \\
&\qquad \cdot P_h(A^n = a^n, \bar{X}^n = \bar{x}^n) \cdot P_h(A_{n+1} = a', \bar{X}_{n+1} = j | A^n = a^n, \bar{X}^n = \bar{x}^n) \Big\} \\
&= E_h[g(A^{n+1}, \bar{X}^{n+1}) \exp(-Z_{hh'}(n+1))],
\end{aligned} \tag{2.58}$$

hence proving (2.49) . □

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

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

$$E^\pi[N_a(\tau, i, j)|C] = E^\pi[N_a(\tau, i)|C] \cdot P_h^a(j|i), \tag{2.59}$$

where $P_h^a(j|i)$ is as given in (2.7). □

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

$$E_h[E_h[N_a(\tau, i, j)|X_0^a]|N_a(\tau)] = E_h[E_h[N_a(\tau, i)|X_0^a]|N_a(\tau)] \cdot P_h^a(j|i). \tag{2.60}$$

Towards this, we note that

$$E_h[E_h[N_a(\tau, i, j)|X_0^a]|N_a(\tau)] = E_h\left[\sum_{m=1}^{N_a(\tau)-1} E_h[\mathbb{I}_{\{X^a_{m-1}=i,\, X^a_m=j\}}|X_0^a] \Big| N_a(\tau) \right]. \tag{2.61}$$

We now simplify the inner conditional expectation term in (2.61) by considering the cases $m = 1$ and $m \geq 2$ separately.

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

$$\begin{aligned} E_h[\mathbb{I}_{\{X_0^a=i,\, X_1^a=j\}}|X_0^a] &= \mathbb{I}_{\{X_0^a=i\}} \cdot E_h[\mathbb{I}_{\{X_1^a=j\}}|X_0^a] \\ &= \mathbb{I}_{\{X_0^a=i\}} \cdot P_h^a(X_1^a = j|X_0^a = i) \\ &= \mathbb{I}_{\{X_0^a=i\}} \cdot P_h^a(j|i). \end{aligned} \tag{2.62}$$

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

$$\begin{aligned} E_h[\mathbb{I}_{\{X_{m-1}^a=i,\, X_m^a=j\}}|X_0^a = k] &= P_h^a(X_{m-1}^a = i,\, X_m^a = j|X_0^a = k) \\ &\overset{(a)}{=} P_h^a(X_{m-1}^a = i|X_0^a = k) \cdot P_h^a(X_1^a = j|X_0^a = i) \\ &= E_h[\mathbb{I}_{\{X_{m-1}^a=i\}}|X_0^a = k] \cdot P_h^a(j|i), \end{aligned} \tag{2.63}$$

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

From the aforementioned cases, it follows that the relation

$$E_h[\mathbb{I}_{\{X_{m-1}^a=i,\, X_m^a=j\}}|X_0^a] = E_h[\mathbb{I}_{\{X_{m-1}^a=i\}}|X_0^a] \cdot P_h^a(j|i) \tag{2.64}$$

holds for all $m \geq 1$. Substituting (2.64) in (2.61) and simplifying, we arrive at (2.60). The lemma then follows by applying expectation $E_h[\cdot]$ to both sides of (2.60). □

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

Lemma 3. *For all $w_1, w_2 \in [0, 1]$ such that $w_1 + w_2 = 1$ and $\nu_1, \nu_2 \in \mathcal{P}(\mathcal{S})$,*

$$\min_{\psi \in \mathcal{P}(\mathcal{S})} [w_1 D(\nu_1||\psi) + w_2 D(\nu_2||\psi)] = w_1 D(\nu_1||\nu^*) + w_2 D(\nu_2||\nu^*), \tag{2.65}$$

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

Proof. This is well known with ν^* viewed as a root of "information centre" and the right-hand side of (2.65) viewed as a mutual information. Here is the proof for completeness.

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

$$w_1 D(\nu_1||\psi) + w_2 D(\nu_2||\psi) = w_1 D(\nu_1\|\nu^*) + w_2 D(\nu_2\|\nu^*) + D(\nu^*||\psi)$$

$$\geq D(\nu_1\|\nu^*) + w_2 D(\nu_2\|\nu^*), \tag{2.66}$$

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

Proof of Proposition 1. Fix an arbitrary $\epsilon > 0$, and let $\pi \in \Pi(\epsilon)$ be a policy whose stopping is $\tau = \tau(\pi)$. Without loss of generality, we assume that $E^\pi[\tau(\pi)|C] < \infty$, for otherwise the inequality (2.18) holds trivially. We organise the proof of the proposition into various sections. In the first of these sections presented below, we lower bound the expected value of $Z_{hh'}(\tau)$ in terms of the error probability ϵ. This uses the above Lemma 1, Lemma 2 and the result of [1, Lemma 19].

2.7.1.1 A Lower Bound on The Expected Value of $Z_{hh'}(\tau)$

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

$$E^\pi[Z_{hh'}(\tau)|C] \geq \sup_{E \in \mathcal{F}_\tau} d(P^\pi(E|C), P^\pi(E|C')), \tag{2.67}$$

where

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

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

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

$$\begin{aligned} d(P^\pi(E|C), P^\pi(E|C')) &\geq d(1-\epsilon, P^\pi(E|C')) \\ &\geq d(1-\epsilon, \epsilon). \end{aligned} \tag{2.68}$$

2. We may minimise both sides of (2.67) over all alternative configurations C' to obtain

$$\min_{C'=(h',P_1',P_2')} E^\pi[Z_{hh'}(\tau)|C] \geq \min_{C'=(h',P_1',P_2')} \sup_{E \in \mathcal{F}_\tau} d(P^\pi(E|C), P^\pi(E|C')). \tag{2.69}$$

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

$$\min_{C'=(h',P_1',P_2')} E^{\pi}[Z_{hh'}(\tau)|C] \geq d(\epsilon, 1-\epsilon). \tag{2.70}$$

2.7.1.2 A Relation Between $E^{\pi}[Z_{hh'}(\tau)|C]$ and $E^{\pi}[\tau|C]$

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

$$E^{\pi}[Z_{hh'}(\tau)|C] = \sum_{a=1}^{K} E^{\pi}\left[\sum_{m=1}^{N_a(\tau)-1} \log\left(\frac{P_h^a(X_m^a|X_{m-1}^a)}{P_{h'}^a(X_m^a|X_{m-1}^a)}\right)\Bigg|C\right], \tag{2.71}$$

where we take inner summation term to be zero whenever $N_a(\tau) < 2$. Focus on the expectation term in (2.71). This term may be written as

$$\begin{aligned} E^{\pi}\left[\sum_{m=1}^{N_a(\tau)-1} \log\left(\frac{P_h^a(X_m^a|X_{m-1}^a)}{P_{h'}^a(X_m^a|X_{m-1}^a)}\right)\Bigg|C\right] &\overset{(a)}{=} E^{\pi}\left[\sum_{m=1}^{N_a(\tau)-1} \sum_{i,j\in S} \mathbb{I}_{\{X_{m-1}^a=i,\, X_m^a=j\}} \log\left(\frac{P_h^a(j|i)}{P_{h'}^a(j|i)}\right)\Bigg|C\right] \\ &= \sum_{i,j\in S} E^{\pi}[N_a(\tau,i,j)|C]\, f_{hh'}^a(j|i), \end{aligned} \tag{2.72}$$

where (a) above follows from the fact that the Markov process of arm a is time homogeneous, and $f_{hh'}^a(j|i) := \log\left(\frac{P_h^a(j|i)}{P_{h'}^a(j|i)}\right)$. Using the result of Lemma 2 in (2.72), we get

$$\begin{aligned} E^{\pi}[Z_{hh'}(\tau)|C] &= \sum_{a=1}^{K} \sum_{i,j\in S} E^{\pi}[N_a(\tau,i)|C] \cdot P_h^a(j|i) \cdot f_{hh'}^a(j|i) \\ &= \sum_{a=1}^{K} \sum_{i\in S} E[N_a(\tau,i)|C]\, D(P_h^a(\cdot|i)||P_{h'}^a(\cdot|i)), \end{aligned} \tag{2.73}$$

where $D(P_h^a(\cdot|i)||P_{h'}^a(\cdot|i)) = \sum_{j\in S} P_h^a(j|i) f_{hh'}^a(j|i)$ denotes the KL divergence between the probability distributions $P_h^a(\cdot|i)$ and $P_{h'}^a(\cdot|i)$. We now express (2.73) by introducing some additional terms as below:

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

$$= (E^{\pi}[\tau+1|C]-K)\Bigg(\sum_{a=1}^{K}\left[\frac{E^{\pi}[N_a(\tau)|C]-1}{E^{\pi}[\tau+1|C]-K}\right]\sum_{i\in\mathcal{S}}p_h^a(i)\cdot D(P_h^a(\cdot|i)||P_{h'}^a(\cdot|i))\Bigg), \tag{2.74}$$

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

2.7.1.3 Asymptotics of Vanishing Error Probability

Since $\sum_{i\in\mathcal{S}} p_h^a(i) = 1$, the inner summation term over i in (2.74) represents the average of the numbers $(D(P_h^a(\cdot|i)||P_{h'}^a(\cdot|i)))_{i\in\mathcal{S}}$ with respect to $(p_h^a(i))_{i\in\mathcal{S}}$. Suppose that at some time, arm a is selected, and it makes a transition from state i to state j, for some $i,j\in\mathcal{S}$. Then, the next time arm a is selected, it makes a transition from state j to state k for some $k\in\mathcal{S}$. For $a\in\mathcal{A}$ and $i\in\mathcal{S}$, let

$$N^a(\tau,i) \coloneqq \sum_{m=2}^{N_a(\tau)} \mathbb{I}_{\{X_{m-1}^a=i\}} \tag{2.75}$$

denote the number of times arm a is observed to occupy state i after a transition. In conjunction with (2.3), it is easy to see that for each $i\in\mathcal{S}$, we have

$$N_a(\tau,i) = N^a(\tau,i) - \mathbb{I}_{\{X_{N_a(\tau)-1}^a=i\}} + \mathbb{I}_{\{X_0^a=i\}}, \tag{2.76}$$

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

$$\frac{E^{\pi}[N^a(\tau,i)|C]-1}{E^{\pi}[N_a(\tau)|C]-1} \le p_h^a(i) \le \frac{E^{\pi}[N^a(\tau,i)|C]+1}{E^{\pi}[N_a(\tau)|C]-1}. \tag{2.77}$$

Using (2.77) in (2.74), we arrive at the form

$$u-\Delta \le E^{\pi}[Z_{hh'}(\tau)] \le u+\Delta, \tag{2.78}$$

where the terms u and Δ are as below:

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

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

$$+\sum_{a\neq h}\sum_{i\in\mathcal{S}} D(P_2(\cdot|i)||P_2'(\cdot|i)). \tag{2.79}$$

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

$$\mu_h^a(i) = \begin{cases} \mu_1(i), & a = h, \\ \mu_2(i), & a \neq h. \end{cases} \tag{2.80}$$

Then, as $\epsilon \downarrow 0$, we have that both the lower and upper bounds in (2.77) converge to $\mu_h^a(i)$. We shall soon exploit this fact below to arrive at the lower bound. Going further, we denote by $(q_h^a(i))_{i\in\mathcal{S}}$ the probability distribution given by

$$q_h^a(i) = \frac{E^\pi[N^a(\tau, i)|C]}{E^\pi[N_a(\tau)|C] - 1}, \quad i \in \mathcal{S}. \tag{2.81}$$

Using the upper bound in (2.78) in combination with (2.70), we have the following chain of inequalities:

$$\begin{aligned} d(\epsilon, 1-\epsilon) &\leq \min_{C'=(h',P_1',P_2')} E^\pi[Z_{hh'}(\tau)|C] \\ &\leq \min_{C'=(h',P_1',P_2')} (u + \Delta) \\ &\leq \min_{C'=(h',P_1',P_2')} u + \min_{C'=(h',P_1',P_2')} \Delta. \end{aligned} \tag{2.82}$$

The first term in (2.82) may be upper bounded as follows:

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

$$\stackrel{(a)}{=} (E^{\pi}[\tau+1|C]-K)\left\{\min_{C'=(h',P_1',P_2')}\left(\sum_{a=1}^{K}\left[\frac{E^{\pi}[N_a(\tau)|C]-1}{E^{\pi}[\tau+1|C]-K}\right]D(P_h^a(\cdot|\cdot)||P_{h'}^a(\cdot|\cdot)|q_h^a)\right)\right\}$$
$$\stackrel{(b)}{\leq} (E^{\pi}[\tau+1|C]-K)\left\{\max_{\lambda\in\mathcal{P}(\mathcal{A})}\min_{C'=(h',P_1',P_2')}\left(\sum_{a=1}^{K}\lambda(a)D(P_h^a(\cdot|\cdot)||P_{h'}^a(\cdot|\cdot)|q_h^a)\right)\right\}, \tag{2.83}$$

where, in (a) above,

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

while (b) follows by noting that maximising over the set $\mathcal{P}(\mathcal{A})$ of all probability distributions on the set of arms $\mathcal{A}$ only increases the right-hand side. The second term in (2.82) may be simplified as

$$\begin{aligned}
&\min_{C'=(h',P_1',P_2')}\Delta\\
&=\min_{P_1',P_2':P_1'\neq P_2'}\left\{\sum_{i\in\mathcal{S}}D(P_1(\cdot|i)||P_2'(\cdot|i))+\sum_{i\in\mathcal{S}}D(P_2(\cdot|i)||P_1'(\cdot|i))+\sum_{a\neq h}\sum_{i\in\mathcal{S}}D(P_2(\cdot|i)||P_2'(\cdot|i))\right\}\\
&\stackrel{(a)}{=}\min_{P_2'}\left\{\sum_{i\in\mathcal{S}}D(P_1(\cdot|i)||P_2'(\cdot|i))+\sum_{a\neq h}\sum_{i\in\mathcal{S}}D(P_2(\cdot|i)||P_2'(\cdot|i))\right\}\\
&=\min\left\{\sum_{i\in\mathcal{S}}D(P_1(\cdot|i)||P_2(\cdot|i)),\ (K-1)\sum_{i\in\mathcal{S}}D(P_2(\cdot|i)||P_1(\cdot|i))\right\},
\end{aligned}\tag{2.84}$$

where (a) above follows by noting that P_1' appears only in the term $D(P_2(\cdot|i)||P_1'(\cdot|i))$, and that for the choice $P_1'=P_2$, we get $D(P_2(\cdot|i)||P_1'(\cdot|i))=0$ for all $i\in\mathcal{S}$. For ease of notation, we shall denote the quantity in (2.84) by Δ', which we note is a constant.

Combining (2.83) with (2.82), we get the following relation after rearrangement:

$$d(\epsilon,1-\epsilon)\leq\Delta'+(E^{\pi}[\tau+1|C]-K)\left\{\max_{\lambda\in\mathcal{P}(\mathcal{A})}\min_{C'=(h',P_1',P_2')}\left[\sum_{a=1}^{K}\lambda(a)D(P_h^a(\cdot|\cdot)||P_{h'}^a(\cdot|\cdot)|q_h^a)\right]\right\}. \tag{2.85}$$

Since (2.85) is valid for any arbitrary choice of $\epsilon>0$ and for all $\pi\in\Pi(\epsilon)$, letting $\epsilon\downarrow 0$ and using $d(\epsilon,1-\epsilon)/\log\frac{1}{\epsilon}\to 1$ as $\epsilon\downarrow 0$, along with the fact that $q_h^a(i)\to\mu_h^a(i)$ for all $i\in\mathcal{S}$ in the regime of vanishing error probabilities, we get

$$\lim_{\epsilon\downarrow 0}\inf_{\pi\in\Pi(\epsilon)}\frac{E^{\pi}[\tau(\pi)|C]}{\log\frac{1}{\epsilon}}\geq\frac{1}{D^*(h,P_1,P_2)}, \tag{2.86}$$

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

$$D^*(h, P_1, P_2) = \max_{\lambda \in \mathcal{P}(\mathcal{A})} \min_{C'=(h', P_1', P_2')} \left(\sum_{a=1}^{K} \lambda(a)\, D(P_h^a(\cdot|\cdot)||P_{h'}^a(\cdot|\cdot)|\mu_h^a) \right). \tag{2.87}$$

We now show that the quantities in (2.87) and (2.19) are the same.

2.7.1.4 The Final Steps

Using (2.7) and (2.80), and using the shorthand notation $D(P_h^a||P_{h'}^a|\mu_h^a)$ to denote the KL divergence term inside the summation in (2.87), we get

$$\begin{aligned} &D^*(h, P_1, P_2) \\ &= \max_{\lambda \in \mathcal{P}(\mathcal{A})} \min_{h' \neq h,\, P_1',\, P_2'} \Big(\lambda(h)\, D(P_1||P_2'|\mu_1) + \lambda(h')\, D(P_2||P_1'|\mu_2) + (1 - \lambda(h) - \lambda(h'))D(P_2||P_2'|\mu_2) \Big). \end{aligned} \tag{2.88}$$

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

$$D^*(h, P_1, P_2) = \max_{\lambda \in \mathcal{P}(\mathcal{A})} \min_{h' \neq h,\, P_2'} \Big(\lambda(h)\, D(P_1||P_2'|\mu_1) + (1 - \lambda(h) - \lambda(h'))D(P_2||P_2'|\mu_2) \Big). \tag{2.89}$$

We now note that

$$\begin{aligned} \min_{h' \neq h}(1 - \lambda(h) - \lambda(h')) &= 1 - \lambda(h) - \max_{h' \neq h} \lambda(h') \\ &\overset{(a)}{\leq} 1 - \lambda(h) - \frac{1 - \lambda(h)}{K - 1} \\ &= (1 - \lambda(h))\frac{(K-2)}{(K-1)}, \end{aligned} \tag{2.90}$$

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

$$D^*(h, P_1, P_2) = \max_{0 \leq \lambda(h) \leq 1} \min_{P_2'} \Big(\lambda(h)\, D(P_1||P_2'|\mu_1) + (1 - \lambda(h))\frac{(K-2)}{(K-1)} D(P_2||P_2'|\mu_2) \Big). \tag{2.91}$$

Using Lemma 3 in (2.91), and recognising that the hand side of (2.91) is not a function of h, we write

$$D^*(h, P_1, P_2) = \max_{0\le\lambda_1\le 1} \left(\lambda_1 \, D(P_1||P|\mu_1) + (1-\lambda_1)\frac{(K-2)}{(K-1)} D(P_2||P|\mu_2) \right), \tag{2.92}$$

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

$$P(j|i) = \frac{\lambda_1 \mu_1(i) P_1(j|i) + (1-\lambda_1)\frac{(K-2)}{(K-1)}\mu_2(i) P_2(j|i)}{\lambda_1 \mu_1(i) + (1-\lambda_1)\frac{(K-2)}{(K-1)}\mu_2(i)}. \tag{2.93}$$

Noting that the right-hand sides of (2.92) and (2.19) are identical, this completes the proof of the proposition. □

2.7.2 Proof of Proposition 2

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

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

$$\frac{N_a(n,i,j)}{N_a(n,i)} \longrightarrow \begin{cases} P_1(j|i), & a = h, \\ P_2(j|i), & a \neq h, \end{cases}, \qquad \frac{\sum_{a\neq h} N_a(n,i,j)}{\sum_{a\neq h} N_a(n,i)} \longrightarrow P_2(j|i). \tag{2.94}$$

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

$$S_a(n) = \sum_{t=0}^{n-1} \left(\mathbb{I}_{\{A_{t+1}=a\}} - P(A_{t+1} = a | A^t, \bar{X}^t) \right), \tag{2.95}$$

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

$$P(A_{t+1} = a|A^t, \bar{X}^t) = \frac{\delta}{K} + (1-\delta)\, \lambda^*(h^*(t), \hat{P}^t_{h^*(t),1}, \hat{P}^t_{h^*(t),2})(a). \tag{2.96}$$

Letting $d^a_{t+1} = \mathbb{I}_{\{A_{t+1}=a\}} - P(A_{t+1} = a|A^t, \bar{X}^t)$, we note that $P(|d_{t+1}| \le 2|A^t, \bar{X}^t) = 1$ for all $t \ge 0$, implying that $\{d_t\}_{t\ge 0}$ is bounded uniformly almost surely. Since $\{d_{t+1}\}_{t\ge 0}$ is a martingale difference sequence, it follows from [23, Th. 1.2A] that for every $\epsilon > 0$, there exists $c_\epsilon > 0$ such

that $P(\frac{S_a(n)}{n} > \epsilon) \leq e^{-nc_\epsilon}$. From this, it follows that $S_a(n)/n \to 0$ almost surely. This implies that the following is true almost surely for sufficiently large values of n:

$$\frac{\delta}{2K} < \frac{N_a(n) - 1}{n} < 1 + \frac{\delta}{2K}. \tag{2.97}$$

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

$$\frac{N_a(n,i)}{N_a(n)} \longrightarrow \mu_h^a(i), \quad \frac{N_a(n,i,j)/N_a(n)}{N_a(n,i)/N_a(n)} \longrightarrow P_h^a(j|i); \tag{2.98}$$

here, $\mu_h^a(i)$ and $P_h^a(j|i)$ are as defined in (2.80) and (2.7) respectively. This establishes the convergence in the first line of (2.94) under the assumption that $C = (h, P_1, P_2)$ is the underlying configuration of the arms.

We then note that almost surely,

$$\begin{aligned}\frac{\sum\limits_{a\neq h} N_a(n,i,j)}{\sum\limits_{a\neq h} N_a(n,i)} &= \frac{\sum\limits_{a\neq h} \frac{N_a(n,i,j)}{N_h^a(n,i)} \frac{N_h^a(n,i)}{N_h^a(n)} \frac{N_h^a(n)}{n}}{\sum\limits_{a\neq h} \frac{N_a(n,i)}{N_h^a(n)} \frac{N_h^a(n)}{n}} \\ &\overset{n\to\infty}{\longrightarrow} P_2(j|i),\end{aligned} \tag{2.99}$$

where the convergence in the last line above follows from (2.98) by noting that for $a \neq h$, when $C = (h, P_1, P_2)$ is the underlying configuration of the arms, $\mu_h^a(i) = \mu_2(i)$ and $P_h^a(j|i) = P_2(j|i)$. This establishes the convergence in the second line of (2.94), thus completing the proof of the lemma. □

Proof of Proposition 2. We now use Lemma 4 to show that (2.39) holds for any $h' \neq h$. Towards this, we show that the quantity on the right-hand side of (2.30) is strictly positive.

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

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

$$\lim_{n\to\infty} \frac{T_1}{n} = 0, \tag{2.100}$$

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

2. From (2.32), we have

$$\frac{T_2(n)}{n} = \frac{1}{n}\sum_{i\in\mathcal{S}} \log B((N_h(n,i,j)+1)_{j\in\mathcal{S}}). \tag{2.101}$$

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

$$\log B((N_h(n,i,j)+1)_{j\in\mathcal{S}}) = \log E\left[\prod_{j\in\mathcal{S}} X_{ij}^{N_h(n,i,j)}\right], \tag{2.102}$$

where the random vector $(X_{ij})_{j\in\mathcal{S}}$ follows Dirichlet distribution with parameters $\alpha_j = 1$ for all $j \in \mathcal{S}$. We now write (2.102) as follows:

$$\frac{1}{N_h(n)} \log B((N_h(n,i,j)+1)_{j\in\mathcal{S}}) = \frac{1}{N_h(n)} \log E\left[\exp\left(N_h(n)\sum_{j\in\mathcal{S}} \frac{N_h(n,i,j)}{N_h(n)} \log X_{ij}\right)\right]. \tag{2.103}$$

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

$$\begin{aligned}
&\frac{1}{N_h(n)} \log B((N_h(n,i,j)+1)_{j\in\mathcal{S}}) \\
&\geq \frac{1}{N_h(n)} \log E\left[\exp\left(N_h(n)\sum_{j\in\mathcal{S}} (\mu_1(i)P_1(j|i)+\epsilon') \log X_{ij}\right)\right].
\end{aligned} \tag{2.104}$$

Noting that $N_h(n)$ converges almost surely to $+\infty$ as $n \to \infty$, by Varadhan's integral lemma [24, Theorem 4.3.1], there exists a positive integer $M_{22} = M_{22}(\epsilon')$ such that for all $n \geq M_2 = \max\{M_{21}, M_{22}\}$, we have

$$\begin{aligned}
\frac{1}{N_h(n)} \log B((N_h(n,i,j)+1)_{j\in\mathcal{S}}) &\overset{(a)}{\geq} \sup_{\{z_j\geq 0, \sum_{j\in\mathcal{S}} z_j = 1\}} \sum_{j\in\mathcal{S}} (\mu_1(i)P_1(j|i)+\epsilon') \log z_j - \frac{\epsilon'}{|\mathcal{S}|} \\
&= \sum_{j\in\mathcal{S}} (\mu_1(i)P_1(j|i)+\epsilon') \log \frac{\mu_1(i)P_1(j|i)+\epsilon'}{\mu_1(i)+\epsilon'|\mathcal{S}|} - \frac{\epsilon'}{|\mathcal{S}|},
\end{aligned} \tag{2.105}$$

where the supremum on the right-hand side of (a) above is computed over all vectors

$(z_j)_{j\in\mathcal{S}}$ such that $z_j \geq 0$ for all $j \in \mathcal{S}$, and $\sum_{j\in\mathcal{S}} z_j = 1$. Plugging (2.105) into (2.101), we get

$$\frac{T_2(n)}{n} \geq \frac{N_h(n)}{n}\left\{\left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}}(\mu_1(i)P_1(j|i)+\epsilon')\log\frac{\mu_1(i)P_1(j|i)+\epsilon'}{\mu_1(i)+\epsilon'|\mathcal{S}|}\right]-\epsilon'\right\} \tag{2.106}$$

for all $n \geq M_2$.

3. From (2.33), we have

$$\frac{T_3(n)}{n} = \frac{1}{n}\sum_{i\in\mathcal{S}}\log B\left(\left(\sum_{a\neq h}N_a(n,i,j)+1\right)_{j\in\mathcal{S}}\right). \tag{2.107}$$

Using the same arguments as those used to simplify (2.101), we obtain the following: there exists a positive integer $M_3 = M_3(\epsilon')$ such that for all $n \geq M_3$, we have

$$\frac{T_3(n)}{n} \geq \frac{\sum_{a\neq h}N_a(n)}{n}\left\{\left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}}(\mu_2(i)P_2(j|i)+\epsilon')\log\frac{\mu_2(i)P_2(j|i)+\epsilon'}{\mu_2(i)+\epsilon'|\mathcal{S}|}\right]-\epsilon'\right\}. \tag{2.108}$$

4. From (2.34), we have

$$\frac{T_4(n)}{n} = -\frac{1}{n}\sum_{i,j\in\mathcal{S}}N_{h'}(n,i,j)\log\frac{N_{h'}(n,i,j)}{N_{h'}(n,i)}. \tag{2.109}$$

If $N_h(n,i) = 0$ for some state $i \in \mathcal{S}$ (in which case it follows that $N_h(n,i,j) = 0$ for all $j \in \mathcal{S}$), or if $N_h(n,i,j) = 0$[1] for some pair of states $i, j \in \mathcal{S}$, then the corresponding terms in the summation in (2.109) will be of the form $0\log\frac{0}{0}$ or $0\log 0$ respectively, which we treat as zero by convention. Thus, without loss of generality, we assume that $N_h(n,i,j) > 0$ for all $i, j \in \mathcal{S}$.

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

$$\begin{aligned}\frac{N_{h'}(n,i,j)}{n} &\to \mu_2(i)P_2(j|i),\\ \frac{N_{h'}(n,i,j)}{N_{h'}(n,i)} &\to P_2(j|i).\end{aligned} \tag{2.110}$$

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

Using these in (2.109), we get that there exists a positive integer $M_4 = M_4(\epsilon')$ such that for all $n \geq M_4$, we have

$$\frac{T_4(n)}{n} \geq \sum_{i,j \in \mathcal{S}} (\mu_2(i) P_2(j|i) - \epsilon') \log \frac{1}{P_2(j|i) + \epsilon'}. \tag{2.111}$$

5. Lastly, we present a simplification of the term $T_5(n)/n$. From (2.35), we have

$$\frac{T_5(n)}{n} = -\frac{1}{n} \sum_{i,j \in \mathcal{S}} \sum_{a \neq h'} N_a(n,i,j) \log \frac{\sum\limits_{a \neq h'} N_a(n,i,j)}{\sum\limits_{a \neq h'} N_a(n,i)}. \tag{2.112}$$

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

$$P_n(j|i) = \frac{\sum\limits_{a \neq h'} N_a(n,i,j)}{\sum\limits_{a \neq h'} N_a(n,i)}. \tag{2.113}$$

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

$$\frac{\sum\limits_{a \neq h,h'} N_a(n,i,j)}{\sum\limits_{a \neq h,h'} N_a(n,i)} \overset{n \to \infty}{\longrightarrow} P_2(j|i), \quad \frac{\sum\limits_{a \neq h,h'} N_a(n,i)}{\sum\limits_{a \neq h,h'} N_a(n)} \overset{n \to \infty}{\longrightarrow} \mu_2(i). \tag{2.114}$$

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

$$\begin{aligned} \frac{T_5(n)}{n} \geq \frac{N_h(n)}{n} \sum_{i,j \in \mathcal{S}} (\mu_1(i) P_1(j|i) - \epsilon') \log \frac{1}{P_n(j|i)} \\ + \frac{\sum\limits_{a \neq h,h'} N_a(n)}{n} \sum_{i,j \in \mathcal{S}} (\mu_2(i) P_2(j|i) - \epsilon') \log \frac{1}{P_n(j|i)}. \end{aligned} \tag{2.115}$$

Combining the results in (2.100), (2.106), (2.108), (2.111) and (2.115), we get that for all $n \geq M(\epsilon') = \max\{M_1, \ldots, M_5\}$, we have

$$\frac{M_{hh'}(n)}{n} \geq f_n(\epsilon'), \tag{2.116}$$

where $f_n(\epsilon')$ denotes the sum of the terms of the right-hand sides of (2.100), (2.106), (2.108), (2.111) and (2.115).

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

$$f_n(0) \coloneqq \frac{N_h(n)}{n} D(P_1||P_n|\mu_1) + \frac{\sum\limits_{a\neq h,h'} N_a(n)}{n} D(P_2||P_n||\mu_2). \tag{2.117}$$

Then, by continuity, we have that for any choice of $\epsilon > 0$, there exists $\epsilon' > 0$ such that $f_n(\epsilon') > f_n(0) - \epsilon$ for all sufficiently large values of n. From (2.116), this implies that

$$\frac{M_{hh'}(n)}{n} > f_n(0) - \epsilon \tag{2.118}$$

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

$$\liminf_{n\to\infty} \left[\frac{M_{hh'}(n)}{n} - f_n(0) \right] \geq -\epsilon. \tag{2.119}$$

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

$$\liminf_{n\to\infty} \frac{M_{hh'}(n)}{n} - \limsup_{n\to\infty} f_n(0) \geq 0, \tag{2.120}$$

from which it follows that

$$\begin{aligned}
&\liminf_{n\to\infty} \frac{M_{hh'}(n)}{n} \geq \limsup_{n\to\infty} f_n(0) \\
&\geq \liminf_{n\to\infty} f_n(0) \\
&\geq \liminf_{n\to\infty} \left\{ \frac{N_h(n)}{n} D(P_1||P_n|\mu_1) + \frac{\sum\limits_{a\neq h,h'} N_a(n)}{n} D(P_2||P_n|\mu_2) \right\} \\
&\geq \liminf_{n\to\infty} \left\{ \frac{N_h(n)}{n} D(P_1||P_n|\mu_1) \right\} + \liminf_{n\to\infty} \left\{ \frac{\sum\limits_{a\neq h,h'} N_a(n)}{n} D(P_2||P_n|\mu_2) \right\}
\end{aligned} \tag{2.121}$$

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

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

$$
\begin{aligned}
&\geq \frac{\sum\limits_{a\neq h'} N_a(n,i,j)}{n} \\
&\geq \frac{N_h(n)}{n}\cdot\frac{N_h(n,i)}{N_h(n)}\cdot\frac{N_h(n,i,j)}{N_h(n,i)} + \frac{\sum\limits_{a\neq h,h'} N_a(n)}{n}\cdot\frac{\sum\limits_{a\neq h,h'} N_a(n,i)}{\sum\limits_{a\neq h,h'} N_a(n)}\cdot\frac{\sum\limits_{a\neq h,h'} N_a(n,i,j)}{\sum\limits_{a\neq h,h'} N_a(n,i)} \\
&\overset{(a)}{\geq} \left(\frac{\delta}{2K}\right)\left(\frac{\mu_1(i)\,P_1(j|i)}{2}\right) + (K-2)\left(\frac{\delta}{2K}\right)\left(\frac{\mu_2(i)\,P_2(j|i)}{2}\right) \\
&\overset{(b)}{\geq} \left(\frac{\delta}{2K}\right)\left(\frac{\mu_1(i)P_1(j|i)+\mu_2(i)P_2(j|i)}{2}\right) \\
&\geq \left(\frac{\delta}{2K}\right)\left(\min\left\{\min_{i\in\mathcal{S}}\mu_1(i),\,\min_{i\in\mathcal{S}}\mu_2(i)\right\}\right)\left(\frac{P_1(j|i)+P_2(j|i)}{2}\right) \text{ almost surely} \qquad (2.122)
\end{aligned}
$$

for all sufficiently large values of n, where (a) follows from (2.97) and Lemma 4, and (b) follows by using the fact that the number of arms $K \geq 3$. It then follows that

$$
\begin{aligned}
&D(P_1||P_n|\mu_1) \qquad (2.123)\\
&\quad= \sum_{i\in\mathcal{S}}\mu_1(i)\sum_{j\in\mathcal{S}}P_1(j|i)\log\frac{P_1(j|i)}{P_n(j|i)} \\
&\quad\leq \sum_{i,j\in\mathcal{S}}\mu_1(i)\,P_1(j|i)\log\frac{P_1(j|i)}{\frac{P_1(j|i)+P_2(j|i)}{2}} + \sum_{i,j\in\mathcal{S}}\mu_1(i)P_1(j|i)\log P_1(j|i) \\
&\qquad\qquad + \log\frac{1}{\left(\frac{\delta}{2K}\right)\left(\min\left\{\min_{i\in\mathcal{S}}\mu_1(i),\,\min_{i\in\mathcal{S}}\mu_2(i)\right\}\right)} \\
&\quad= D\left(P_1\middle|\middle|\frac{P_1+P_2}{2}\middle|\mu_1\right) + \sum_{i\in\mathcal{S}}\mu_1(i)(-H(P_1(\cdot|i)) \\
&\qquad\qquad + \log\frac{1}{\left(\frac{\delta}{2K}\right)\left(\min\left\{\min_{i\in\mathcal{S}}\mu_1(i),\,\min_{i\in\mathcal{S}}\mu_2(i)\right\}\right)} \\
&\quad< \infty \text{ almost surely.} \qquad (2.124)
\end{aligned}
$$

On similar lines, it can be shown that $D(P_2||P_n|\mu_1)$ is bounded uniformly almost surely for all $n \geq 0$. Using the uniform boundedness property just proved, we may express (2.121) as

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

$$\geq \left(\frac{\delta}{2K}\right)\left(\liminf_{n\to\infty} D(P_1||P_n|\mu_1) + (K-2)\liminf_{n\to\infty} D(P_2||P_n|\mu_2)\right) \text{ almost surely,} \tag{2.125}$$

where the last line follows from (2.97).

Finally, we show that the first limit infimum term in (2.125) is strictly positive, and note that an exactly parallel argument may be used to show that the second limit infimum term is also strictly positive. Suppose that $\liminf_{n\to\infty} D(P_1||P_n|\mu_1) = 0$ almost surely. By the property that KL divergence is zero if and only if the argument probability distributions are identical, it follows that there exists a subsequence $(n_k)_{k\geq 1}$ such that $P_{n_k}(j|i) \to P_1(j|i)$ as $k \to \infty$ almost surely for all $i, j \in \mathcal{S}$. We now fix attention to this subsequence, and note that by the property that the sequences $(N_h(n_k)/n_k)_{k\geq 1}$ and $(\sum_{a\neq h,h'} N_a(n_k)/n_k)_{k\geq 1}$ are bounded, there exists a further subsequence $(n_{k_l})_{l\geq 1}$ of $(n_k)_{k\geq 1}$ such that the aforementioned bounded sequences admit limits, say α and β respectively. From Lemma 4, we then have the following convergence almost surely as $l \to \infty$:

$$P_{n_{k_l}}(j|i) \to \frac{\alpha\,\mu_1(i)\,P_1(j|i) + \beta\,\mu_2(i)\,P_2(j|i)}{\alpha\,\mu_1(i) + \beta\,\mu_2(i)}. \tag{2.126}$$

However, we note that the right-hand side of (2.126) is not equal to $P_1(j|i)$ whenever $P_2(j|i) > 0$, thus resulting in a contradiction. This completes the proof of the proposition. □

2.7.3 Proof of Proposition 3

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

1. The policy never stops in finite time.

2. The policy stops in finite time and declares $h' \neq h$ as the true index of the odd arm.

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

$$P_e^\pi = P^\pi(I(\pi) \neq h|C) = P^\pi\left(\exists\ n \text{ and } \ h' \neq h \text{ such that } I(\pi) = h' \text{ and } \tau(\pi) = n \middle| C\right). \tag{2.127}$$

We now let

$$\mathcal{R}_{h'}(n) \coloneqq \{\omega : \tau(\pi)(\omega) = n,\ I(\pi)(\omega) = h'\} \tag{2.128}$$

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

$$
\begin{aligned}
P_e^{\pi} &= P^{\pi}\left(\bigcup_{h' \neq h} \bigcup_{n=0}^{\infty} \mathcal{R}_{h'}(n) \middle| C\right) \\
&= \sum_{h' \neq h} \sum_{n=0}^{\infty} P^{\pi}(\tau(\pi) = n, I(\pi) = h' | C) \\
&= \sum_{h' \neq h} \sum_{n=0}^{\infty} \int\limits_{\mathcal{R}_{h'}(n)} dP^{\pi}(\omega | C) \\
&\overset{(a)}{=} \sum_{h' \neq h} \sum_{n=0}^{\infty} \int\limits_{\mathcal{R}_{h'}(n)} f(A^n(\omega), \bar{X}^n(\omega) | H_h) \left[\prod_{t=0}^{n} P_h(A_t | A^{t-1}, \bar{X}^{t-1})\right] d(A^n(\omega), \bar{X}^n(\omega)) \\
&\overset{(b)}{\leq} \sum_{h' \neq h} \sum_{n=0}^{\infty} \int\limits_{\mathcal{R}_{h'}(n)} \hat{f}(A^n(\omega), \bar{X}^n(\omega) | H_h) \left[\prod_{t=0}^{n} P_h(A_t | A^{t-1}, \bar{X}^{t-1})\right] d(A^n(\omega), \bar{X}^n(\omega)) \\
&\overset{(c)}{=} \sum_{h' \neq h} \sum_{n=0}^{\infty} \left\{ \int\limits_{\mathcal{R}_{h'}(n)} e^{-M_{h'h}(n)}\, f(A^n(\omega), \bar{X}^n(\omega) | H_{h'}) \left[\prod_{t=0}^{n} P_{h'}(A_t | A^{t-1}, \bar{X}^{t-1})\right] d(A^n(\omega), \bar{X}^n(\omega)) \right\} \\
&\leq \sum_{h' \neq h} \sum_{n=0}^{\infty} \left\{ \int\limits_{\mathcal{R}_{h'}(n)} \frac{1}{(K-1)L}\, dP^{\pi}(\omega | C') \right\} \\
&= \sum_{h' \neq h} \frac{1}{(K-1)L}\, P^{\pi}\left(\bigcup_{n=0}^{\infty} \mathcal{R}_{h'}(n) \middle| C'\right) \leq \frac{1}{L}, \qquad (2.129)
\end{aligned}
$$

where in (a) above, $P_h(A_t | A^{t-1}, \bar{X}^{t-1})$ denotes the probability of selecting arm A_t at time t when the index of the odd arm is h, with the convention that at time $t = 0$, this term represents $P_h(A_0)$; (b) above follows by the definition of $\hat{f}$ in (2.29), and (c) follows by using the fact that the probability of selecting an arm at any time t, based on the history of past arm selections and observations, is independent of the odd arm index, and is thus the same when the arm indexed by either h or h' is the odd arm. Setting $L = 1/\epsilon$ gives $P_e^{\pi} \leq \epsilon$, thus proving that $\pi = \pi^{\star}(L, \delta) \in \Pi(\epsilon)$. This completes the proof of the proposition. □

2.7.4 Proof of Proposition 4

Before we present the proof of Proposition 4, we show that the odd arm chosen by the non-stopping version of policy $\pi^{\star}(L, \delta)$ is indeed the correct one. Further, we show that the arm

selection frequencies under the same policy converge to the respective optimal values given in (2.22).

Proposition 6. *Let $C = (h, P_1, P_2)$ denote the underlying configuration of the arms. Fix $L \geq 1$ and $\delta \in (0,1)$, and consider the non-stopping version of policy $\pi^\star(L, \delta)$. For any $h' \neq h$ and $i, j \in \mathcal{S}$, let $P_n(j|i)$ be defined as in (2.113), Then, the following convergences hold almost surely as $n \to \infty$.*

$$\theta(n) \to h, \tag{2.130}$$

$$\lambda_{opt}(\theta(n), \hat{P}^n_{\theta(n),1}, \hat{P}^n_{\theta(n),2}) \to \lambda_{opt}(h, P_1, P_2), \tag{2.131}$$

$$\frac{N_a(n)}{n} \to \lambda^*_\delta(h, P_1, P_2)(a) \text{ for all } a \in \mathcal{A}, \tag{2.132}$$

$$P_n(j|i) \to P_\delta(j|i) \text{ for all } i, j \in \mathcal{S}, \tag{2.133}$$

where for each $a \in \mathcal{A}$ and each $i, j \in \mathcal{S}$, the quantity $\lambda^*_\delta(h, P_1, P_2)(a)$ and the term $P_\delta(j|i)$ in (2.133) are as defined in the statement of Proposition 4. □

Proof. We already established that (2.41) holds for all sufficiently large n. This establishes (2.130), which in turn implies that

$$\lambda_{opt}(\theta(n), \hat{P}^n_{\theta(n),1}, \hat{P}^n_{\theta(n),2}) \to \lambda_{opt}(h, P_1, P_2), \tag{2.134}$$

because of the convergence of the maximum likelihood estimates shown in (2.94), and the fact that $\lambda^*(h, P, Q)$ is jointly continuous in the pair (P, Q), a fact that follows from Berge's Maximum Theorem [25]. This establishes (2.131).

We now proceed to show (2.132). Towards this, we observe that from (2.38) and the convergence in (2.131), we have

$$\begin{aligned} P(A_{n+1} = a | A^n, \bar{X}^n) &= \frac{\delta}{K} + (1-\delta)\, \lambda_{opt}(\theta(n), \hat{P}^n_{\theta(n),1}, \hat{P}^n_{\theta(n),2})(a) \\ &\to \frac{\delta}{K} + (1-\delta)\lambda_{opt}(h, P_1, P_2)(a). \end{aligned} \tag{2.135}$$

We revisit the quantity $S_a(n)$ defined in (2.95), and use the fact that $\frac{S_a(n)}{n} \to 0$ almost surely as $n \to \infty$ to obtain

$$\frac{N_a(n)}{n} \to \frac{1}{n} \sum_{t=0}^{n-1} P(A_{t+1} = a | A^t, \bar{X}^t)$$

$$\to \frac{\delta}{K} + (1-\delta)\lambda_{opt}(h, P_1, P_2)(a). \tag{2.136}$$

This establishes (2.132).

Defining

$$\alpha_n := \frac{N_h(n)}{n}, \quad \beta_n := \frac{\sum_{a \neq h, h'} N_a(n)}{n}, \tag{2.137}$$

we note that the convergence in (2.132) implies in particular that

$$\begin{aligned}
\alpha_n &\to \lambda^*_\delta(h, P_1, P_2)(h) = \frac{\delta}{K} + (1-\delta)\lambda^* = \lambda^*_\delta, \\
\beta_n &\to (K-2)\left(\frac{\delta}{K} + (1-\delta)\frac{1-\lambda^*}{K-1}\right) \\
&= \frac{(K-2)}{(K-1)}(1-\lambda^*_\delta).
\end{aligned} \tag{2.138}$$

Taking limits as $n \to \infty$ on both sides of (2.113), and using the above limits for α_n and β_n, we get the convergence in (2.133), hence completing the proof of the proposition. □

Proof of Proposition 4. We recall from (2.121) and (2.117) that

$$\begin{aligned}
\liminf_{n\to\infty} \frac{M_{hh'}(n)}{n} &\geq \liminf_{n\to\infty} \alpha_n D(P_1||P_n|\mu_1) + \liminf_{n\to\infty} \beta_n D(P_2||P_n|\mu_2) \\
&= \lambda^*_\delta D(P_1||P_\delta|\mu_1) + \frac{(K-2)}{(K-1)}(1-\lambda^*_\delta) D(P_2||P_\delta||\mu_2),
\end{aligned} \tag{2.139}$$

where the terms α_n and β_n are as given in (2.137). Using Varadhan's integral lemma [24, Theorem 4.3.1] to write

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log B((N_h(n,i,j)+1)_{j\in\mathcal{S}}) &\leq \limsup_{n\to\infty} \frac{N_h(n)}{n}\mu_1(i) \sup_{\{z_j \geq 0, \sum_{j\in\mathcal{S}} z_j = 1\}} \sum_{j\in\mathcal{S}} P_1(j|i) \log z_j \\
&= \lim_{n\to\infty} \frac{N_h(n)}{n}\mu_1(i)(-H(P_1(\cdot|i))),
\end{aligned} \tag{2.140}$$

and following similar steps leading to (2.106), we obtain

$$\begin{aligned}
\limsup_{n\to\infty} \frac{M_{hh'}(n)}{n} &\leq \lim_{n\to\infty} \alpha_n D(P_1||P_n|\mu_1) + \lim_{n\to\infty} \beta_n D(P_2||P_n|\mu_2) \\
&= \lambda^*_\delta D(P_1||P_\delta|\mu_1) + \frac{(K-2)}{(K-1)}(1-\lambda^*_\delta) D(P_2||P_\delta||\mu_2).
\end{aligned} \tag{2.141}$$

Combining (2.139) and (2.141), we get the desired result. □

2.7.5 Proof of Proposition 5

This section is organised as follows. We first show in Lemma 5 that the stopping time of policy $\pi^\star(L,\delta)$ goes to infinity as the error probability vanishes (or as $L \to \infty$). We then exploit this to show that under policy $\pi^\star(L,\delta)$, the modified GLR statistic has the correct drift (see Lemma 6). That is, we build on the result of Proposition 2 and obtain the explicit limit for the modified GLR statistic for the regime of vanishing error probability. We then use the result of Lemma 6 to show in Lemma 7 that the stopping time of policy $\pi^*(L,\delta)$ satisfies an asymptotic almost sure upper bound that matches with the right-hand side of (2.46). Finally, we establish that for any fixed $\delta \in (0,1)$, the family $\{\tau(\pi^\star(L,\delta))/\log L : L \geq 1\}$ is uniformly integrable, and as an intermediate step towards this, we establish in Lemma 8 an exponential upper bound for a certain probability term. Combining the almost sure limit of Lemma 7 along with the uniform integrability result then yields the desired upper bound in (2.46).

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

$$\liminf_{L\to\infty} \tau(\pi^\star(L,\delta)) = \infty \text{ almost surely.} \tag{2.142}$$

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

$$\text{for each } m \geq K, \quad \lim_{L\to\infty} P^\pi(\tau(\pi) \leq m|C) = 0. \tag{2.143}$$

Fix $m \geq K$, and note that

$$\begin{aligned}
&\limsup_{L\to\infty} P^\pi(\tau(\pi) \leq m|C) \\
&= \limsup_{L\to\infty} P^\pi\left(\exists\ K \leq n \leq m \text{ and } \tilde{h} \in \mathcal{A} \text{ such that } M_{\tilde{h}}(n) > \log((K-1)L)\middle| C\right) \\
&\leq \limsup_{L\to\infty} \sum_{\tilde{h}\in\mathcal{A}} \sum_{n=K}^{m} P^\pi(M_{\tilde{h}}(n) > \log((K-1)L)|C) \\
&\leq \limsup_{L\to\infty} \frac{1}{\log((K-1)L)} \sum_{\tilde{h}\in\mathcal{A}} \sum_{n=K}^{m} E^\pi[M_{\tilde{h}}(n)|C],
\end{aligned} \tag{2.144}$$

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

We now show that for each $m \in \{K, \ldots, n\}$, the expectation term inside the summation in (2.144) is finite. Towards this, we have

$$\begin{aligned} M_{\tilde{h}}(n) &= \log \left(\frac{f(A^n, \bar{X}^n | H_{\tilde{h}})}{\max\limits_{h' \neq \tilde{h}} \hat{f}(A^n, \bar{X}^n | H_{h'})} \right) \\ &\leq \log \left(\frac{\hat{f}(A^n, \bar{X}^n | H_{\tilde{h}})}{\hat{f}(A^n, \bar{X}^n | H_{h'})} \right) \text{ for all } h' \neq \tilde{h}. \end{aligned} \tag{2.145}$$

Fix an arbitrary $h' \neq \tilde{h}$. We recognise that the logarithmic term in (2.145) is the classical GLR test statistic of hypothesis $H_{\tilde{h}}$ with respect to hypothesis $H_{h'}$, given by

$$\log \left(\frac{\hat{f}(A^n, \bar{X}^n | H_{\tilde{h}})}{\hat{f}(A^n, \bar{X}^n | H_{h'})} \right) = S_1(n) + S_2(n) + S_3(n) + S_4(n), \tag{2.146}$$

where the terms $S_1(n), \ldots, S_4(n)$ appearing in (2.146) are as below.

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

$$S_1(n) = \sum_{i,j \in \mathcal{S}} N_{\tilde{h}}(n, i, j) \log \frac{N_{\tilde{h}}(n, i, j)}{N_{\tilde{h}}(n, i)}. \tag{2.147}$$

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

$$S_2(n) = \sum_{i,j \in \mathcal{S}} \sum_{a \neq \tilde{h}} N_a(n, i, j) \log \frac{\sum\limits_{a \neq \tilde{h}} N_a(n, i, j)}{\sum\limits_{a \neq \tilde{h}} N_a(n, i)}. \tag{2.148}$$

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

$$S_3(n) = -\sum_{i,j \in \mathcal{S}} N_{h'}(n, i, j) \log \frac{N_{h'}(n, i, j)}{N_{h'}(n, i)}. \tag{2.149}$$

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

$$S_4(n) = -\sum_{i,j \in \mathcal{S}} \sum_{a \neq h'} N_a(n, i, j) \log \frac{\sum\limits_{a \neq h'} N_a(n, i, j)}{\sum\limits_{a \neq h'} N_a(n, i)}. \tag{2.150}$$

We now obtain an almost surely upper bound for (2.146). We recognise that $S_1(n)$ and $S_2(n)$ are non-positive, and thus upper bound each of these terms by zero. Let

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

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

$$\begin{aligned} S_3(n) &= (N_{h'}(n)-1)\sum_{i\in\mathcal{S}}\left[\frac{N_{h'}(n,i)}{N_{h'}(n)-1}\right]H(A(i)) \\ &\leq (N_{h'}(n)-1)\ H\left(\sum_{i\in\mathcal{S}}\left[\frac{N_{h'}(n,i)}{N_{h'}(n)-1}\right]A(i)\right) \\ &\leq N_{h'}(n)\log|\mathcal{S}|, \end{aligned} \tag{2.151}$$

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

$$S_4(n) \leq \left(\sum_{a\neq h'} N_a(n)\right)\log|\mathcal{S}|. \tag{2.152}$$

Using in (2.146) the results of (2.151) and (2.152), along with the zero upper bound for the non-positive terms in (2.147) and (2.148) and the relation (2.5c), we get

$$M_{\tilde{h}}(n) \leq (n+1)\log|\mathcal{S}| \text{ almost surely}, \tag{2.153}$$

from which it follows that

$$\begin{aligned} \limsup_{L\to\infty} P^{\pi}(\tau(\pi)\leq m|C) &\leq \limsup_{L\to\infty}\frac{1}{\log((K-1)L)}\sum_{\tilde{h}\in\mathcal{A}}\sum_{n=K}^{m}(n+1)\log|\mathcal{S}| \\ &= 0. \end{aligned} \tag{2.154}$$

This completes the proof of the lemma. □

Lemma 6. *Fix $\delta\in(0,1)$. Under the policy $\pi=\pi^{\star}(L,\delta)$ and under the arms configuration*

$C = (h, P_1, P_2)$, *for any* $h' \neq h$, *we have*

$$\lim_{L\to\infty} \frac{M_{hh'}(\tau(\pi))}{\tau(\pi)} = D^*_\delta(h, P_1, P_2) \text{ almost surely} \tag{2.155}$$

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

Lemma 7. *Fix* $\delta \in (0, 1)$. *Under the arms configuration* $C = (h, P_1, P_2)$, *the stopping time of the policy* $\pi = \pi^*(L, \delta)$ *satisfies*

$$\limsup_{L\to\infty} \frac{\tau(\pi)}{\log L} \leq \frac{1}{D^*_\delta(h, P_1, P_2)} \quad \text{almost surely.} \tag{2.156}$$

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

1. Case 1: Suppose that arm h is selected at time n. Then, noting that in the expression for $M_{hh'}(n)$, the only terms that depend on the arm index h are those in (2.32) and (2.35), we have

$$M_{hh'}(n) - M_{hh'}(n-1) = \Big[T_2(n) - T_2(n-1)\Big] + \Big[T_5(n) - T_5(n-1)\Big]. \tag{2.157}$$

Suppose that at time n, the Markov process of arm h undergoes a transition from state i to state j, where $i, j \in \mathcal{S}$ are such that $\max\{P_1(j|i), P_2(j|i)\} > 0$[1]. Then, noting that

$$\begin{aligned} N_a(n, i', j') &= N_a(n-1, i', j') \quad \text{for all } a \in \mathcal{A},\ i' \neq i,\ j' \neq j, \\ N_h(n, i, j) &= N_h(n-1, i, j) + 1, \\ N_a(n, i') &= N_a(n-1, i') \quad \text{for all } a \in \mathcal{A},\ i' \neq i, \\ N_h(n, i) &= N_h(n-1, i) + 1, \end{aligned} \tag{2.158}$$

it can be shown after some simplification that

$$\begin{aligned} T_2(n) - T_2(n-1) &= \log \frac{B(N_h(n-1,i,j)+2, (N_h(n-1,i,j')+1)_{j'\neq j})}{B(N_h(n-1,i,j')+1)_{j'\in\mathcal{S}}} \\ &\overset{(a)}{=} \frac{N_h(n-1,i,j)}{\sum_{j'\in\mathcal{S}} N_h(n-1,i,j')} \\ &\leq 1 \quad \text{almost surely,} \end{aligned} \tag{2.159}$$

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

where (a) above follows by using the relation

$$B(\alpha_1, \ldots, \alpha_{|\mathcal{S}|}) = \left(\prod_{k=1}^{|\mathcal{S}|} \Gamma(\alpha_k) \right) \Big/ \Gamma \left(\sum_{k=1}^{|\mathcal{S}|} \alpha_k \right). \tag{2.160}$$

Also, we have

$$\begin{aligned} T_5(n) - T_5(n-1) &= \left(\sum_{a \neq h'} N_a(n-1,i,j) \right) \log \frac{\sum\limits_{a \neq h'} N_a(n-1,i,j)}{\sum\limits_{a \neq h'} N_a(n-1,i)} \\ &\quad - \left(1 + \sum_{a \neq h'} N_a(n-1,i,j) \right) \log \frac{1 + \sum\limits_{a \neq h'} N_a(n-1,i,j)}{1 + \sum\limits_{a \neq h'} N_a(n-1,i)} \\ &\leq \log \frac{\sum\limits_{a \neq h'} N_a(n-1,i)}{\sum\limits_{a \neq h'} N_a(n,i,j)} \\ &\to \log \frac{1}{P_\delta(j|i)} \quad \text{almost surely,} \end{aligned} \tag{2.161}$$

where the convergence in the last line follows from (2.133). Thus, it follows that the increment $M_{hh'}(n) - M_{hh'}(n-1)$ is bounded for all $n \geq 1$.

2. Case 2: Suppose that arm h' is sampled at time n. Noting that the only terms that depend on the arm index h' are those in (2.33) and (2.34), the analysis for this case proceeds on the exactly same lines as that of Case 1 presented above, and is omitted.

3. Case 3: Suppose that arm a' is sampled at time n, where $a' \in \mathcal{A} \setminus \{h, h'\}$. Noting that the only terms that depend on the arm index a' are those in (2.33) and (2.35), the analysis for this case proceeds on the exactly same lines as that of Case 1 presented above, and is omitted.

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

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

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

$$= 1 \quad \text{almost surely,} \tag{2.162}$$

where (a) above is due to boundedness of the increments of the modified GLR process established above. Then, using Lemma 6 along with the relation (2.162) yields

$$\begin{aligned}
\limsup_{L\to\infty} \frac{\tau(\pi)}{\log L} &= \limsup_{L\to\infty} \left\{ \left(\frac{\tau(\pi)}{M_{hh'}(\tau(\pi))} \right) \left(\frac{M_{hh'}(\tau(\pi))}{\log L} \right) \right\} \\
&= \left(\lim_{L\to\infty} \frac{\tau(\pi)}{M_{hh'}(\tau(\pi))} \right) \left(\limsup_{L\to\infty} \frac{M_{hh'}(\tau(\pi))}{\log L} \right) \\
&\leq \frac{1}{D^*_\delta(h, P_1, P_2)} \quad \text{almost surely,}
\end{aligned} \tag{2.163}$$

thus completing the proof of the lemma. □

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

$$\limsup_{L\to\infty} E^\pi \left[\exp\left(\frac{\tau(\pi)}{\log L} \right) \middle| C \right] < \infty. \tag{2.164}$$

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

$$l(L,\delta) \coloneqq \frac{3\log((K-1)L)}{\frac{\delta}{2K}\Big(D(P_1||P_\delta|\mu_1) + D(P_2||P_\delta|\mu_2) \Big)}. \tag{2.165}$$

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

$$u(L) \coloneqq \exp\left(\frac{1 + l(L,\delta)}{\log L} \right) \tag{2.166}$$

Clearly, we have $\tau(\pi^\star_h) \geq \tau(\pi)$ almost surely. Then,

$$\begin{aligned}
&\limsup_{L\to\infty} E^\pi \left[\exp\left(\frac{\tau(\pi)}{\log L} \right) \middle| C \right] = \limsup_{L\to\infty} \int_0^\infty P^\pi \left(\frac{\tau(\pi)}{\log L} > \log x \middle| C \right) dx \\
&\leq \limsup_{L\to\infty} \int_0^\infty P^\pi \Big(\tau(\pi^\star_h) \geq \lceil (\log x)(\log L) \rceil \Big| C \Big) \, dx
\end{aligned}$$

$$\overset{(a)}{\leq} \limsup_{L\to\infty} \left\{ u(L) + \int_{u(L)}^{\infty} P^{\pi}\left(\tau(\pi_h^{\star}) \geq \lceil (\log x)(\log L)\rceil \middle| C\right) dx \right\}$$

$$\leq \exp\left(\frac{3}{\frac{\delta}{2K}(D(P_1||P_\delta|\mu_1) + D(P_2||P_\delta|\mu_2))}\right) + \limsup_{L\to\infty} \sum_{n\geq l(L,\delta)} \exp\left(\frac{n+1}{\log L}\right) P^{\pi}(M_h(n) < \log((K-1)L)|C), \tag{2.167}$$

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

We now show that for all $n \geq l(L,\delta)$, the probability term in (2.167) decays exponentially in n. This is a strengthening of the result in Proposition 2 which only establishes that when $C = (h, P_1, P_2)$ is the underlying configuration of the arms, $M_h(n) \to \infty$ as $n \to \infty$.

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

$$P^{\pi}(M_h(n) < \log((K-1)L)|C) \leq Be^{-\theta n}. \tag{2.168}$$

Proof. Since

$$P^{\pi}(M_h(n) < \log((K-1)L)|C) = P^{\pi}\left(\min_{h'\neq h} M_{hh'}(n) < \log((K-1)L)\middle| C\right) \leq \sum_{h'\neq h} P^{\pi}\left(M_{hh'}(n) < \log((K-1)L)\middle| C\right), \tag{2.169}$$

in order to prove the lemma, it suffices to show that each term inside the summation in (2.169) is exponentially bounded. Going further, we drop the superscript π and the conditioning on configuration C in $P^{\pi}(\cdot|C)$ for ease of notation. For all $i, j \in \mathcal{S}$, let

$$\tilde{P}_n(j|i) := \frac{\alpha_n \mu_1(i) P_1(j|i) + \beta_n \mu_2(i) P_2(j|i)}{\alpha_n \mu_1(i) + \beta_n \mu_2(i)}, \tag{2.170}$$

where α_n and β_n are as in (2.137). Fix $h' \neq h$ and $\epsilon > 0$ arbitrarily. Then, using (2.30) and triangle inequality, we have

$$P(M_{hh'}(n) < \log((K-1)L)) \leq U_1 + U_2 + U_3 + U_4 + U_5 + U_6 + U_7, \tag{2.171}$$

where the terms $U_1, \ldots, U_7$ in (2.171) are as below.

1. The term U_1 is given by

$$U_1 = P\left(\frac{T_1(n)}{n} < -\epsilon\right), \tag{2.172}$$

where T_1 is given by (2.31).

2. The term U_2 is given by

$$U_2 = P\left(\frac{T_2(n)}{n} - \frac{N_h(n)}{n}\sum_{i\in\mathcal{S}}\mu_1(i)(-H(P_1(\cdot|i))) < -\epsilon\right), \tag{2.173}$$

where $T_2(n)$ is given by (2.32).

3. The term U_3 is given by

$$U_3 = P\left(\frac{T_3(n)}{n} - \frac{\sum_{a\neq h} N_a(n)}{n}\sum_{i\in\mathcal{S}}\mu_2(i)(-H(P_2(\cdot|i))) < -\epsilon\right), \tag{2.174}$$

where $T_3(n)$ is given by (2.33).

4. The term U_4 is given by

$$U_4 = P\left(\frac{T_4(n)}{n} - \frac{N_{h'}(n)}{n}\sum_{i\in\mathcal{S}}\mu_2(i)H(P_2(\cdot|i)) < -\epsilon\right), \tag{2.175}$$

where $T_4(n)$ is given by (2.34).

5. The term U_5 is given by

$$U_5 = P\left(\frac{T_5(n)}{n} - \sum_{i\in\mathcal{S}}(\alpha_n\mu_1(i) + \beta_n\mu_2(i))H(\tilde{P}_n(\cdot|i)) < -\epsilon\right), \tag{2.176}$$

where $T_5(n)$ is given by (2.35).

6. The term U_6 is given by

$$U_6 = P\Big(\alpha_n\Big[D(P_1||\tilde{P}_n|\mu_1) - D(P_1||P_\delta|\mu_1)\Big] + \beta_n\Big[D(P_2||\tilde{P}_n|\mu_2) - D(P_2||P_\delta|\mu_2)\Big] < -\epsilon\Big), \tag{2.177}$$

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

7. The term U_7 is given by

$$U_7 = P\bigg(\alpha_n D(P_1||P_\delta|\mu_1) + \beta_n D(P_2||P_\delta|\mu_2) - 6\epsilon < \frac{\log((K-1)L)}{n}\bigg). \tag{2.178}$$

In (2.173), the term $H(P_1(\cdot|i))$ refers to the Shannon entropy of the probability distribution $(P_1(j|i))_{j\in\mathcal{S}}$ on set $\mathcal{S}$; the terms $H(P_2(\cdot|i))$ and $H(\tilde{P}_n(\cdot|i))$ are defined similarly.

We now obtain a bound for the terms in (2.172)-(2.178).

1. We begin by showing an exponential upper bound for (2.178). We choose $0 < \epsilon' < \frac{2}{3}$, and then select $\epsilon > 0$ such that the following holds:

$$\frac{\delta}{2K}(1-\epsilon')\bigg(D(P_1||P_\delta|\mu_1) + D(P_2||P_\delta|\mu_2)\bigg) - 6\epsilon > \frac{1}{3}\cdot\frac{\delta}{2K}\bigg(D(P_1||P_\delta|\mu_1) + D(P_2||P_\delta|\mu_2)\bigg). \tag{2.179}$$

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

$$P\bigg(\alpha_n D(P_1||P_\delta|\mu_1) + \beta_n D(P_2||P_\delta|\mu_2) - 6\epsilon < \frac{\log((K-1)L)}{n}, \\ \frac{N_a(n)}{n} > \frac{\delta}{2K}(1-\epsilon') \text{ for all } a \in \mathcal{A}\bigg) = 0. \tag{2.180}$$

Writing the probability term in (2.178) as a sum of the probability term in (2.180) and a second probability term given by

$$P\bigg(\alpha_n D(P_1||P_\delta|\mu_1) + \beta_n D(P_2||P_\delta|\mu_2) - 6\epsilon < \frac{\log((K-1)L)}{n}, \\ \frac{N_a(n)}{n} \leq \frac{\delta}{2K}(1-\epsilon') \text{ for some } a \in \mathcal{A}\bigg), \tag{2.181}$$

and upper bounding (2.181) by $P(N_a(n)/n \leq (\delta/2K)(1-\epsilon')$ for some $a \in \mathcal{A})$, an application of the union bound yields

$$P\bigg(\alpha_n D(P_1||P_\delta|\mu_1) + \beta_n D(P_2||P_\delta|\mu_2) - 6\epsilon < \frac{\log((K-1)L)}{n}\bigg) \\ \leq \sum_{a=1}^{K} P\left(\frac{N_a(n)}{n} \leq \frac{\delta}{2K}(1-\epsilon')\right). \tag{2.182}$$

Noting that for each $a \in \mathcal{A}$, the sequence $\left(N_a(n) - n\frac{\delta}{2K}\right)_{n\geq 0}$ is a submartingale, with the

absolute value of the difference between any two successive terms of the submartingale sequence being of value at most 1, we use the Azuma-Hoeffding inequality to obtain

$$\begin{aligned} P\left(\frac{N_a(n)}{n} \leq \frac{\delta}{2K}(1-\epsilon')\right) &= P\left(N_a(n) - n\frac{\delta}{2K} \leq -n\epsilon'\frac{\delta}{2K}\right) \\ &= P\left(\left[N_a(n) - n\frac{\delta}{2K}\right] - N_a(0) \leq -n\epsilon'\frac{\delta}{2K} - N_a(0)\right) \\ &\leq P\left(\left[N_a(n) - n\frac{\delta}{2K}\right] - N_a(0) \leq -n\epsilon'\frac{\delta}{2K}\right) \\ &\leq \exp\left(-\frac{n(\epsilon')^2\delta^2}{8K^2}\right). \end{aligned} \tag{2.183}$$

Plugging (2.183) back in (2.182), we arrive at

$$P\left(\alpha_n D(P_1||P_\delta|\mu_1) + \beta_n D(P_2||P_\delta|\mu_2) - 6\epsilon < \frac{\log((K-1)L)}{n}\right) \leq K\exp\left(-\frac{n(\epsilon')^2\delta^2}{8K^2}\right). \tag{2.184}$$

2. We now turn attention to (2.175), which we upper bound as follows:

$$\begin{aligned} &P\left(\frac{T_4(n)}{n} - \frac{N_{h'}(n)}{n}\sum_{i\in\mathcal{S}}\mu_2(i)H(P_2(\cdot|i)) < -\epsilon\right) \\ &= P\left(\frac{N_{h'}(n)}{n}\left\{\sum_{i\in\mathcal{S}}\frac{N_{h'}(n,i)}{N_{h'}(n)}H\left(\frac{N_{h'}(n,i,\cdot)}{N_{h'}(n,i)}\right) - \mu_2(i)H(P_2(\cdot|i))\right\} < -\epsilon\right) \\ &\leq P\left(\frac{N_{h'}(n)}{n}\left\{\sum_{i\in\mathcal{S}}\frac{N_{h'}(n,i)}{N_{h'}(n)}H\left(\frac{N_{h'}(n,i,\cdot)}{N_{h'}(n,i)}\right) - \sum_{i\in\mathcal{S}}\mu_2(i)H(P_2(\cdot|i))\right\} < -\epsilon,\right. \\ &\qquad\qquad \left.\frac{N_a(n)}{n} > \frac{\delta}{2K}(1-\epsilon') \text{ for all } a\in\mathcal{A}\right) \\ &+ \sum_{a=1}^{K} P\left(\frac{N_a(n)}{n} \leq \frac{\delta}{2K}(1-\epsilon')\right). \end{aligned} \tag{2.185}$$

From the analysis using the Azuma-Hoeffding inequality for bounded difference submartingales presented earlier, we know that each term inside the summation in (2.185) is exponentially bounded. The first term in (2.185) may be written as

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

$$\frac{N_a(n)}{n} > \frac{\delta}{2K}(1-\epsilon') \text{ for all } a \in \mathcal{A}\Bigg)$$
$$\leq P\Bigg(\Bigg\{\sum_{i\in\mathcal{S}} \frac{N_{h'}(n,i)}{N_{h'}(n)} H\left(\frac{N_{h'}(n,i,\cdot)}{N_{h'}(n,i)}\right) - \sum_{i\in\mathcal{S}} \mu_2(i) H(P_2(\cdot|i))\Bigg\} < -\epsilon,$$
$$\frac{N_a(n)}{n} > \frac{\delta}{2K}(1-\epsilon') \text{ for all } a \in \mathcal{A}\Bigg). \tag{2.186}$$

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

$$\frac{N_{h'}(n,i,j)}{N_{h'}(n,i)} \to P_2(j|i), \text{ for all } i,j \in \mathcal{S},$$
$$\frac{N_{h'}(n,i)}{N_{h'}(n)} \to \mu_2(i), \text{ for all } i \in \mathcal{S}. \tag{2.187}$$

Using the above convergences and the continuity of the Shannon entropy functional $H(\cdot)$, we get that there exist constants $\delta_1 = \delta_1(\epsilon)$ and $\delta_2 = \delta_2(\epsilon)$ such that the probability in (2.186) may be upper bounded by the probability

$$P\Bigg(\exists\, i,j \in \mathcal{S} \text{ such that } \left|\frac{N_{h'}(n,i,j)}{N_{h'}(n,i)} - P_2(j|i)\right| > \delta_1, \left|\frac{N_{h'}(n,i)}{N_{h'}(n)} - \mu_2(i)\right| > \delta_2,$$
$$\frac{N_a(n)}{n} > \frac{\delta}{2K}(1-\epsilon') \text{ for all } a \in \mathcal{A}\Bigg). \tag{2.188}$$

Noting that $(N_{h'}(n,i,j) - N_{h'}(n,i)P_2(j|i))_{n\geq 0}$ and $(N_{h'}(n,i) - N_{h'}(n)\mu_2(j|i))_{n\geq 0}$ are martingale sequences for all $i,j \in \mathcal{S}$, we may then express (2.188) as a probability of deviation of martingale sequences from zero, which may be exponentially bounded by using results from [23, Theorem 1.2A].

3. We now upper bound the term in (2.173). Towards this, we first pick $\epsilon_1 > 0$ satisfying

$$0 < \epsilon_1 \leq \frac{\epsilon}{1 + 2\sum_{i\in\mathcal{S}} \mu_1(i) H(P_1(\cdot|i))}. \tag{2.189}$$

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

$$\frac{N_h(n)}{n} \to \lambda_\delta^*,$$
$$\frac{N_h(n,i,j)}{N_h(n)} \to \mu_1(i)P_1(j|i). \tag{2.190}$$

Following the steps leading up to (2.106), we note that for every choice of $\epsilon' > 0$, there

exists $M = M(\epsilon')$ such that (2.106) holds. We now choose ϵ' such that

$$\begin{aligned}\frac{T_2(n)}{n} &\geq \frac{N_h(n)}{n}\left\{\left[\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}}(\mu_1(i)P_1(j|i)+\epsilon')\log\frac{\mu_1(i)P_1(j|i)+\epsilon'}{\mu_1(i)+\epsilon'|\mathcal{S}|}\right]-\epsilon'\right\}\\ &\geq \frac{N_h(n)}{n}\left(\sum_{i\in\mathcal{S}}\mu_1(i)(-H(P_1(\cdot|i)))\right)-\epsilon_1 \qquad (2.191)\end{aligned}$$

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

$$\begin{aligned}&P\left(\frac{T_2(n)}{n}-\frac{N_h(n)}{n}\sum_{i\in\mathcal{S}}\mu_1(i)(-H(P_1(\cdot|i)))<-\epsilon\right)\\ &\leq P\Bigg(\frac{T_2(n)}{n}-\frac{N_h(n)}{n}\sum_{i\in\mathcal{S}}\mu_1(i)(-H(P_1(\cdot|i)))<-\epsilon,\left|\frac{N_h(n)}{n}-\lambda_\delta^*\right|\leq\epsilon_1,\\ &\qquad\qquad\left|\frac{N_h(n,i,j)}{N_h(n)}-\mu_1(i)P_1(j|i)\right|\leq\epsilon' \text{ for all } i,j\in\mathcal{S}\Bigg)\\ &+P\left(\left|\frac{N_h(n)}{n}-\lambda_\delta^*\right|>\epsilon_1\right)+\sum_{i,j\in\mathcal{S}}P\left(\left|\frac{N_h(n,i,j)}{N_h(n)}-\mu_1(i)P_1(j|i)\right|>\epsilon'\right). \qquad (2.192)\end{aligned}$$

We now focus on the first term in (2.192), and notice that for all sufficiently large values of n, this term may be upper bounded as

$$\begin{aligned}&P\left((\lambda_\delta^*+\epsilon_1)\sum_{i\in\mathcal{S}}\mu_1(i)(-H(P_1(\cdot|i)))-\epsilon_1<-\epsilon+(\lambda_\delta^*-\epsilon_1)\sum_{i\in\mathcal{S}}\mu_1(i)(-H(P_1(\cdot|i)))\right)\\ &\leq P\left(\epsilon_1>\frac{\epsilon}{1+2\sum_{i\in\mathcal{S}}\mu_1(i)H(P_1(\cdot|i))}\right)\\ &=0, \qquad (2.193)\end{aligned}$$

where the last line follows from the choice of ϵ_1 in (2.189). Exponential bounds for the remaining terms in (2.192) can be obtained similarly as in the analysis of the first term in (2.185).

Lastly, for the terms in (2.172), (2.174), (2.176) and (2.177), noting that the left-hand sides of the inequality inside the probability expression in all the three terms converge to zero almost surely, similar procedures as used above for (2.173) and (2.175) may be used to obtain exponential upper bounds.

This completes the proof of the lemma. □

Using the result of Lemma 8 in (2.167), we get that there exist constants $\theta > 0$ and $0 < B < \infty$ independent of L such that the following holds:

$$\begin{aligned}
&\limsup_{L\to\infty} E^{\pi}\left[\exp\left(\frac{\tau(\pi)}{\log L}\right)\middle| C\right] \\
&\leq \exp\left(\frac{3}{\frac{\delta}{2K}(D(P_1||P_\delta|\mu_1)+D(P_2||P_\delta|\mu_2))}\right) + \limsup_{L\to\infty}\sum_{n\geq l(L,\delta)} B\exp\left(\frac{n+1}{\log L}-n\theta\right) \\
&< \infty, \qquad (2.194)
\end{aligned}$$

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

Combining the above result on uniform integrability along with the asymptotic bound in (2.156) yields the desired upper bound in (2.46). □

2.8 Summary

We analysed the asymptotic behaviour of policies for the problem of odd arm identification in a multi-armed bandit setting with rested arms. The asymptotics is in the regime of vanishing error probabilities. We focused on the particular case when the transition probability matrix of neither the odd arm nor the non-odd arm Markov processes is known beforehand. We derived an asymptotic lower bound on the growth rate of the expected stopping time of any policy as a function of error probability. We identified an explicit configuration-dependent constant in the lower bound. Furthermore, we proposed a scheme that (a) is a modification of the classical GLRT, and (b) uses an idea of "forced exploration" from [10]. This scheme takes as inputs two parameters: $L > 1$ and $\delta \in (0,1)$. We showed that (a) for a suitable choice of L, the probability of error of our scheme can be controlled to any desired tolerance level, and (b) by tuning δ, the performance of our scheme can be made arbitrarily close to the lower bound for vanishingly small error probabilities. In proving the above results, we highlighted how to overcome some of the key challenges that the Markov setting offers in the analysis. Our analysis of the rested Markov setting is a key first step in understanding the difficult case of restless arms.

Chapter 3

Restless Arms with Known TPMs

3.1 Preamble

In the previous chapter, we analysed the setting when the arms are rested, i.e., the unobserved arms remain frozen and do not evolve. In this chapter, we analyse the more difficult setting of restless arms in which the unobserved arms continue to evolve. To begin with, we consider the simpler case when the TPMs are known beforehand. All the essential conceptual difficulties related to the setting of restless arms remain despite this simplification. Formally, given a multi-armed bandit with $K \geq 3$ arms and an arms configuration $C = (h, P_1, P_2)$ in which h is the index of the odd arm, the TPM of arm h is P_1, and the TPM of each of the remaining arms is $P_2 \neq P_1$, we wish to characterise

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

for the setting of restless arms when P_1 and P_2 are known beforehand.

As we shall see later in the chapter, the continued evolution of the Markov process of each arm makes it necessary to keep a record of (a) the time elapsed since each arm was previously selected (called the arm's *delay*), and (b) the state of each arm as observed at its previous selection time (called the *last observed state* of the arm). The notion of arm delays is superfluous when the arms are rested because the unobserved arms remain frozen at their previously observed states. It is superfluous also in the special case of the restless setting when each arm yields iid observations as in the prior works [4, 5, 6, 8] because at any given time, the current state of an arm is independent of its last observed state. Therefore, the notions of arm delays and last observed states are strikingly new features of the setting of restless arms.

3.1.1 Motivation and the Notion of a Trembling Hand

As alluded to in the previous chapters, our motivation to study the problem of odd arm identification in the setting of restless arms comes from the desire to extend the analysis of Vaidhiyan et al. in [4, 5, 6] to more general settings. It is often the case in such visual search experiments that though the subject intends to focus his/her attention at a certain location, the actual focus location differs from the intended focus location with a small probability. We model this in our multi-armed bandit setting as a *trembling hand* for the decision entity: with probability $1-\eta$, the decision entity samples the intended arm, but with probability η, the decision entity samples a uniformly randomly chosen arm. Up to Section 3.6, we assume that $\eta > 0$, as is often the case in visual search experiments such as that described above. The case when $\eta = 0$ is dealt with separately in Section 3.7.

Our assumption about the uniform sampling of the arms under the trembling hand model is merely for convenience, and any probability distribution on the arms that puts a strictly positive mass on each of the arms may be used in place of the uniform distribution. The values of all the expectations and probabilities (in particular, the lower bound of Section 3.4), which rely on the uniform sampling assumption, will accordingly differ.

For a related example in the cognitive radio setting (no trembling) in which the number of anomalous arms may be more than one, see [26].

3.1.2 Prior Works on Restless Arms

The topic of restless arms has been studied extensively in the literature in the context of reward maximisation (or equivalently, regret minimisation). In such works, each arm is assumed to yield, upon being sampled, an immediate 'reward' based on the arm's current state. Regret is then defined as the difference between the expected sum of rewards obtained under a particular arm selection scheme and that obtained by a scheme that knows which arm yields the highest expected reward. Whittle [27] refined and extended the results of Gittins [18] on the optimality, in the setting of rested arms, of a certain index-based policy. Whittle [27] demonstrated that Gittins's policy in [18] is not necessarily optimal in the context of restless arms, introduced a new index (now called *Whittle's index*) which could be computed if each arm satisfied an *indexability* condition, and demonstrated that the new index coincides with Gittins's index in the rested setting. Yet, as Whittle showed, the new index-based policy is not necessarily optimal for the general setting of restless arms.

Whittle's results require the Markov transition laws of each of the arms to be known beforehand. Extensions of Whittle's results to the case when the laws are not known beforehand

appear in Liu et al. [28]. Ortner et al. [29] provide a policy that, when the transition laws of the arms are unknown, gives a regret of the order $O(\sqrt{T})$ after T time steps in relation to a policy that knows the Markov transition laws of all the arms. As Ortner et al. show in [29], an optimal policy for the restless bandit problem does not necessarily pick the arm with the largest stationary mean at each time instant[1], but instead switches between the arms in an optimal fashion. Working on this key idea, Grünelwalder et al. [31] provide conditions under which the problem of finding the arm with the largest stationary mean serves as a "good" approximation to the original problem of finding the optimal arm switching strategy when each arm is a stationary ϕ-mixing process and the arms are restless. The works [29] and [31] deal with general state spaces (i.e., not necessarily finite or countable) and address the associated technical challenges.

While the above mentioned works focus on maximising rewards (or equivalently, minimising regret), our focus is on the stopping problem of identifying the index of the odd arm as quickly as possible. For a related problem of best arm identification instead of odd arm identification, see [20, 32]. The recent works [33], [34] and [35] deal with more general problems of sequential hypothesis testing in multi-armed bandits, special cases of which are the problems of best arm identification and odd arm identification, in the context of iid observations from each arm. In contrast to these works, we study in this chapter the specific problem of odd arm identification in the more difficult setting of restless arms.

3.1.3 A Brief Overview of Our Contributions

Below, we highlight our contributions and bring out the challenges that we need to overcome in the analysis of the setting of restless arms when the TPMs of the arms are assumed to be known beforehand.

1. We show that given a pre-specified error probability threshold $\epsilon > 0$, the expected time taken by the decision maker to identify the index of the odd arm with probability of error at most ϵ grows as $\Theta(\log(1/\epsilon))$. We give a precise characterisation of the best (smallest) constant multiplying $\log(1/\epsilon)$, which we call $R^*(P_1, P_2)$, in terms of the Markov transition probability matrices P_1 and P_2. This is the first known characterisation of this constant for the setting of restless Markov arms. See Section 3.4 for an exact mathematical expression. We prove this by first showing a lower bound in Section 3.4 and then a matching asymptotic upper bound in Section 3.5.

[1] This is indeed the case in a multi-armed bandit problem with iid observations from each arm, as was shown in [30].

2. An examination of the lower bounds in the prior works [4, 5, 6, 36] reveals that the best constant multiplier in these works is the solution to an optimisation problem having an outer supremum over all (unconditional) probability distributions on the arms, followed by an inner minimum over all alternative odd arm locations (i.e., a sup-min optimisation problem). A further examination reveals that when arm h is the odd arm, there exists a probability distribution λ_h^* on the arms, possibly depending on the odd arm location h, that (a) attains the outer supremum, and (b) puts equal mass on each of the non-odd arm locations.

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

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

3. In order to derive the constant $R^*(P_1, P_2)$, we use the fact that the arm delays and the last observed states form a *controlled Markov process*, with the arm selections playing the role of *controls*. This approach of ours takes into account the delays and the last observed states of *all* the arms jointly. In contrast, the approaches of [4, 5, 6, 36] suggest dealing with the delays and the last observed states of each of the arms separately, which we view as a 'local' perspective of the arm delays and the last observed states. In Section 3.9.1, we show that this local perspective of arm delays and last observed states leads to an infinite dimensional, constrained, linear programming problem (LPP). The drawback of this approach is that it is not easy to find the tightest set of constraints for the LPP. As a consequence, the constant multiplier obtained as the solution to the LPP may not necessarily be the best (smallest).

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

4. We show that under a *stationary* arm selection policy (in which at each time, the arms are selected according to a certain conditional probability distribution on the arms, conditioned on the delays and last observed states at that time), the aforementioned controlled Markov process is, in fact, a Markov process. Additionally, we show that under every stationary arm selection policy, this Markov process is *ergodic* when the trembling hand parameter $\eta > 0$ (Lemma 9). It is this ergodicity property, together with the strict positivity of the trembling hand parameter η, that plays a crucial role in our analysis of the lower and the upper bounds. The case $\eta = 0$ demands a careful examination since, in this case, such an ergodicity property is not readily available for every stationary arm selection policy.

5. We show that for every arm selection policy of the decision maker, stationary or otherwise, what enters into the analyses of the lower and the upper bounds is the following statistic: for each possible value of arm delays $\underline{d}$, last observed states $\underline{i}$ and arm a, the long-term fraction of times the aforementioned controlled Markov process visits the state $(\underline{d}, \underline{i})$ and arm a is selected. This fact, together with Theorem 3, enables us to restrict attention only to stationary arm selection policies in arriving at the best constant multiplier $R^*(P_1, P_2)$.

   In spite of the above simplification, the computability of $R^*(P_1, P_2)$ remains an issue since it involves a search over the space of all stationary arm selection policies. One must resort to Q-learning in the context of restless Markov arms (see, for instance, [37]) to compute $R^*(P_1, P_2)$. Under some circumstances, good approximations to $R^*(P_1, P_2)$ may be possible; see Section 3.10.

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

   Notwithstanding the additional technical difficulty encountered in the setting of restless arms due to the presence of the countably infinite-valued arm delays, we show that the

supremum in the expression for $R^*(P_1, P_2)$ may be approached arbitrarily closely by stitching together certain parameterised solutions to the MDPs mentioned in point 3 above. We present the details in Section 3.5.

7. The trembling hand model (with $\eta > 0$) may be viewed as a regularisation that ensures stability of the aforementioned controlled Markov process (of arm delays and last observed states) for free. If $\eta = 0$, one could deliberately add some regularisation parameterised by η, re-label the constant $R^*(P_1, P_2)$ in this case as $R^*_\eta(P_1, P_2)$ for each $\eta > 0$, and analyse the limiting value of $R^*_\eta(P_1, P_2)$ as $\eta \downarrow 0$. We show that in this case, (a) the limit of $R^*_\eta(P_1, P_2)$ as $\eta \downarrow 0$ exists, and (b) the upper bound is governed by $\lim_{\eta \downarrow 0} R^*_\eta(P_1, P_2)$, while the lower bound is governed by $R^*_0(P_1, P_2)$ (which is obtained by plugging $\eta = 0$ in the expression for $R^*_\eta(P_1, P_2)$). So, the question then is, do these lower and the upper bounds match? In Section 3.7, we are only able to establish that $\lim_{\eta \downarrow 0} R^*_\eta(P_1, P_2) \leq R^*_0(P_1, P_2)$. A key tool needed to establish equality in this inequality is the "envelope theorem" [38, Theorem 2]. A verification of the hypotheses of the envelope theorem for the setting of restless arms still remains open.

8. We verify that the envelope theorem holds in the iid and rested Markov settings of the prior works [4, 5, 6, 36], thus leading to matching upper and lower bounds in these works. Thus, sufficient conditions for the upper and the lower bounds to match are either (a) $\eta > 0$, or (b) $\eta = 0$ and the observations come from either iid or rested Markov arms.

3.1.4 Chapter Organisation

The rest of this chapter is organised as follows. In Section 3.2, we set up the notations and provide some preliminaries that will be useful for the rest of the chapter. In Section 3.4, we present the lower bound on the growth rate of the expected time to find the odd arm as a function of the error probability. In the same section, we also show that by following the conventional approaches available in the prior works, we arrive at an infinite-dimensional linear programming problem (LPP) with countably infinitely many constraints that is difficult to solve. In Section 3.5, we present a sequence of strategies whose expected times to find the index of the odd arm approach the lower bound in the limit of vanishing error probabilities, following which we state the main result of this chapter in Section 3.6. We discuss the no trembling hand case in Section 3.7. The proofs of all the results are contained in Section 3.8. Section 3.9.2 contains the statement of an important theorem that is used in several places in the main body of the chapter.

3.2 Notations and Preliminaries

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

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

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

We note here that the observations $\{\bar{X}_t : t \geq 0\}$ are noiseless. The case of noisy observations, e.g., hidden Markov models, is important and is left for future work.

3.2.1 Policy

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

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

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

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

$$\Pi(\epsilon) := \{\pi : P_h(\theta(\pi) \neq h) \leq \epsilon \text{ for all } h \in \mathcal{A}\} \tag{3.1}$$

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

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

Our interest is in characterising the constant factor multiplying $\log(1/\epsilon)$ in the limit as $\epsilon \downarrow 0$. For simplicity, we assume that every policy starts with the observation that arm 1 is observed at time $t = 0$, arm 2 is observed at time $t = 1$, etc., and arm K is observed at time $t = K - 1$. This can be effected by sampling the arms uniformly until this event occurs. Clearly, for $\eta > 0$, this requirement will result in a finite delay almost surely which does not affect the asymptotic analysis as $\epsilon \downarrow 0$.

3.2.2 Delays and Last Observed States

Recall that at each time $t \in \{0, 1, \ldots\}$, the decision maker observes only one of the arms, while the unobserved arms continue to undergo state evolution. Therefore, the probability of the observation $\bar{X}_t$ on the selected arm A_t is a function of (a) the time elapsed since the previous time instant of selection of arm A_t (called the *delay* of arm A_t), and (b) the state of arm A_t at its previous selection time instant (called the *last observed state* of arm A_t). Notice that when

the arms are *rested*, the notion of arm delays is superfluous since each arm remains frozen at its previously observed state until its next selection time instant. Also, the notion of arm delays is redundant in the setting of iid observations since, in this special case, the current state of the arm selected is independent of the state at its previous selection. Thus, the notion of arm delays is a key distinguishing feature of the setting of restless arms.

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

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

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

$$d_a(t+1) = \begin{cases} d_a(t)+1, & a \neq a', \\ 1, & a = a', \end{cases} \qquad i_a(t+1) = \begin{cases} i_a(t), & a \neq a', \\ \bar{X}_t, & a = a', \end{cases} \tag{3.2}$$

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

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

3.2.3 Controlled Markov Process and the Resulting Markov Decision Problem

From the update rule in (3.2), it is clear that the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ takes values in a subset $\mathbb{S}$ of the countable set $\mathbb{N}^K \times \mathcal{S}^K$, where $\mathbb{N} = \{1, 2, \ldots\}$ denotes the set of natural numbers. The subset $\mathbb{S}$ is formed based on the constraint that at 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. Note that for

all $(\underline{d}, \underline{i}) \in \mathbb{S}$ and $t \geq K$,

$$\begin{aligned} &P(\underline{d}(t+1) = \underline{d}, \underline{i}(t+1) = \underline{i} \mid (\underline{d}(s), \underline{i}(s)), B_s,\ K \leq s \leq t) \\ &= P(\underline{d}(t+1) = \underline{d}, \underline{i}(t+1) = \underline{i} \mid (\underline{d}(t), \underline{i}(t)), B_t). \end{aligned} \tag{3.3}$$

On account of (3.3) being satisfied, we say that under any policy π, the evolution of the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is *controlled* by the sequence $\{B_t\}_{t\geq 0}$ of intended arm selections under policy π. Alternatively, we say that $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is a controlled Markov process, with $\{B_t\}_{t\geq 0}$ as the sequence of controls; the terminology used here follows that of Borkar [39]. Thus, we are in a Markov decision problem (MDP) setting. We now make precise the state space, the action space, the transition probabilities and our objective.

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

1. the trembling hand rule

$$P(A_t = a | B_t) = \frac{\eta}{K} + (1-\eta)\ \mathbb{I}_{\{B_t = a\}}, \quad \forall a \in \mathcal{A}, \tag{3.4}$$

2. the law associated with arm A_t, and

3. the update rule (3.2).

In (3.4), $\mathbb{I}$ denotes the indicator function. In order to write the transition probabilities of the MDP precisely, let us introduce some notations. Given $h, a \in \mathcal{A}$, let P_h^a denote the transition probability matrix of the Markov process of arm a under the hypothesis $\mathcal{H}_h$. That is,

$$P_h^a = \begin{cases} P_1, & a = h, \\ P_2, & a \neq h. \end{cases} \tag{3.5}$$

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

$$P(\underline{d}(t+1) = \underline{d}', \underline{i}(t+1) = \underline{i}' \mid \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, B_t = b)$$

$$= \begin{cases} \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b\}}(a)\right)\,(P_h^a)^{d_a}(i'_a|i_a), & \text{if } d'_a = 1 \text{ and } d'_{\tilde{a}} = d_{\tilde{a}} + 1 \text{ for all } \tilde{a} \neq a, \\ & i'_{\tilde{a}} = i_{\tilde{a}} \text{ for all } \tilde{a} \neq a, \\ 0, & \text{otherwise}, \end{cases} \tag{3.6}$$

where d'_a and i'_a in (3.6) denote the component corresponding to arm a in $\underline{d}'$ and $\underline{i}'$ respectively. Note that the transition probabilities defined in (3.6) are stationary and independent of time. Also, for $a \in \mathcal{A}$, we have

$$P(\underline{d}(t+1) = \underline{d}', \underline{i}(t+1) = \underline{i}' \mid \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, A_t = a)$$
$$= \begin{cases} (P_h^a)^{d_a}(i'_a|i_a), & \text{if } d'_a = 1 \text{ and } d'_{\tilde{a}} = d_{\tilde{a}} + 1 \text{ for all } \tilde{a} \neq a, \\ & i'_{\tilde{a}} = i_{\tilde{a}} \text{ for all } \tilde{a} \neq a, \\ 0, & \text{otherwise}. \end{cases} \tag{3.7}$$

The left-hand sides of (3.6) and (3.7) differ in that B_t in (3.6) is replaced by A_t in (3.7). We shall write $Q(\underline{d}', \underline{i}'|\underline{d}, \underline{i}, a)$ to denote the quantity in (3.7).

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

$$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)}. \tag{3.8}$$

In the next section, we provide some preliminaries on MDPs. The terminologies used follow Borkar [39].

3.3 Preliminaries on MDPs

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

$$P(\underline{d}(t+1) = \underline{d}, \underline{i}(t+1) = \underline{i} \mid B^{t-1}, \{(\underline{d}(s), \underline{i}(s)),\ K \leq s \leq t\})$$
$$= \sum_{b=1}^{K} \Big[P(B_t = b \mid B^{t-1}, \{(\underline{d}(s), \underline{i}(s)),\ K \leq s \leq t\})$$
$$\cdot P(\underline{d}(t+1) = \underline{d}, \underline{i}(t+1) = \underline{i} \mid B_t = b,\ B^{t-1}, \{(\underline{d}(s), \underline{i}(s)),\ K \leq s \leq t\}) \Big]$$
$$= \sum_{b=1}^{K} \Big[P(B_t = b \mid B^{t-1}, \{(\underline{d}(s), \underline{i}(s)),\ K \leq s \leq t\})$$

$$\cdot P(\underline{d}(t+1) = \underline{d}, \underline{i}(t+1) = \underline{i} \mid (\underline{d}(t), \underline{i}(t)), B_t = b)\Big], \tag{3.9}$$

where the last line above follows from (3.3). From (3.9), it is evident that the policy π may be described completely by specifying $P(B_t|B^{t-1}, \{(\underline{d}(s), \underline{i}(s)), K \leq s \leq t\})$ for all $t \geq K$. We say that a policy π is a *stationary randomised strategy* (SRS) if there exists a Cartesian product λ of the form

$$\lambda = \bigotimes_{(\underline{d},\underline{i}) \in \mathbb{S}} \lambda_{(\underline{d},\underline{i})}, \tag{3.10}$$

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

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

Such an SRS π will be denoted π^λ. Note that $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is indeed a *Markov process* under the SRS π^λ. This follows from the relation (3.9) where the first probability term inside the summation in (3.9) is now a function only of $(\underline{d}(t), \underline{i}(t))$. Let Π_{SRS} denote the set of all SRS policies.

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

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

Lemma 9. *Let $\eta \in (0, 1]$. For every $\pi^\lambda \in \Pi_{\mathsf{SRS}}$, the controlled Markov process $\{\underline{d}(t), \underline{i}(t) : t \geq K\}$ under the policy π^λ is irreducible, aperiodic, positive recurrent, and hence ergodic.*

Proof. See Section 3.8.1. □

The proof of Lemma 9 relies on the hypothesis that the trembling hand parameter $\eta > 0$.

As a consequence of Lemma 9, it follows that under every SRS policy, a unique stationary distribution exists for the Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$. Let us call this stationary distribution μ^λ corresponding to the SRS policy π^λ.

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

3.4 Converse: Lower Bound

We now present a lower bound for (3.8).

Proposition 7. *Fix $h \in \mathcal{A}$, and assume that $\mathcal{H}_h$ is the true hypothesis. Let P_1 be the transition probability matrix of the Markov process of arm h, and for each $a \neq h$, let P_2 be the transition probability matrix of the Markov process arm a. Then,*

$$\liminf_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} \geq \frac{1}{R^*(P_1, P_2)}, \tag{3.11}$$

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

$$R^*(P_1, P_2) \coloneqq \sup_{\pi^\lambda \in \Pi_{SRS}} \min_{h' \neq h} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu^\lambda(\underline{d}, \underline{i}, a)\, k_{hh'}(\underline{d}, \underline{i}, a), \tag{3.12}$$

with

$$k_{hh'}(\underline{d}, \underline{i}, a) \coloneqq \begin{cases} D(P_1^{d_a}(\cdot|i_a) \| P_2^{d_a}(\cdot|i_a)), & a = h, \\ D(P_2^{d_a}(\cdot|i_a) \| P_1^{d_a}(\cdot|i_a)), & a = h', \\ 0, & a \neq h, h', \end{cases} \tag{3.13}$$

and

$$\nu^\lambda(\underline{d}, \underline{i}, a) \coloneqq \mu^\lambda(\underline{d}, \underline{i}) \left(\frac{\eta}{K} + (1 - \eta)\, \lambda(a|\underline{d}, \underline{i}) \right), \quad \forall (\underline{d}, \underline{i}, a) \in \mathbb{S} \times \mathcal{A}. \tag{3.14}$$

Proof. See Section 3.8.2. □

The proof of the lower bound follows the outline in [32], with necessary modifications for the setting of restless arms. The key ingredients are the data processing inequality for relative entropies, a Wald-type Lemma for Markov processes, and a recognition that, for any $(\underline{d}, \underline{i})$, the long-term fraction of exits from the state $(\underline{d}, \underline{i})$ matches the long-term fraction of entries into the state $(\underline{d}, \underline{i})$. This forces the long-term probability of seeing the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ in the state $(\underline{d}, \underline{i})$ to be that under its unique stationary distribution, by ergodicity (Lemma 9). These observations lead to (3.11).

Observe that the left-hand side of (3.11) is evaluated by taking into consideration *all* policies, including those that are not necessarily SRS policies, whereas the supremum in (3.12) is only over SRS policies. This is a consequence of [40, Theorem 8.8.2], a formal statement of which appears in Theorem 3 of Section 3.9.2 as applicable to the context of this chapter. For details on how Theorem 3 is used in the proof, see Section 3.8.2.

Finally, note that the constant $R^*(P_1, P_2)$ in (3.12) does not depend on the odd arm location h. This is due to symmetry in the structure of the arms.

3.4.1 Our 'Lift' Approach

It may be a little surprising to the reader as to why the summation on the right-hand side of (3.12) is over the delays and the last observed states of *all* the arms when the function $k_{hh'}(\underline{d}, \underline{i}, a)$, as given in (3.13), is a function only of d_a and i_a, the delay and the last observed state of arm a. In fact, the prior works [4, 5, 6, 36] suggest that it suffices to use (d_a, i_a) in place of $(\underline{d}, \underline{i})$ for deriving the lower bound. Relabelling $k_{hh'}(\underline{d}, \underline{i}, a)$ as $k_{hh'}(d_a, i_a, a)$ and proceeding to derive the lower bound as suggested by the prior works leads to a linear programming problem (LPP) with countably infinitely many linear constraints; see Section 3.9.1 for the details. However, it is not clear if the constraints of the above LPP constitute the tightest set of constraints. This is important because the optimal value of the LPP, say $R_1^*(P_1, P_2)$, may not necessarily be the smallest (best) constant for the problem at hand if the constraints are not tight, in which case we can only assert that $R_1^*(P_1, P_2) \geq R^*(P_1, P_2)$. In this case, it is not clear if this inequality is indeed an equality.

In contrast to the approach of using only (d_a, i_a) as suggested by the prior works, our 'lift' approach of using $(\underline{d}, \underline{i})$ automatically captures all the constraints of the LPP and makes the problem amenable to analysis, thereby enabling us to assert that $R^*(P_1, P_2)$ is the best (smallest) constant for the problem at hand. For more details on the LPP, see Section 3.9.1.

3.5 Achievability

The question of whether the supremum in (13) is a maximum, i.e., whether there exists an SRS policy that obtains the supremum value, is under study. Recall that this supremum is over all $\pi^\lambda \in \Pi_{\mathsf{SRS}}$ for $\lambda(\cdot|\cdot)$ which are conditional probability distributions on the arms, conditioned on the arm delays and the last observed states. This is in contrast to the works [4, 5, 6, 36] where the corresponding supremum is over all *unconditional* probability distributions on the arms. This is because, in those works, the arm delays are superfluous. The unconditional probability measures are elements of the probability simplex on $\mathcal{A}$, whereas the conditional probability measures are more complex due to the countably many possible values for the arm delays. In spite of this added complexity, we can come arbitrarily close to the supremum in (3.12). We shall use this fact in our achievability result, which is the topic of this section.

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

hypothesis $\mathcal{H}_{h'}$. Then, $Z_{hh'}(n)$ may be expressed as

$$\begin{aligned} Z_{hh'}(n) &= \log \frac{P_h(B^n, A^n, \bar{X}^n)}{P_{h'}(B^n, A^n, \bar{X}^n)} \\ &= \log \frac{P_h(B_0)}{P_{h'}(B_0)} + \log \frac{P_h(A_0|B_0)}{P_{h'}(A_0|B_0)} + \log \frac{P_h(\bar{X}_0|B_0, A_0)}{P_{h'}(\bar{X}_0|B_0, A_0)} \qquad (3.15) \\ &\quad + \sum_{t=1}^{n} \log \left(\frac{P_h(B_t|B^{t-1}, A^{t-1}, \bar{X}^{t-1})}{P_{h'}(B_t|B^{t-1}, A^{t-1}, \bar{X}^{t-1})} \right) \qquad (3.16) \\ &\quad + \sum_{t=1}^{n} \log \left(\frac{P_h(A_t|B^t, A^{t-1}, \bar{X}^{t-1})}{P_{h'}(A_t|B^t, A^{t-1}, \bar{X}^{t-1})} \right) \qquad (3.17) \\ &\quad + \sum_{t=1}^{n} \log \left(\frac{P_h(\bar{X}_t|A_t, B^t, A^{t-1}, \bar{X}^{t-1})}{P_{h'}(\bar{X}_t|A_t, B^t, A^{t-1}, \bar{X}^{t-1})} \right). \qquad (3.18) \end{aligned}$$

We now note that under the policy π, the probability of choosing arm B_t at time t, based on the history up to time t, cannot be a function of the underlying odd arm location (which is unknown to π), and must therefore be the same under hypotheses $\mathcal{H}_h$ and $\mathcal{H}_{h'}$. Thus, the first term in (3.15) and the expression in (3.16) are 0. Also, we note that $P_h(A_0|B_0) = P_{h'}(A_0|B_0)$, and for each t,

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

since A_t, the arm that is actually pulled at time t, is a function only of B_t and is related to B_t through (3.4). Therefore, given the history, the choice of A_t is not a function of the odd arm location, and is the same under hypotheses $\mathcal{H}_h$ and $\mathcal{H}_{h'}$, implying that the second term in (3.15) and the expression in (3.17) are 0. Finally, the probabilities in (3.18) do not depend on the intended arm pulls $\{B_t\}$ since the state $\bar{X}_t$ observed on arm A_t is a function only of the delay and the last observed state of arm A_t. Letting X_t^a denote the state of arm $A_t = a$, and defining

$$N(n, \underline{d}, \underline{i}, a) := \sum_{t=K}^{n} \mathbb{I}_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a\}}, \qquad (3.19)$$

$$N(n, \underline{d}, \underline{i}, a, j) := \sum_{t=K}^{n} \mathbb{I}_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a, X_t^a=j\}}, \qquad (3.20)$$

for all $(\underline{d}, \underline{i}, a) \in \mathbb{S} \times \mathcal{A}$, and using the assumption that arm 1 is selected at time $t = 0$, arm 2 at time $t = 1$ and so on until arm K at time $t = K - 1$, we have

$Z_{hh'}(n)$

$$= \sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)} + \sum_{t=K}^{n} \log \frac{P_h(\bar{X}_t|A_t, B^t, A^{t-1}, \bar{X}^{t-1})}{P_{h'}(\bar{X}_t|A_t, B^t, A^{t-1}, \bar{X}^{t-1})}$$

$$= \sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)} + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \sum_{a=1}^{K} \sum_{t=K}^{n} \mathbb{I}_{\{\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a,X_t^a=j\}} \log \frac{P_h(\bar{X}_t = j|A_t = a, B^t, A^{t-1}, \bar{X}^{t-1})}{P_{h'}(\bar{X}_t = j|A_t = a, B^t, A^{t-1}, \bar{X}^{t-1})}$$

$$= \sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)} + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \sum_{a=1}^{K} \sum_{t=K}^{n} \mathbb{I}_{\{\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a,X_t^a=j\}} \log \frac{P_h(X_t^a = j|A_t = a, X_{t-d_a}^a = i_a)}{P_{h'}(X_t^a = j|A_t = a, X_{t-d_a}^a = i_a)}$$

$$= \sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)} + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \sum_{a=1}^{K} \sum_{t=K}^{n} \mathbb{I}_{\{\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a,X_t^a=j\}} \log \frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)}$$

$$= \sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)} + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \sum_{a=1}^{K} N(n, \underline{d}, \underline{i}, a, j) \log \frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)} \tag{3.21}$$

$$= \sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)} + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \left[N(n, \underline{d}, \underline{i}, h, j) \log \frac{P_1^{d_h}(j|i_h)}{P_2^{d_h}(j|i_h)} + N(n, \underline{d}, \underline{i}, h', j) \log \frac{P_2^{d_{h'}}(j|i_{h'})}{P_1^{d_{h'}}(j|i_{h'})} \right]. \tag{3.22}$$

In the above set of equations, $P_h(X_{a-1}^a)$ denotes the law of the observation X_{a-1}^a obtained from arm a at time $a-1$ when the true hypothesis is $\mathcal{H}_h$; $P_{h'}(X_{a-1}^a)$ is defined similarly. Also, (3.22) follows by noting that

$$P_h^a = \begin{cases} P_1, & a = h, \\ P_2, & a \neq h, \end{cases} \qquad P_{h'}^a = \begin{cases} P_1, & a = h', \\ P_2, & a \neq h', \end{cases} \tag{3.23}$$

and thus the only nonzero terms in the summation over the arms in (3.21) are those corresponding to $a = h$ and $a = h'$.

To describe our policy, we first fix constants $\delta > 0$ and $L > 1$. These will be the parameters of our policy. Recall that the supremum in (3.12) is over all SRS policies. By the definition of this supremum, we know that for any fixed hypothesis $\mathcal{H}_h$ and given $\delta > 0$, there exists

$\lambda(\cdot \mid \cdot) = \lambda_{h,\delta}(\cdot \mid \cdot)$ such that under the SRS policy $\pi^{\lambda_{h,\delta}}$, we have

$$\min_{h' \neq h} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu^{\lambda_{h,\delta}}(\underline{d}, \underline{i}, a)\ k_{hh'}(\underline{d}, \underline{i}, a) \geq \frac{R^*(P_1, P_2)}{1+\delta}. \tag{3.24}$$

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

Our policy, which we call $\pi_1^\star(L, \delta)$, is then as below.

Policy $\pi_1^\star(L, \delta)$:

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

1. Let $\theta(n) \in \arg\max_{h \in \mathcal{A}} M_h(n)$. Resolve ties, if any, uniformly at random.
2. If $M_{\theta(n)}(n) \geq \log((K-1)L)$, stop further arm selections and declare $\theta(n)$ as the odd arm index.
3. If $M_{\theta(n)}(n) < \log((K-1)L)$, decide to pull arm B_n according to $\lambda_{\theta(n),\delta}(\cdot \mid \underline{d}(n), \underline{i}(n))$. Update $n \leftarrow n+1$ and go back to step 1.

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

Recall that the supremum in (3.12) is only over SRS policies. However, the policy $\pi_1^\star(L, \delta)$ described above is *not* an SRS policy since the distribution in item 3 is a function of $\theta(n)$ and could potentially depend on the entire history of arm selections and observations up to time n. Yet, as we show below, its performance comes arbitrarily close to the lower bound.

3.5.1 Performance of Policy $\pi_1^\star(L, \delta)$

We now present results on the performance of our policy.

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

Lemma 10. *Fix parameters $L > 1$ and $\delta > 0$. Under the non-stopping version of the policy $\pi_1^\star(L, \delta)$ which runs indefinitely (i.e., even if the condition in item 2 is true, it moves to item 3), for all $h, h' \in \mathcal{A}$ such that $h' \neq h$,*

$$\liminf_{n\to\infty} \frac{Z_{hh'}(n)}{n} > 0 \quad \textit{almost surely.} \tag{3.25}$$

Proof. See Section 3.8.3. □

Thanks to Lemma 10, we have $\liminf_{n\to\infty} M_h(n)/n > 0$ almost surely under the hypothesis $\mathcal{H}_h$. This implies that for any $h' \neq h$, almost surely,

$$\begin{aligned}
\limsup_{n\to\infty} M_{h'}(n) &= \limsup_{n\to\infty} \min_{a\neq h'} Z_{h'a}(n) \\
&\leq \limsup_{n\to\infty} Z_{h'h}(n) \\
&= \limsup_{n\to\infty} -Z_{hh'}(n) \\
&= -\liminf_{n\to\infty} Z_{hh'}(n) \\
&\leq -\liminf_{n\to\infty} M_h(n) \\
&< 0 \quad \text{almost surely.}
\end{aligned} \tag{3.26}$$

From the above set of inequalities, it follows that under policy $\pi_1^\star(L, \delta)$, almost surely,

$$\theta(n) = h \quad \forall\ n \text{ sufficiently large.} \tag{3.27}$$

Let $\pi_h^\star(L, \delta)$ denote a version of the policy $\pi_1^\star(L, \delta)$ that stops only at declaration h. It then follows that the stopping times of policies $\pi_1^\star(L, \delta)$ and $\pi_h^\star(L, \delta)$ are almost surely related as $\tau(\pi_h^\star(L, \delta)) \geq \tau(\pi_1^\star(L, \delta))$, as a consequence of which we have the following set of almost sure inequalities:

$$\begin{aligned}
&\tau(\pi_1^\star(L, \delta) \\
&\leq \tau(\pi_h^\star(L, \delta)) \\
&= \inf\{n \geq 1 : M_h(n) \geq \log((K-1)L)\} \\
&\leq \inf\left\{n \geq 1 : Z_{hh'}(n') \geq \log((K-1)L) \text{ for all } n' \geq n \text{ and for all } h' \neq h\right\} \\
&< \infty,
\end{aligned} \tag{3.28}$$

where the last line follows as a consequence of Lemma 10. This establishes that the policy $\pi_1^\star(L,\delta)$ stops in finite time almost surely.

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

Lemma 11. *Fix error probability $\epsilon > 0$. If $L = 1/\epsilon$, then for every $\delta > 0$, $\pi_1^\star(L,\delta) \in \Pi(\epsilon)$. Here, $\Pi(\epsilon)$ is as defined in (3.1).*

Proof. The proof uses the fact that the policy stops in finite time almost surely. See Section 3.8.4 for the details. □

With the above ingredients in place, we state the main result of this section, which is that the expected stopping time of our policy satisfies an asymptotic upper bound that comes arbitrarily close to the lower bound in (3.11).

Proposition 8. *Fix $h \in \mathcal{A}$ and $\delta > 0$, and let $\mathcal{H}_h$ be the true hypothesis. The stopping time of the policy $\pi_1^\star(L,\delta)$ satisfies*

$$\limsup_{L\to\infty} \frac{E_h[\tau(\pi_1^\star(L,\delta))]}{\log L} \leq \frac{1+\delta}{R^*(P_1,P_2)}. \tag{3.29}$$

Proof. In the proof, which we provide in Section 3.8.5, we first show that as $L \to \infty$ (equivalently $\epsilon \downarrow 0$), the ratio $\tau(\pi_1^\star(L,\delta))/\log L$ satisfies an almost sure upper bound that matches with the right-hand side of (3.29). We then show that the family $\{\tau(\pi_1^\star(L,\delta))/\log L : L > 1\}$ is uniformly integrable. Combining the almost sure upper bound with the uniform integrability result yields (3.29). □

3.6 Main Result

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

Theorem 2. *Consider a multi-armed bandit with $K \geq 3$ arms in which each arm is a time homogeneous and ergodic Markov process on the finite state space $\mathcal{S}$. Fix $h \in \mathcal{A}$, and suppose that h is the odd arm. Let P_1 be the transition probability matrix of the Markov process of arm h. Further, for all $a \neq h$, let the transition probability matrix of arm a be P_2, where $P_2 \neq P_1$. Fix $\eta \in (0,1]$, and suppose that the decision entity has a trembling hand with parameter η. Assuming that P_1 and P_2 are known to the decision entity, the expected time required by the decision entity to find the index of the odd arm satisfies the asymptotic relation*

$$\lim_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} = \lim_{\delta\downarrow 0} \lim_{L\to\infty} \frac{E_h[\tau(\pi_1^\star(L,\delta))]}{\log L} = \frac{1}{R^*(P_1,P_2)}. \tag{3.30}$$

Proof. From Lemma 11, we see that given any error tolerance parameter $\epsilon > 0$, by setting $L = 1/\epsilon$, we have $\pi_1^\star(L, \delta) \in \Pi(\epsilon)$ for all $\delta > 0$. Therefore, it follows that for all $\epsilon, \delta > 0$,

$$\inf_{\pi \in \Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log\left(\frac{1}{\epsilon}\right)} \leq \frac{E_h[\tau(\pi_1^\star(L,\delta))]}{\log L}. \tag{3.31}$$

Fixing $\delta > 0$ and letting $\epsilon \downarrow 0$ (which is identical to letting $L \to \infty$) in (3.31), and using the upper bound in (3.29), we get

$$\limsup_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} \leq \limsup_{L \to \infty} \frac{E_h[\tau(\pi_1^\star(L,\delta))]}{\log L} \leq \frac{1+\delta}{R^*(P_1, P_2)}. \tag{3.32}$$

Letting $\delta \downarrow 0$ in (3.32) and noting that the leftmost term in (3.32) does not depend on δ, we get

$$\limsup_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} \leq \lim_{\delta \downarrow 0} \limsup_{L \to \infty} \frac{E_h[\tau(\pi_1^\star(L,\delta))]}{\log L} \leq \frac{1}{R^*(P_1, P_2)}. \tag{3.33}$$

Combining the result in (3.33) with the lower bound in (3.11), we get

$$\begin{aligned} \frac{1}{R^*(P_1, P_2)} \leq \liminf_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} &\leq \limsup_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} \\ &\leq \lim_{\delta \downarrow 0} \limsup_{L \to \infty} \frac{E_h[\tau(\pi_1^\star(L,\delta))]}{\log L} \leq \frac{1}{R^*(P_1, P_2)}. \end{aligned} \tag{3.34}$$

Thus, it follows that the limit infimum and the limit suprema in (3.34) are indeed limits, thereby yielding (3.30). This completes the proof of the theorem. □

We thus see that the policy $\pi_1^\star(L, \delta)$ is asymptotically optimal. As noted in Lemma 11, the parameter L may be set appropriately so as to ensure that the policy meets the desired error probability at stoppage. Furthermore, the parameter δ may be set so as to ensure that the upper bound in (3.29) is within a desired accuracy from the lower bound in (3.11). Finally, we emphasise here that our analysis of the lower and upper bounds crucially relies on the trembling hand parameter η being strictly positive.

3.7 The Case $\eta = 0$

We now investigate the case $\eta = 0$. Let us first recall that the key result of Lemma 9, which states that under every SRS policy the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is an ergodic Markov process, crucially relies on the trembling hand parameter η being strictly positive. Such an ergodicity property may not be available when $\eta = 0$. While, in principle,

we may consider plugging $\eta = 0$ in (3.11) and treating the resulting expression as the lower bound for the case when $\eta = 0$, it is not clear if this new lower bound can be approached asymptotically through a sequence of strategies (policies) in the sense of (3.29). Therefore, it is a priori not clear if the results of this chapter extend directly to the case $\eta = 0$.

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

1. Writing $R^*(P_1, P_2)$ of (3.12) more explicitly as $R^*_\eta(P_1, P_2)$ for $\eta \in (0, 1]$, we show that $\lim_{\eta \downarrow 0} R^*_\eta(P_1, P_2)$ exists. This is based on a key monotonicity property which we elaborate upon in Section 3.7.1.

2. Writing $R^*_0(P_1, P_2)$ to denote the constant obtained by plugging $\eta = 0$ in (3.12), we demonstrate that

$$\lim_{\eta \downarrow 0} R^*_\eta(P_1, P_2) \leq R^*_0(P_1, P_2). \tag{3.35}$$

   It is not clear if, in general, the inequality in (3.35) is an equality.

3. We show in Section 3.7.2 and Section 3.7.3 that the lower bounds for the settings when either (a) each arm yields iid observations from a common finite alphabet, or (b) each arm yields Markov observations from a common finite state space and the arms are rested, may be recovered from (3.12) by plugging $\eta = 0$ in (3.12). Our proof of this is based on verifying that the hypotheses of the envelope theorem [38, Theorem 2] are satisfied for these settings. Thus, we show that the inequality in (3.35) is an equality for each of the above settings, thereby implying that the lower bounds for these settings may be approached asymptotically through a sequence of "trembling-hand" based policies similar to that presented in this chapter; the policy of [4, Section II.B] is an example case in point. This demonstrates that our analysis of the setting of restless arms carries over to the settings of the prior works with minor modifications.

3.7.1 A Key Monotonicity Property

Fix $\eta \in (0, 1]$, and assume that the decision entity possesses a trembling hand with parameter η. Let $\lambda = \lambda(\cdot \mid \cdot)$ be any conditional probability distribution on the arms, conditioned on the arm delays and the last observed states, as described in Section 3.3, and let Λ denote the set of all such conditional distributions. Define

$$\Lambda^\eta := \left\{ \frac{\eta}{K} + (1 - \eta)\, \lambda(\cdot \mid \cdot) : \lambda(\cdot \mid \cdot) \in \Lambda \right\}. \tag{3.36}$$

Note that for any $\lambda(\cdot \mid \cdot) \in \Lambda$, the corresponding element of Λ^η is the probability distribution according to which arms are *actually* selected, when the decision entity *intends* to pull the arms

according to $\lambda(\cdot \mid \cdot)$. Notice that $\Lambda^\eta \subset \Lambda$ for all $\eta \in (0, 1]$.

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

Lemma 12. *$\Lambda^\eta \subset \Lambda^{\eta'}$ for all $0 < \eta' < \eta \leq 1$.*

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

$$\begin{aligned}
\frac{\eta}{K} + (1-\eta)\ \lambda(a|\underline{d}, \underline{i}) &= \frac{\eta'}{K} + \frac{\eta - \eta'}{K} + (1-\eta)\ \lambda(a|\underline{d}, \underline{i}) \\
&= \frac{\eta'}{K} + (1-\eta') \left[\frac{\eta - \eta'}{1-\eta'} \cdot \frac{1}{K} + \frac{1-\eta}{1-\eta'}\ \lambda(a|\underline{d}, \underline{i}) \right] \\
&= \frac{\eta'}{K} + (1-\eta') \left[\frac{\eta''}{K} + (1-\eta'')\ \lambda(a|\underline{d}, \underline{i}) \right] \qquad (3.37) \\
&\in \Lambda^{\eta'}, \qquad (3.38)
\end{aligned}$$

where in (3.37), $\eta'' = \frac{\eta - \eta'}{1-\eta'} \in (0, 1]$, and (3.38) follows by noting that the term inside the square brackets in (3.37) is a valid element of Λ. The relation in (3.38) implies that every element of Λ^η is also an element of $\Lambda^{\eta'}$ whenever $\eta' < \eta$. This completes the proof. □

Plugging $\eta = 0$ in (3.36), and denoting the resulting set as Λ^0, we see that $\Lambda^0 = \Lambda$. Thus, it follows from Lemma 12 that

$$\bigcup_{\eta \downarrow 0} \Lambda^\eta \subset \Lambda. \qquad (3.39)$$

Let us now turn our attention to (3.14), and note that the right-hand side of (3.14) represents the long-term probability of seeing the state $(\underline{d}, \underline{i})$ and selecting arm a subsequently with probability $\frac{\eta}{K} + (1-\eta)\,\lambda(a|\underline{d}, \underline{i})$. Defining $\lambda^\eta(\cdot \mid \cdot) := \frac{\eta}{K} + (1-\eta)\,\lambda(\cdot \mid \cdot)$, and writing ν^λ in (3.14) as ν^{λ^η}, we may express the right-hand side of (3.12) equivalently as

$$R^*_\eta(P_1, P_2) := \sup_{\lambda^\eta(\cdot|\cdot) \in \Lambda^\eta} \min_{h' \neq h} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu^{\lambda^\eta}(\underline{d}, \underline{i}, a)\, k_{hh'}(\underline{d}, \underline{i}, a). \qquad (3.40)$$

It follows from Lemma 12 that $R^*_\eta(P_1, P_2)$ is non-decreasing in η; thus, $\lim_{\eta \downarrow 0} R^*_\eta(P_1, P_2)$ exists.

Finally, denoting by $R^*_0(P_1, P_2)$ the quantity obtained by plugging $\eta = 0$ in (3.40), it follows from (3.39) that (3.35) holds.

3.7.2 IID Observations From The Arms

We now show that when each arm yields iid observations coming from a finite alphabet common across the arms, the inequality in (3.35) is indeed an equality. Fix $h \in \mathcal{A}$, and suppose that

$\mathcal{H}_h$ is the true hypothesis. Let arm h be associated with an iid process whose underlying law is ν_1. Further, for all $h' \neq h$, let arm h' be associated with an iid process whose law is ν_2, where $\nu_2 \neq \nu_1$. Assume that the iid process of any given arm is independent of the iid process of each of the remaining arms. Let ν_h^a denote the marginal law of the iid process of arm a under the hypothesis $\mathcal{H}_h$, i.e.,

$$\nu_h^a = \begin{cases} \nu_1, & a = h, \\ \nu_2, & a \neq h. \end{cases} \tag{3.41}$$

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

$$P_1^d(j|i) = \nu_1(j), \quad P_2^d(j|i) = \nu_2(j). \tag{3.42}$$

Thus, when each arm yields iid observations, the function $k_{hh'}(\underline{d}, \underline{i}, a)$ in (3.13) may be expressed as

$$k_{hh'}(\underline{d}, \underline{i}, a) = \begin{cases} D(\nu_1 \| \nu_2), & a = h, \\ D(\nu_2 \| \nu_1), & a = h', \\ 0, & \text{otherwise.} \end{cases} \tag{3.43}$$

In other words, the function k does not depend on either the arm delays or the last observed states. Noting that the right-hand side of (3.43) may be written compactly as $D(\nu_h^a \| \nu_{h'}^a)$, and plugging this in (3.40), we get

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

$$= \sup_{\lambda \in \mathcal{P}(\mathcal{A})} \min_{h' \neq h} \frac{\eta}{K} \sum_{a=1}^{K} D(\nu_h^a \| \nu_{h'}^a) + (1-\eta) \sum_{a=1}^{K} \lambda(a)\, D(\nu_h^a \| \nu_{h'}^a) \tag{3.44}$$

$$= \sup_{\lambda \in \mathcal{P}(\mathcal{A})} \frac{\eta}{K} \left[D(\nu_1 \| \nu_2) + D(\nu_2 \| \nu_1) \right] + (1-\eta) \left[\lambda(h)\, D(\nu_1 \| \nu_2) + \left(\min_{h' \neq h} \lambda(h') \right) D(\nu_2 \| \nu_1) \right], \tag{3.45}$$

where in (a) above, μ^{λ^η} is the long-term probability of observing the state $(\underline{d}, \underline{i})$ when the arms are selected according to the distribution $\lambda^\eta(\cdot \mid \cdot)$, (b) above follows by using the fact that ν^{λ^η} is a probability distribution on $\mathbb{S} \times \mathcal{A}$, and the term $\lambda(a)$ in (3.44) is given by

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

with $\mathcal{P}(\mathcal{A})$ in (3.44) denoting the set of all probability distributions on the set $\mathcal{A}$. Lastly, (3.45) follows by noting that

$$\nu_h^a = \begin{cases} \nu_1, & a = h, \\ \nu_2, & a \neq h, \end{cases} \qquad \nu_{h'}^a = \begin{cases} \nu_1, & a = h', \\ \nu_2, & a \neq h', \end{cases} \tag{3.46}$$

and therefore the only non-zero terms in the summation over the arms in (3.44) are those corresponding to $a = h$ and $a = h'$.

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

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

is bounded and linear (hence absolutely continuous) for all $\eta \in [0,1]$. Using the envelope theorem [38, Theorem 2], we get that the mapping $\eta \mapsto R_\eta^*(P_1, P_2)$ is absolutely continuous for all $\eta \in [0,1]$, thereby implying that $\lim_{\eta \downarrow 0} R_\eta^*(P_1, P_2) = R_0^*(P_1, P_2)$. This establishes that the inequality in (3.35) holds with equality.

3.7.3 Rested Markov Arms

We now show that when each arm is a Markov process on a finite state space that is common across the arms, and the arms are rested, the inequality in (3.35) is indeed an equality. Fix $h \in \mathcal{A}$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Let each arm be associated with a time-homogeneous and ergodic discrete-time Markov process on a common, finite state space $\mathcal{S}$. Let P_1 be the transition probability matrix of the odd arm, and let P_2 be the transition probability

matrix of each of the non-odd arms. Let μ_1 and μ_2 denote the unique stationary distributions of P_1 and P_2 respectively. Assume that the Markov process of any given arm is independent of the Markov process of each of the remaining arms.

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

$$P_h^a = \begin{cases} P_1, & a = h, \\ P_2, & a \neq h, \end{cases} \qquad \mu_h^a = \begin{cases} \mu_1, & a = h, \\ \mu_2, & a \neq h. \end{cases} \tag{3.47}$$

When the arms are rested, as noted at the beginning of this section, the delay parameter for every arm is identically equal to 1, i.e., $d_a(t) \equiv 1$ for all $a \in \mathcal{A}$ and $t \geq K$. Thus, we may omit the summation over $\underline{d}$ in (3.40). Writing $\lambda(a|\underline{i})$ in place of $\lambda(a|\underline{d}, \underline{i})$, writing $\nu^{\lambda^\eta}(\underline{i})$ in place of $\nu^{\lambda^\eta}(\underline{d}, \underline{i})$, and the following the steps presented earlier for the case of iid observations, we have

$$\begin{aligned}
&R_\eta^*(P_1, P_2) \\
&= \sup_{\lambda^\eta(\cdot|\cdot)\in\Lambda^\eta} \min_{h'\neq h} \sum_{a=1}^K \sum_{\underline{i}\in\mathcal{S}^K} \nu^{\lambda^\eta}(\underline{i}, a)\, D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a)) \\
&= \sup_{\lambda(\cdot|\cdot)\in\Lambda} \min_{h'\neq h} \sum_{a=1}^K \sum_{\underline{i}\in\mathcal{S}^K} \mu^{\lambda^\eta}(\underline{i}) \left[\frac{\eta}{K} + (1-\eta)\,\lambda(a|\underline{i})\right] D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a)) \\
&= \sup_{\lambda(\cdot|\cdot)\in\Lambda} \min_{h'\neq h} \left[\frac{\eta}{K} \sum_{a=1}^K \sum_{\underline{i}\in\mathcal{S}^K} \mu^{\lambda^\eta}(\underline{i})\, D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a))\right. \\
&\qquad\qquad \left. + (1-\eta) \sum_{a=1}^K \sum_{\underline{i}\in\mathcal{S}^K} \mu^{\lambda^\eta}(\underline{i})\,\lambda(a|\underline{i})\, D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a))\right] \\
&\overset{(a)}{=} \sup_{\lambda(\cdot|\cdot)\in\Lambda} \min_{h'\neq h} \left[\frac{\eta}{K} \sum_{a=1}^K \sum_{\underline{i}\in\mathcal{S}^K} \mu^{\lambda^\eta}(\underline{i})\, D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a))\right. \\
&\qquad\qquad \left. +(1-\eta) \sum_{a=1}^K \sum_{i_a\in\mathcal{S}} \left(\sum_{\underline{i}^{-a}\in\mathcal{S}^{K-1}} \mu^{\lambda^\eta}(\underline{i})\,\lambda(a|\underline{i}) \right) D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a))\right] \\
&\overset{(b)}{=} \sup_{\lambda(\cdot|\cdot)\in\Lambda} \min_{h'\neq h} \left[\frac{\eta}{K} \sum_{a=1}^K \sum_{i_a\in\mathcal{S}} \mu^{\lambda^\eta}(i_a)\, D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a))\right. \\
&\qquad\qquad \left. +(1-\eta) \sum_{a=1}^K \sum_{i_a\in\mathcal{S}} \mu^{\lambda^\eta}(i_a)\,\lambda(a|i_a)\, D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a))\right],
\end{aligned} \tag{3.48}$$

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

$$\sum_{\underline{i}^{-a} \in \mathcal{S}^{K-1}} \mu^{\lambda^\eta}(\underline{i})\, \lambda(a|\underline{i}) = \mu^{\lambda^\eta}(i_a)\, \lambda(a|i_a). \tag{3.49}$$

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

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

$$\mu^{\lambda^\eta}(i_a)\, \lambda(a|i_a) = \lambda(a) \cdot \mu_h^a(i_a), \tag{3.50}$$

where in (3.50), $\lambda(a) = \sum\limits_{i_a \in \mathcal{S}} \mu^{\lambda^\eta}(i_a)\, \lambda(a|i_a)$. Using (3.50) in (3.48), we have

$$\begin{aligned}
&R_\eta^*(P_1, P_2) \\
&= \sup_{\lambda(\cdot|\cdot) \in \Lambda} \min_{h' \neq h} \Bigg[\frac{\eta}{K} \sum_{a=1}^{K} \sum_{i_a \in \mathcal{S}} \mu^{\lambda^\eta}(i_a)\, D(P_h^a(\cdot|i_a) \| P_{h'}^a(\cdot|i_a)) \\
&\qquad\qquad + (1-\eta) \sum_{a=1}^{K} \sum_{i_a \in \mathcal{S}} \lambda(a)\, \mu_h^a(i_a)\, D(P_h^a(\cdot|i_a) \| P_{h'}^a(\cdot|i_a)) \Bigg] \\
&\overset{(a)}{=} \sup_{\lambda \in \mathcal{P}(\mathcal{A})} \min_{h' \neq h} \left[\frac{\eta}{K} \sum_{a=1}^{K} D(P_h^a(\cdot|\cdot) \| P_{h'}^a(\cdot|\cdot) | \mu_h^a) + (1-\eta) \sum_{a=1}^{K} \lambda(a)\, D(P_h^a(\cdot|\cdot) \| P_{h'}^a(\cdot|\cdot) | \mu_h^a) \right] \\
&= \sup_{\lambda \in \mathcal{P}(\mathcal{A})} \Bigg[\frac{\eta}{K} \Big(D(P_1(\cdot|\cdot) \| P_2(\cdot|\cdot) \| \mu_1) + D(P_2(\cdot|\cdot) \| P_1(\cdot|\cdot) \| \mu_2) \Big) \\
&\qquad\qquad + (1-\eta) \Big(\lambda(h)\, D(P_1(\cdot|\cdot) \| P_2(\cdot|\cdot) \| \mu_1) + \Big(\min_{h' \neq h} \lambda(h') \Big) D(P_2(\cdot|\cdot) \| P_1(\cdot|\cdot) \| \mu_2) \Big) \Bigg]
\end{aligned} \tag{3.51}$$

where in (a) above,

$$D(P_h^a(\cdot|\cdot)\|P_{h'}^a(\cdot|\cdot)|\mu_h^a) := \sum_{i_a \in \mathcal{S}} \mu_h^a(i_a)\, D(P_h^a(\cdot|i_a)\|P_{h'}^a(\cdot|i_a)),$$

and (3.51) follows by noting that

$$P_h^a = \begin{cases} P_1, & a = h, \\ P_2, & a \neq h, \end{cases} \qquad \mu_h^a = \begin{cases} \mu_1, & a = h, \\ \mu_2, & a \neq h, \end{cases} \qquad P_{h'}^a = \begin{cases} P_1, & a = h', \\ P_2, & a \neq h', \end{cases} \qquad \mu_{h'}^a = \begin{cases} \mu_1, & a = h', \\ \mu_2, & a \neq h', \end{cases} \tag{3.52}$$

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

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

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

is bounded and linear (hence absolutely continuous) for all $\eta \in [0,1]$. Using the envelope theorem [38, Theorem 2], we get that the mapping $\eta \mapsto R_\eta^*(P_1, P_2)$ is absolutely continuous for all $\eta \in [0,1]$, thereby implying that $\lim_{\eta \downarrow 0} R_\eta^*(P_1, P_2) = R_0^*(P_1, P_2)$. This establishes that the inequality in (3.35) holds with equality.

3.7.4 A Subtle Remark on the Interpretation of $R_0^*(P_1, P_2)$

Recall that $R_0^*(P_1, P_2)$ denotes the constant obtained by plugging $\eta = 0$ in (3.12). The correct interpretation of this constant deserves some explanation, which is the content of this section. Recall that when $\eta > 0$, under every SRS policy, the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is an ergodic Markov process (Lemma 9). This, in conjunction with Theorem 3 of Section 3.9.2, leads to the supremum over the set of SRS policies in (3.12). However, it is important to note that Theorem 3 crucially relies on the ergodicity property given by Lemma 9, the proof of which in turn holds only for the case $\eta > 0$. Such an ergodicity property may not be available when $\eta = 0$. Therefore, it is not clear how, after plugging $\eta = 0$, the right hand side of (3.12) is to be interpreted; for e.g., ν^λ, the ergodic state-action occupancy measure under the SRS policy π^λ when $\eta > 0$, may no longer be interpreted so when $\eta = 0$.

In order to address the above mentioned issue, we appeal to the literature and note that a common assumption that appears in works that deal with controlled Markov processes is one

of "under every SRS policy, the underlying controlled Markov process is an ergodic Markov process"; see, for instance, [39, pp. 58, Section II] or [40]. Such an assumption readily holds for the case $\eta > 0$. Thus, $R_0^*(P_1, P_2)$ must be interpreted as the constant obtained by plugging $\eta = 0$ in (3.12), under the assumption[1] that every SRS policy makes the underlying controlled Markov process an ergodic Markov process.

3.8 Proofs

3.8.1 Proof of Lemma 9

Fix $h \in \mathcal{A}$ and an SRS policy $\pi^\lambda \in \Pi_{\mathsf{SRS}}$, and let $\mathcal{H}_h$ be the true hypothesis. Recall that under π^λ, the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is, in fact, a Markov process.

Proof of Irreducibility. Fix any two states $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$, and suppose that the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is in the state $(\underline{d}, \underline{i})$ at some time $t = T_0$. We shall now demonstrate that there exists N such that the state $(\underline{d}', \underline{i}')$ may be reached starting from the state $(\underline{d}, \underline{i})$ after N steps under π^λ. Recall that at any time t, the arm intended to be pulled is B_t, while the arm actually pulled at time t is its trembled version A_t; the arms A_t and B_t are related through the trembling hand relation in (3.4) as a consequence of which for all $a \in \mathcal{A}$, we have

$$
\begin{aligned}
P(A_t = a \mid B^{t-1}, A^{t-1}, \bar{X}^{t-1}) &= \sum_{b=1}^{K} P(B_t = b,\, A_t = a | B^{t-1}, A^{t-1}, \bar{X}^{t-1}) \\
&= \sum_{b=1}^{K} P(B_t = b \mid B^{t-1}, A^{t-1}, \bar{X}^{t-1}) \cdot P(A_t = a \mid B_t = b, B^{t-1}, A^{t-1}, \bar{X}^{t-1}) \\
&\overset{(a)}{=} \sum_{b=1}^{K} \lambda(b \mid \underline{d}(t), \underline{i}(t)) \cdot \left(\frac{\eta}{K} + (1-\eta)\, \mathbb{I}_{\{a=b\}} \right) \\
&= \frac{\eta}{K} + (1-\eta)\, \lambda(a \mid \underline{d}(t), \underline{i}(t)) \\
&\geq \frac{\eta}{K}, \qquad (3.53)
\end{aligned}
$$

where (a) above follows from (3.4) and the fact that under π^λ, the intended arm B_t is selected according to $\lambda(\cdot \mid \underline{d}(t), \underline{i}(t))$.

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 and P_2 are transition probability matrices on the finite set $\mathcal{S}$, we use [41, Proposition 1.7] for finite state Markov processes to

[1] Or any assumption that in turn guarantees ergodicity of the underlying controlled Markov process under every SRS policy.

deduce that there exists an integer M such that for all $m \geq M$,

$$P_1^m(j|i) > 0 \text{ for all } i, j \in \mathcal{S}, \quad P_2^m(j|i) > 0 \text{ for all } i, j \in \mathcal{S}. \tag{3.54}$$

Consider the following sequence of actions and observations: starting from the state $(\underline{d}, \underline{i})$ at time $t = T_0$, let the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ evolve for $M-1$ time instants. Thereafter, let arm 1 be selected at the (T_0+M)th time instant and let the state observed on arm 1 be i'_1; let arm 2 be selected at the $(T_0+M+d'_1-d'_2)$th time instant and let the state observed on arm 2 be i'_2, and so on. Finally, let arm K be observed at the $(T_0+M+d'_1-d'_K)$th time instant, and let the state observed on arm K be i'_K. Additionally, let arm 1 not be selected for all $T_0+M < t < T_0+M+d'_1$; let arm 2 not be selected for all $T_0+M+d'_1-d'_2 < t < T_0+M+d'_1$ and so on. Clearly, this sequence of actions and observations leads to the state $(\underline{d}', \underline{i}')$ after $M+d'_1-d'_K$ time instants. Thus, the probability of starting from the state $(\underline{d}, \underline{i})$ and reaching the state $(\underline{d}', \underline{i}')$ may be lower bounded by the probability that the above sequence of actions and observations occurs under π^λ which, when the true hypothesis is $\mathcal{H}_h$, is given by

$$\begin{aligned}
&\left(\prod_{a=1}^{K} P(A_{T_0+M+d'_1-d'_a} = a \mid B^{T_0+M+d'_1-d'_a-1}, A^{T_0+M+d'_1-d'_a-1}, \bar{X}^{T_0+M+d'_1-d'_a-1})\right) \\
&\qquad \cdot \left(\prod_{a=1}^{K} (P_h^a)^{M+d'_1-d'_a+d_a}(i'_a|i_a)\right) \\
&\qquad \cdot \left(\prod_{a=1}^{K-1} \prod_{t=T_0+M+d'_1-d'_a+1}^{T_0+M+d'_1-d'_{a+1}} P(A_t \notin \{1,\ldots,a\} \mid B^{t-1}, A^{t-1}, \bar{X}^{t-1})\right) \\
&\overset{(a)}{\geq} \left(\frac{\eta}{K}\right)^K \cdot \left[\prod_{a=1}^{K} (P_h^a)^{M+d'_1-d'_a+d_a}(i'_a|i_a)\right] \cdot \left[\prod_{a=1}^{K-1} \prod_{t=T_0+M+d'_1-d'_a+1}^{T_0+M+d'_1-d'_{a+1}} \frac{\eta(K-a)}{K}\right] \\
&\overset{(b)}{\geq} \left(\frac{\eta}{K}\right)^K \cdot \left[\prod_{a=1}^{K} (P_h^a)^{M+d'_1-d'_a+d_a}(i'_a|i_a)\right] \cdot \left[\prod_{a=1}^{K-1} \prod_{t=T_0+M+d'_1-d'_a+1}^{T_0+M+d'_1-d'_{a+1}} \frac{\eta}{K}\right] \\
&> 0,
\end{aligned} \tag{3.55}$$

where (a) above follows from the observation that the right-hand side of (3.53), for each t, is $\geq \eta/K$ and the fact that

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

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

Proof of Aperiodicity. Fix an arbitrary $(\underline{d}, \underline{i}) \in \mathbb{S}$. We shall now demonstrate that starting from the state $(\underline{d}, \underline{i})$, there is a strictly positive probability of the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ returning back to the state $(\underline{d}, \underline{i})$ after M' steps as well as after $(M'+1)$ steps, where M' is sufficiently large and such that (3.54) holds for all $m \geq M'$. This will then establish the desired aperiodicity property since the period of the state $(\underline{d}, \underline{i})$ is equal to the gcd of M' and $M'+1$, which is 1. Assume, without loss of generality, that $\underline{d}$ is such that $d_1 > d_2 > \cdots d_K = 1$. Let M be such that (3.54) holds for all $m \geq M$. Using arguments similar to those presented above in the proof of irreducibility, the probability of starting from the state $(\underline{d}, \underline{i})$ at some time $t = T_0$ and returning back to the state $(\underline{d}, \underline{i})$ after $M + d_1 - d_K$ time instants may be lower bounded, under the hypothesis $\mathcal{H}_h$, by

$$\left(\frac{\eta}{K}\right)^K \cdot \left[\prod_{a=1}^{K} (P_h^a)^{M+d_1}(i'_a|i_a)\right] \cdot \left[\prod_{a=1}^{K-1} \prod_{t=T_0+M+d_1-d_{a+1}}^{T_0+M+d_1-d_a+1} \frac{\eta}{K}\right] > 0. \tag{3.56}$$

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

Proof of positive recurrence. Let

$$p_\eta \coloneqq \frac{\eta}{K} \min\left\{ \min\{P_1^M(j|i) : i, j \in \mathcal{S}\}, \min\{P_2^M(j|i) : i, j \in \mathcal{S}\}\right\}; \tag{3.57}$$

here, once again, M is such that (3.54) holds for all $m \geq M$. Therefore, it follows that $p_\eta > 0$. Let

$$r(\pi^\lambda) \coloneqq \min\{t > K : \underline{d}(t) = \underline{d}(K), \underline{i}(t) = \underline{i}(K)\} \tag{3.58}$$

denote the first return time of the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ to its initial state (i.e., the state at time $t = K$) under π^λ. We may then upper bound $r(\pi^\lambda)$ as

$$r(\pi^\lambda) \leq M \cdot K \cdot \tau_\eta \quad \text{almost surely}, \tag{3.59}$$

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

that

$$\begin{aligned} E\left[r(\pi^{\lambda})\right] &\leq M \cdot K \cdot E[\tau_{\eta}] \\ &= M \cdot K \cdot \frac{1}{p_{\eta}} \\ &< \infty, \end{aligned} \tag{3.60}$$

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

3.8.2 Proof of Proposition 7

This proof is organised as follows. Given $\epsilon > 0$, we first obtain a lower bound for $E_h[Z_{hh'}(\tau(\pi)]$ for all $\pi \in \Pi(\epsilon)$ using a change of measure argument of Kaufmann et al. [32]. Following this, we obtain an upper bound for $E_h[Z_{hh'}(\tau(\pi)]$ in terms of $E_h[\tau(\pi)]$. Combining the upper and the lower bounds, and letting $\epsilon \downarrow 0$, we arrive at the desired result. The ergodicity property established in Lemma 9 plays a crucial role in deriving the final lower bound of (3.11).

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

As a first step towards deriving the lower bound, we use a result of Kaufmann et al. [32] to obtain a lower bound for $E_h[Z_{hh'}(\tau(\pi))]$ in terms of the error probability parameter ϵ. This is based on a generalisation of [32, Lemma 18], a change of measure argument for iid observations from the arms, to the setting of restless arms with Markov observations. We present this generalisation in the following lemma.

Lemma 13. *Fix $\pi \in \Pi(\epsilon)$, and let $\tau(\pi)$ be the stopping time of policy π. Let $\mathcal{F}_{\tau(\pi)}$ be the σ-algebra*

$$\mathcal{F}_{\tau(\pi)} = \{E \in \mathcal{F} : E \cap \{\tau(\pi) = t\} \in \mathcal{F}_t \text{ for all } t \geq 0\}, \tag{3.61}$$

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

$$P_{h'}(E) = E_h[1_E \exp(-Z_{hh'}(\tau(\pi)))] \tag{3.62}$$

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

Proof of Lemma 13. We prove the lemma by demonstrating, through mathematical induction, that the relation

$$E_{h'}[g(B^t, A^t, \bar{X}^t)] = E_h[g(B^t, A^t, \bar{X}^t) \exp(-Z_{hh'}(t))] \tag{3.63}$$

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

$$\begin{aligned}
E_{h'}[g(B_0, A_0, \bar{X}_0)] &= \sum_{b=1}^{K} \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} g(b, a, i)\ P_{h'}(B_0 = b, A_0 = a, \bar{X}_0 = i) \\
&= \sum_{b=1}^{K} \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} g(b, a, i)\ P_{h'}(B_0 = b)\ P_{h'}(A_0 = a | B_0 = b)\ P_{h'}(\bar{X}_0 = i | B_0 = b, A_0 = a) \\
&\overset{(a)}{=} \sum_{b=1}^{K} \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} g(b, a, i)\ P_{h}(B_0 = b)\ P_{h}(A_0 = a | B_0 = b)\ P_{h'}(\bar{X}_0 = i | A_0 = a) \\
&= \sum_{b=1}^{K} \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} g(b, a, i)\ P_{h}(B_0 = b)\ P_{h}(A_0 = a | B_0 = b)\ P_{h'}(X_0^a = i), \qquad (3.64)
\end{aligned}$$

where (a) follows using the facts that $P_h(B_0 = b) = P_{h'}(B_0 = b)$ and $P_h(A_0 = a|B_0 = b) = P_{h'}(A_0 = a|B_0 = b)$ (see Section 3.5). Assuming that $X_0^a \sim \nu$, where ν is a probability distribution on $\mathcal{S}$, independent of the true hypothesis (to the knowledge of which π is oblivious), we have

$$\begin{aligned}
E_{h'}[g(B_0, A_0, \bar{X}_0)] &= \sum_{b=1}^{K} \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} g(b, a, i)\ P_{h}(B_0 = b)\ P_{h}(A_0 = a | B_0 = b)\ \nu(i) \qquad (3.65) \\
&= \sum_{b=1}^{K} \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} g(b, a, i)\ P_{h}(B_0 = b)\ P_{h}(A_0 = a | B_0 = b)\ P_{h}(X_0^a = i | A_0 = a) \qquad (3.66) \\
&= \sum_{b=1}^{K} \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} g(b, a, i)\ P_{h}(B_0 = b)\ P_{h}(A_0 = a | B_0 = b)\ P_{h}(X_0^a = i | A_0 = a, B_0 = b). \qquad (3.67)
\end{aligned}$$

Also, we have (see Section 3.5)

$$Z_{hh'}(0) = \log \frac{P_h(B_0, A_0, \bar{X}_0)}{P_{h'}(B_0, A_0, \bar{X}_0)} = 0. \qquad (3.68)$$

Combining (3.67) and (3.68), we get $E_{h'}[g(B_0, A_0, \bar{X}_0)] = E_h[g(B_0, A_0, \bar{X}_0) \exp(-Z_{hh'}(0))]$, thus proving (3.63) for $t = 0$.

We now assume that (3.63) is true for some $t > 0$, and demonstrate that it also true for $t + 1$. By the law of iterated expectations,

$$E_{h'}[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})] = E_{h'}[E_{h'}[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1}) | \mathcal{F}_t]]. \qquad (3.69)$$

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

$$E_{h'}[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})|\mathcal{F}_t] = E_h[E_{h'}[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})|\mathcal{F}_t] \exp(-Z_{hh'}(t))]. \tag{3.70}$$

We now note that

$$\begin{aligned}
&E_{h'}[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})|\mathcal{F}_t] \exp(-Z_{hh'}(t)) \\
&\overset{(a)}{=} E_{h'}[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1}) \exp(-Z_{hh'}(t))|\mathcal{F}_t] \\
&= \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}} \Big[g(B^t, A^t, \bar{X}^t, b, a, i) \cdot P_{h'}(B_{t+1} = b|B^t, A^t, \bar{X}^t) \\
&\qquad\qquad \cdot P_{h'}(A_{t+1} = a|B^{t+1} = b, B^t, A^t, \bar{X}^t) \\
&\qquad\qquad \cdot P_{h'}(\bar{X}_{t+1} = i|B^{t+1} = b, A_{t+1} = a, B^t, A^t, \bar{X}^t) \cdot \exp(-Z_{hh'}(t)) \Big] \\
&\overset{(b)}{=} \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}} \Big[g(B^t, A^t, \bar{X}^t, b, a, i) \cdot P_h(B_{t+1} = b|B^t, A^t, \bar{X}^t) \cdot P_h(A_{t+1} = a|B^{t+1} = b, B^t, A^t, \bar{X}^t) \\
&\qquad\qquad \cdot P_{h'}(\bar{X}_{t+1} = i|A_{t+1} = a, A^t, \bar{X}^t) \cdot \exp(-Z_{hh'}(t)) \Big],
\end{aligned} \tag{3.71}$$

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

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

See Section 3.5 for a justification of why the above facts are true. It then follows that

$$\begin{aligned}
&\sum_{i\in\mathcal{S}} P_{h'}(\bar{X}_{t+1} = i|A_{t+1} = a, A^t, \bar{X}^t) \exp(-Z_{hh'}(t)) \\
&= \sum_{i\in\mathcal{S}} \frac{P_{h'}(\bar{X}_{t+1} = i|A_{t+1} = a, A^t, \bar{X}^t)}{P_h(\bar{X}_{t+1} = i|A_{t+1} = a, A^t, \bar{X}^t)} \exp(-Z_{hh'}(t))\, P_h(\bar{X}_{t+1} = i|A_{t+1} = a, A^t, \bar{X}^t) \\
&= \sum_{i\in\mathcal{S}} \exp(-Z_{hh'}(t+1, a, i))\, P_h(\bar{X}_{t+1} = i|A_{t+1} = a, A^t, \bar{X}^t).
\end{aligned} \tag{3.72}$$

where in (3.72), the quantity $Z_{hh'}(t+1,a,i)$ is defined as

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

Substituting (3.72) in (3.71) and simplifying, we get

$$\begin{aligned}
&E_{h'}[g(B^{t+1},A^{t+1},\bar{X}^{t+1})|\mathcal{F}_t]\ \exp(-Z_{hh'}(t)) \\
&= \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}}\Big[g(B^t,A^t,\bar{X}^t,b,a,i)\cdot P_h(B_{t+1}=b|B^t,A^t,\bar{X}^t)\cdot P_h(A_{t+1}=a|B_{t+1}=b,B^t,A^t,\bar{X}^t) \\
&\qquad\qquad \cdot P_h(\bar{X}_{t+1}=i|B_{t+1}=b,A_{t+1}=a,B^t,A^t,\bar{X}^t)\cdot\exp(-Z_{hh'}(t+1,a,i))\Big] \qquad (3.73)\\
&= E_h[g(B^{t+1},A^{t+1},\bar{X}^{t+1})\ \exp(-Z_{hh'}(t+1))|\mathcal{F}_t]. \qquad (3.74)
\end{aligned}$$

Applying $E_h[\cdot]$ to both sides of (3.74) and using the law of total expectations, we arrive at the desired relation. This proves (3.63) for all $t \geq 0$.

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

$$\begin{aligned}
P_{h'}(E) &= E_{h'}[1_E] \\
&= E_{h'}\left[\sum_{t\geq 0} 1_{E\cap\{\tau(\pi)=t\}}\right] \\
&\overset{(a)}{=} \sum_{t\geq 0} E_{h'}\left[1_{E\cap\{\tau(\pi)=t\}}\right] \\
&\overset{(b)}{=} \sum_{t\geq 0} E_h\left[1_{E\cap\{\tau(\pi)=t\}}\ \exp(-Z_{hh'}(t))\right] \\
&= \sum_{t\geq 0} E_h\left[1_{E\cap\{\tau(\pi)=t\}}\ \exp(-Z_{hh'}(\tau(\pi)))\right] \\
&= E_h\left[1_E\ \exp(-Z_{hh'}(\tau(\pi)))\right], \qquad (3.75)
\end{aligned}$$

where (a) is due to monotone convergence theorem, and (b) above follows from (3.63) and the fact that $E\cap\{\tau(\pi)=t\} \in \mathcal{F}_t$ for all $t \geq 0$ since $E \in \mathcal{F}_{\tau(\pi)}$. This completes the proof of the lemma. □

Lemma 13, in conjunction with [32, Lemma 19], yields the following inequality for all policies

$\pi \in \Pi(\epsilon)$ and all $h' \neq h$:

$$E_h[Z_{hh'}(\tau(\pi))] \geq \sup_{E \in \mathcal{F}_{\tau(\pi)}} d(P_h(E), P_{h'}(E)), \tag{3.76}$$

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

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

is the binary relative entropy function. As noted in [32], $x \mapsto d(x,y)$ is monotone increasing for $x < y$ and the $y \mapsto d(x,y)$ is monotone decreasing for any fixed x. Also, for any $\pi \in \Pi(\epsilon)$, we have

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

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

$$\min_{h' \neq h} E_h[Z_{hh'}(\tau(\pi))] \geq d(\epsilon, 1-\epsilon). \tag{3.77}$$

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

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

We now obtain an upper bound for the left-hand side of (3.77). Fix $\pi \in \Pi(\epsilon)$ and $h, h' \in \mathcal{A}$ such that $h' \neq h$ arbitrarily. Then, from (3.22),

$$\begin{aligned} &E_h[Z_{hh'}(\tau(\pi))] \\ &= E_h\left[\sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)}\right] + E_h\left[\sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \sum_{j \in \mathcal{S}} N(\tau(\pi), \underline{d}, \underline{i}, a, j) \log \frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)}\right]. \end{aligned} \tag{3.78}$$

To simplify the second expectation term in (3.78), we use the following lemma.

Lemma 14. *Fix $h \in \mathcal{A}$. For every $(\underline{d}, \underline{i}) \in \mathbb{S}$, $a \in \mathcal{A}$ and $j \in \mathcal{S}$,*

$$E_h[E_h[N(\tau(\pi), \underline{d}, \underline{i}, a, j)|X_{a-1}^a]|\tau(\pi)] = E_h[E_h[N(\tau(\pi), \underline{d}, \underline{i}, a)|X_{a-1}^a]|\tau(\pi)]\,(P_h^a)^{d_a}(j|i_a). \tag{3.79}$$

Proof of Lemma 14. Substituting $n = \tau(\pi)$ in (3.20), we have

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

$$= E_h\left[E_h\left[\sum_{t=K}^{\tau(\pi)} 1_{\{\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a,X_t^a=j\}}\middle| X_{a-1}^a\right]\middle|\tau(\pi)\right]$$
$$= E_h\left[\sum_{t=K}^{\tau(\pi)} P_h(\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a,X_t^a=j|X_{a-1}^a)\middle|\tau(\pi)\right]. \tag{3.80}$$

For each t in the range of the summation in (3.80), the conditional probability term for t may be expressed as

$$\begin{aligned}
&P_h(\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a,X_t^a=j|X_{a-1}^a)\\
&= P_h(\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a|X_{a-1}^a)\cdot P_h(X_t^a=j|A_t=a,\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},X_{a-1}^a)\\
&= P_h(\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a|X_{a-1}^a)\cdot (P_h^a)^{d_a}(j|i_a).
\end{aligned} \tag{3.81}$$

Plugging (3.81) back in (3.80) and simplifying, we arrive at the desired relation in (3.79). □

Using Lemma 14, the second expectation term on the right-hand side of (3.78) can be simplified as follows.

$$\begin{aligned}
&E_h\left[\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^{K}\sum_{j\in\mathcal{S}} N(\tau(\pi),\underline{d},\underline{i},a,j)\log\frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)}\right]\\
&= E_h\left[\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^{K}\sum_{j\in\mathcal{S}} N(\tau(\pi),\underline{d},\underline{i},a,j)\log\frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)}\right]\\
&= E_h\left[\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^{K}\sum_{j\in\mathcal{S}} E_h[E_h[N(\tau(\pi),\underline{d},\underline{i},a,j)|X_{a-1}^a]|\tau(\pi)]\log\frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)}\right]\\
&\overset{(a)}{=} E_h\left[\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^{K}\sum_{j\in\mathcal{S}} E_h[E_h[N(\tau(\pi),\underline{d},\underline{i},a)|X_{a-1}^a]|\tau(\pi)]\cdot(P_h^a)^{d_a}(j|i)\cdot\log\frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)}\right]\\
&= E_h\left[\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^{K} E_h[E_h[N(\tau(\pi),\underline{d},\underline{i},a)|X_{a-1}^a]|\tau(\pi)]\cdot D((P_h^a)^{d_a}(\cdot|i_a)\|(P_{h'}^a)^{d_a}(\cdot|i_a))\right]\\
&= \sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^{K} E_h[N(\tau(\pi),\underline{d},\underline{i},a)]\cdot D((P_h^a)^{d_a}(\cdot|i_a)\|(P_{h'}^a)^{d_a}(\cdot|i_a)),
\end{aligned} \tag{3.82}$$

where in the above set of equations, (a) follows from lemma 14, and (3.82) is due to monotone

convergence theorem and the fact that

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

Plugging (3.82) back in (3.78), we get

$$\begin{aligned} &E_h[Z_{hh'}(\tau(\pi))] \\ &= E_h\left[\sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)}\right] + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} E_h[N(\tau(\pi), \underline{d}, \underline{i}, a)] \cdot D((P_h^a)^{d_a}(\cdot|i_a)\|(P_{h'}^a)^{d_a}(\cdot|i_a)). \end{aligned} \tag{3.83}$$

Noting that

$$\begin{aligned} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} E_h[N(\tau(\pi), \underline{d}, \underline{i}, a)] &\overset{(a)}{=} E_h\left[\sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} N(\tau(\pi), \underline{d}, \underline{i}, a)\right] \\ &= E_h\left[\sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \sum_{t=K}^{\tau(\pi)} 1_{\{\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a\}}\right] \\ &= E_h\left[\sum_{t=K}^{\tau(\pi)} 1\right] \qquad (3.84) \\ &= E_h[\tau(\pi) - K + 1], \qquad (3.85) \end{aligned}$$

where (a) above is due to monotone convergence theorem, we write (3.83) as

$$\begin{aligned} &E_h[Z_{hh'}(\tau(\pi))] \\ &= E_h\left[\sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)}\right] \\ &+ \left(E_h[\tau(\pi) - K + 1]\right) \cdot \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \frac{E_h[N(\tau(\pi), (\underline{d}, \underline{i}), a)]}{E_h[\tau(\pi) - K + 1]} \cdot D((P_h^a)^{d_a}(\cdot|i_a)\|(P_{h'}^a)^{d_a}(\cdot|i_a)). \end{aligned} \tag{3.86}$$

Combining (3.77) and (3.86), and noting that (3.86) holds for all $h' \neq h$, we get

$$d(\epsilon, 1-\epsilon) \leq \min_{h'\neq h} \left\{ E_h\left[\sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)}\right] \right.$$

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

$$\leq \sup_{\nu}\min_{h'\neq h}\Bigg\{E_h\left[\sum_{a=1}^{K}\log\frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)}\right]$$

$$+\left(E_h[\tau(\pi)-K+1]\right)\cdot\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^{K}\nu(\underline{d},\underline{i},a)\,D((P_h^a)^{d_a}(\cdot|i_a)\|(P_{h'}^a)^{d_a}(\cdot|i_a))\Bigg\}, \tag{3.87}$$

where the supremum in (3.87) is over all state-action occupancy measures satisfying

$$\sum_{a=1}^{K}\nu(\underline{d}',\underline{i}',a)=\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^{K}\nu(\underline{d},\underline{i},a)\,Q(\underline{d}',\underline{i}'|\underline{d},\underline{i},a)\quad\text{for all }(\underline{d}',\underline{i}')\in\mathbb{S}, \tag{3.88}$$

$$\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^{K}\nu(\underline{d},\underline{i},a)=1, \tag{3.89}$$

$$\nu(\underline{d},\underline{i},a)\geq 0\quad\text{for all }(\underline{d},\underline{i},a)\in\mathbb{S}\times\mathcal{A}. \tag{3.90}$$

Recall that Q in (3.88) denotes the transition probability matrix given by (3.7). The left-hand side of (3.88) represents the long-term probability of leaving the state $(\underline{d},\underline{i})$, while the right-hand side of (3.89) represents the long-term probability of entering into the state $(\underline{d},\underline{i})$. Thus, (3.88) is the *global balance equation* for the controlled Markov process $\{(\underline{d}(t),\underline{i}(t)):t\geq K\}$. Equations (3.89) and (3.90) together imply that ν is a probability measure on $\mathbb{S}\times\mathcal{A}$.

As outlined in Section 3.3, the controlled Markov process $\{(\underline{d}(t),\underline{i}(t)):t\geq K\}$, together with the sequence $\{B_t:t\geq 0\}$ of intended arm selections (or equivalently the sequence $\{A_t:t\geq 0\}$ of actual arm selections), defines a Markov decision problem (MDP) with state space $\mathbb{S}$ and action space $\mathcal{A}$. From lemma 9, we know that $\{(\underline{d}(t),\underline{i}(t)):t\geq K\}$ is an ergodic Markov process under every SRS policy. This suffices to apply Theorem 3 of Section 3.9.2 to deduce a one-one correspondence between feasible solutions to (3.88)-(3.90) and policies in Π_{SRS}. In other words, Theorem 3 implies that for any given ν satisfying (3.88)-(3.90), we can find an SRS policy $\pi^\lambda\in\Pi_{\mathsf{SRS}}$ such that $\nu^\lambda(\underline{d},\underline{i},a)=\nu(\underline{d},\underline{i},a)$ for all $(\underline{d},\underline{i},a)\in\mathbb{S}\times\mathcal{A}$. Recall that under the SRS policy π^λ, the stationary distribution of the Markov process $\{(\underline{d}(t),\underline{i}(t)):t\geq K\}$ is μ^λ. The associated ergodic state occupancy measure, ν^λ, is then defined according to (3.14).

On account of Theorem 3, we may replace the supremum in (3.87) by a supremum over all

SRS policies. Doing so leads us to the relation

$$d(\epsilon, 1-\epsilon) \leq \sup_{\pi^\lambda \in \Pi_{\mathsf{SRS}}} \min_{h' \neq h} \Bigg\{ E_h \left[\sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)} \right] + \Big(E_h[\tau(\pi) - K + 1] \Big) \cdot \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu^\lambda(\underline{d}, \underline{i}, a)\, D((P_h^a)^{d_a}(\cdot|i_a) \| (P_{h'}^a)^{d_a}(\cdot|i_a)) \Bigg\}. \tag{3.91}$$

for all $\pi \in \Pi(\epsilon)$. Observe that the constant term multiplying $E_h[\tau(\pi) - K + 1]$ in (3.91) is finite; further, it is not a function of either ϵ or of $\pi \in \Pi(\epsilon)$. The finiteness of this constant follows from the following observation: denote by μ_h^a the stationary distribution of the transition probability matrix P_h^a (i.e., $\mu_h^a = \mu_1$ for $a = h$ and $= \mu_2$ for all $a \neq h$). An application of the ergodic theorem to the Markov process of arm a yields

$$D((P_h^a)^{d_a}(\cdot|i_a) \| (P_{h'}^a)^{d_a}(\cdot|i_a)) \longrightarrow D(\mu_h^a \| \mu_{h'}^a) < \infty \quad \text{as } d_a \to \infty. \tag{3.92}$$

Since every convergent sequence is bounded, we may write $D((P_h^a)^{d_a}(\cdot|i_a) \| (P_{h'}^a)^{d_a}(\cdot|i_a)) \leq C$ for all $(\underline{d}, \underline{i}, a) \in \mathbb{S} \times \mathcal{A}$, where $0 < C < \infty$. Using (3.89), it follows that the constant term multiplying $E_h[\tau(\pi) - K + 1]$ in (3.91) is bounded above by C.

Let us also note that the first term inside the braces in (3.91) does not depend on ϵ. Since $d(\epsilon, 1-\epsilon) \to d(0,1) = +\infty$ as $\epsilon \downarrow 0$, the boundedness of $R^*(P_1, P_2)$ shows that $\epsilon \downarrow 0$ is equivalent to $E_h[\tau(\pi)] \to \infty$ for all $\pi \in \Pi(\epsilon)$. Letting $\epsilon \downarrow 0$, and using $d(\epsilon, 1-\epsilon)/\log(1/\epsilon) \longrightarrow 1$ as $\epsilon \downarrow 0$, we arrive at the lower bound in (3.11). This completes the proof of the proposition.

3.8.3 Proof of Lemma 10

We note that the key ingredient in the proof of Lemma 9 is the strict positivity of the probability term in (3.53) when the trembling hand parameter $\eta > 0$. Clearly, this is satisfied even under the non-stopping version of the policy $\pi_1^\star(L, \delta)$. We leverage this to first show that under the non-stopping version of the policy $\pi_1^\star(L, \delta)$,

$$\liminf_{n \to \infty} \frac{N(n, \underline{d}, \underline{i})}{n} > 0 \quad \text{almost surely} \tag{3.93}$$

for every $(\underline{d}, \underline{i}) \in \mathbb{S}$, where for each $n \geq K$,

$$N(n, \underline{d}, \underline{i}) := \sum_{t=K}^{n} \mathbb{I}_{\{\underline{d}(t) = \underline{d},\ \underline{i}(t) = \underline{i}\}} \tag{3.94}$$

denotes the number of times the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ visits the state $(\underline{d}, \underline{i})$. Noting that P_1 and P_2 are transition probability matrices on the finite set $\mathcal{S}$, we use [41, Proposition 1.7] for finite state Markov processes to deduce that there exists an integer M such that for all $m \geq M$,

$$P_1^m(j|i) > 0 \text{ for all } i, j \in \mathcal{S}, \quad P_2^m(j|i) > 0 \text{ for all } i, j \in \mathcal{S}. \tag{3.95}$$

Fix an arbitrary $(\underline{d}, \underline{i}) \in \mathbb{S}$, and assume without loss of generality that $\underline{d}$ is such that $d_1 > d_2 > \cdots > d_K = 1$. Also assume, again without loss of generality, that the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ starts in the state $(\underline{d}, \underline{i})$, i.e., $\underline{d}(K) = \underline{d}$, $\underline{i}(K) = \underline{i}$. Let $p(\underline{d}, \underline{i})$ 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 $(\underline{d}, \underline{i})$. The analysis presented in Section 3.8.1 shows that this probability is lower bounded by the probability of returning after $M + d_1$ time instants given by (3.56). Since (3.56) is strictly positive, it follows that $p(\underline{d}, \underline{i}) > 0$.

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

$$N(n, \underline{d}, \underline{i}) \geq \text{Bin}\left(\frac{n - K + 1}{M + d_1}, \quad p(\underline{d}, \underline{i})\right) \quad \text{almost surely,} \tag{3.96}$$

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

$$\begin{aligned}
\liminf_{n\to\infty} \frac{N(n, \underline{d}, \underline{i})}{n} &\geq \liminf_{n\to\infty} \frac{\text{Bin}\left(\frac{n-K+1}{M+d_1}, \quad p(\underline{d}, \underline{i})\right)}{n} \\
&= \liminf_{n\to\infty} \frac{\text{Bin}\left(\frac{n-K+1}{M+d_1}, \quad p(\underline{d}, \underline{i})\right)}{\frac{n-K+1}{M+d_1}} \cdot \frac{n - K + 1}{n} \cdot \frac{1}{M + d_1} \\
&\overset{(a)}{=} \frac{p(\underline{d}, \underline{i})}{M + d_1} \\
&> 0,
\end{aligned} \tag{3.97}$$

where (a) above is due to the strong law of large numbers. This establishes (3.93).

We now show that for all $(\underline{d}, \underline{i}) \in \mathbb{S}$ and $a \in \mathcal{A}$,

$$\liminf_{n\to\infty} \frac{N(n, \underline{d}, \underline{i}, a)}{n} > 0 \quad \text{almost surely.} \tag{3.98}$$

We shall then use (3.98) to establish (3.25). Fix an arbitrary $a \in \mathcal{A}$, and define

$$S(n, \underline{d}, \underline{i}, a) := \sum_{t=K}^{n} \left[\mathbb{I}_{\{A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}\}} - P(A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | B^{t-1}, A^{t-1}, \bar{X}^{t-1}) \right]. \tag{3.99}$$

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

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

the collection $\{\mathbb{I}_{\{A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}\}} - P(A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | B^{t-1}, A^{t-1}, \bar{X}^{t-1})\}_{t \geq K}$ is a bounded martingale difference sequence. Using the concentration result [23, Theorem 1.2A] for bounded martingale difference sequences and subsequently applying the Borel-Cantelli lemma, we get that

$$\frac{S(n, \underline{d}, \underline{i}, a)}{n} \longrightarrow 0 \quad \text{as } n \to \infty, \quad \text{almost surely.} \tag{3.100}$$

This implies that for every choice of $\varepsilon > 0$, there exists N_ε sufficiently large such that

$$\frac{N(n, \underline{d}, \underline{i}, a)}{n} \geq \frac{1}{n} \sum_{t=K}^{n} P(A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | B^{t-1}, A^{t-1}, \bar{X}^{t-1}) - \varepsilon \quad \forall n \geq N_\varepsilon \text{ almost surely.} \tag{3.101}$$

Now, for each $t \geq K$, almost surely,

$$\begin{aligned}
&P(A_t = a, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | B^{t-1}, A^{t-1}, \bar{X}^{t-1}) \\
&= P(A_t = a | \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, B^{t-1}, A^{t-1}, \bar{X}^{t-1}) \cdot P(\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | B^{t-1}, A^{t-1}, \bar{X}^{t-1}) \\
&= \left[\frac{\eta}{K} + (1 - \eta) \lambda_{\theta(t), \delta}(a \mid \underline{d}, \underline{i}) \right] P(\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i} | B^{t-1}, A^{t-1}, \bar{X}^{t-1}) \\
&\geq \frac{\eta}{K} \cdot \mathbb{I}_{\{(\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}\}},
\end{aligned} \tag{3.102}$$

where (3.102) follows from the fact that $\underline{d}(t)$ and $\underline{i}(t)$ are measurable with respect to the history $(B^{t-1}, A^{t-1}, \bar{X}^{t-1})$. Plugging (3.102) in (3.101), we get

$$\frac{N(n, \underline{d}, \underline{i}, a)}{n} \geq \frac{\eta}{K} \cdot \frac{N(n, \underline{d}, \underline{i})}{n} - \varepsilon \quad \forall\, n \geq N_\varepsilon \quad \text{almost surely.} \tag{3.103}$$

Using (3.97) in (3.103), we get

$$\frac{N(n,\underline{d},\underline{i},a)}{n-K+1} \geq \frac{\eta}{K}\cdot\frac{p(\underline{d},\underline{i})}{2(M+d_1)} - \varepsilon \tag{3.104}$$

for all sufficiently large values of n, almost surely. Setting $\varepsilon = \frac{\eta}{2K}\cdot\frac{p(\underline{d},\underline{i})}{2(M+d_1)}$ establishes (3.98).

Proof of Lemma 10. For all $h, h' \in \mathcal{A}$ such that $h' \neq h$, we have

$$\frac{Z_{hh'}(n)}{n} = \sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},h,j)}{n}\log\frac{P_1^{d_h}(j|i_h)}{P_2^{d_h}(j|i_h)} + \frac{N(n,\underline{d},\underline{i},h',j)}{n}\log\frac{P_1^{d_{h'}}(j|i_{h'})}{P_2^{d_{h'}}(j|i_{h'})}. \tag{3.105}$$

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

$$\frac{N(n,\underline{d},\underline{i},a,j)}{N(n,\underline{d},\underline{i},a)} \longrightarrow (P_h^a)^{d_a}(j|i_a) \quad \text{as } n\to\infty \quad \text{almost surely.} \tag{3.106}$$

Using (3.106) in (3.105), we get that for every choice of ε, there exists N_ε sufficiently large such that for all $n\geq N_\varepsilon$, almost surely,

$$\begin{aligned}
\frac{Z_{hh'}(n)}{n} \geq{}& \sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},h)}{n}(P_1^{d_h}(j|i_h)+\varepsilon)\log P_1^{d_h}(j|i_h)\\
&+ \sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},h)}{n}(P_1^{d_h}(j|i_h)-\varepsilon)\log\frac{1}{P_2^{d_h}(j|i_h)}\\
&+ \sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},h')}{n}(P_2^{d_{h'}}(j|i_{h'})+\varepsilon)\log P_2^{d_{h'}}(j|i_{h'})\\
&+ \sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},h')}{n}(P_2^{d_{h'}}(j|i_{h'})-\varepsilon)\log\frac{1}{P_1^{d_{h'}}(j|i_{h'})}\\
={}& \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \frac{N(n,\underline{d},\underline{i},h)}{n} D(P_1^{d_h}(\cdot|i_h)\|P_2^{d_h}(\cdot|i_h)) + \frac{N(n,\underline{d},\underline{i},h')}{n} D(P_2^{d_{h'}}(\cdot|i_{h'})\|P_1^{d_{h'}}(\cdot|i_{h'}))\\
&+\varepsilon\Bigg[\sum_{(\underline{d},\underline{i})\in\mathbb{S}} \frac{N(n,\underline{d},\underline{i},h)}{n}\left(\sum_{j\in\mathcal{S}}\log P_1^{d_h}(j|i_h)\,P_2^{d_h}(j|i_h)\right)\\
&\qquad + \frac{N(n,\underline{d},\underline{i},h')}{n}\left(\sum_{j\in\mathcal{S}}\log P_1^{d_{h'}}(j|i_{h'})\,P_2^{d_{h'}}(j|i_{h'})\right)\Bigg].
\end{aligned} \tag{3.107}$$

As a consequence of the convergence theorem for finite state Markov processes [41, Theorem

4.9], we have

$$P_1^d(j|i) \longrightarrow \mu_1(j) > 0 \quad \text{as } d \to \infty$$
$$P_2^d(j|i) \longrightarrow \mu_2(j) > 0 \quad \text{as } d \to \infty \tag{3.108}$$

for all $i, j \in \mathcal{S}$. This implies that the term inside the square brackets in (3.107) is bounded from below (say by a constant $C < 0$). We then have

$$\begin{aligned}
&\frac{1}{n} Z_{hh'}(n) \\
&\geq \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \frac{N(n, \underline{d}, \underline{i}, h)}{n} D(P_1^{d_h}(\cdot|i_h) \| P_2^{d_h}(\cdot|i_h)) + \frac{N(n, \underline{d}, \underline{i}, h')}{n} D(P_2^{d_{h'}}(\cdot|i_{h'}) \| P_1^{d_{h'}}(\cdot|i_{h'})) + C\varepsilon \\
&\geq \frac{N(n, \underline{d}, \underline{i}, h)}{n} D(P_1^{d_h}(\cdot|i_h) \| P_2^{d_h}(\cdot|i_h)) + \frac{N(n, \underline{d}, \underline{i}, h')}{n} D(P_2^{d_{h'}}(\cdot|i_{h'}) \| P_1^{d_{h'}}(\cdot|i_{h'})) + C\varepsilon
\end{aligned} \tag{3.109}$$

for all $(\underline{d}, \underline{i}) \in \mathbb{S}$ and for all $n \geq N_\varepsilon$, almost surely. Now, fix an arbitrary $(\underline{d}, \underline{i}) \in \mathbb{S}$ such that $d_1 > d_2 > \cdots > d_K = 1$. From (3.103), we know that there exist constants $N_h, N_{h'}$ sufficiently large such that

$$\frac{N(n, \underline{d}, \underline{i}, h)}{n} \geq \frac{\eta}{K} \cdot \frac{p(\underline{d}, \underline{i})}{2(M + d_1)} - \varepsilon, \quad \frac{N(n, \underline{d}, \underline{i}, h')}{n} \geq \frac{\eta}{K} \cdot \frac{p(\underline{d}, \underline{i})}{2(M + d_1)} - \varepsilon \tag{3.110}$$

for all $n \geq \max\{N_h, N_{h'}, N_\varepsilon\}$, almost surely. Combining (3.110) and (3.109), we may choose $\varepsilon > 0$ appropriately so that the right-hand side of (3.109) is strictly positive. This establishes the desired result. □

3.8.4 Proof of Lemma 11

The policy $\pi_1^\star(L, \delta)$ commits error under the hypothesis $\mathcal{H}_h$ if one of the following is true:

1. The policy does not stop in finite time.
2. The policy stops in finite time and declares $h' \neq h$ as the true index of the odd arm.

The event in item 1 above has zero probability thanks to Lemma 10. Thus, the probability of error of policy $\pi = \pi_1^\star(L, \delta)$ may be evaluated as follows: under the hypothesis $\mathcal{H}_h$,

$$P_h(\theta(\tau(\pi)) \neq h) = P_h\Big(\exists\ n \text{ and } \ h' \neq h \text{ such that } \tau(\pi) = n \text{ and } \theta(n) = h'\Big). \tag{3.111}$$

We now let

$$\mathcal{R}_{h'}(n) \coloneqq \{\omega : \tau(\pi)(\omega) = n,\, \theta(\tau(\pi))(\omega) = h'\} \tag{3.112}$$

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

$$\begin{aligned}
&P_h(\theta(\tau(\pi)) \neq h) \\
&= P_h\left(\bigcup_{h' \neq h} \bigcup_{n=0}^{\infty} \mathcal{R}_{h'}(n)\right) \\
&= \sum_{h' \neq h} \sum_{n=0}^{\infty} P_h(\tau(\pi) = n,\ \theta(\tau(\pi)) = h') \\
&= \sum_{h' \neq h} \sum_{n=0}^{\infty} \int\limits_{\mathcal{R}_{h'}(n)} dP_h(\omega) \\
&\overset{(a)}{=} \sum_{h' \neq h} \sum_{n=0}^{\infty} \int\limits_{\mathcal{R}_{h'}(n)} \exp(Z_h(n)(\omega)) \quad d(B^n(\omega), A^n(\omega), \bar{X}^n(\omega)) \\
&\overset{(b)}{=} \sum_{h' \neq h} \sum_{n=0}^{\infty} \int\limits_{\mathcal{R}_{h'}(n)} \exp(-Z_{h'h}(n)(\omega)) \quad \exp(Z_{h'}(n)(\omega)) \quad d(B^n(\omega), A^n(\omega), \bar{X}^n(\omega)) \\
&\overset{(c)}{\leq} \sum_{h' \neq h} \sum_{n=0}^{\infty} \left\{ \int\limits_{\mathcal{R}_{h'}(n)} \frac{1}{(K-1)L}\, dP_{h'}(\omega) \right\} \\
&= \sum_{h' \neq h} \frac{1}{(K-1)L}\, P_{h'}\left(\bigcup_{n=0}^{\infty} \mathcal{R}_{h'}(n)\right) \leq \frac{1}{L},
\end{aligned} \tag{3.113}$$

where in (a) above,

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

denotes the log-likelihood of all the intended arm pulls, the actual arm pulls and the observations up to time n under the hypothesis $\mathcal{H}_h$, (b) above follows by noting that $Z_{hh'}(n) = Z_h(n) - Z_{h'}(n) = -Z_{h'h}(n)$, and (c) follows from the fact that when $\mathcal{H}_{h'}$ is the true hypothesis, the condition $M_{h'}(n) \geq \log((K-1)L)$ is satisfied when the policy $\pi = \pi_1^\star(L, \delta)$ stops at time $\tau(\pi) = n$, which in particular implies that $Z_{h'h}(n) \geq \log((K-1)L)$. Setting $L = 1/\epsilon$ yields the desired result.

3.8.5 Proof of Proposition 8

This section is organised as follows. First, we show in Proposition 9 that under the policy $\pi_1^\star(L,\delta)$, the test statistic $M_h(n)$ has the correct drift, one that comes from the ergodic occupancy measure corresponding to $\pi^{\lambda_{h,\delta}}$ when $\mathcal{H}_h$ is the true hypothesis. We then show in Lemma 15 that the stopping time of the policy $\pi_1^\star(L,\delta)$ grows with L (i.e., lower probability of error implies more time required to stop and declare the odd arm location correctly with high confidence). Further, we show in Lemma 16 that ratio $\tau(\pi)/\log L$ has, in the limit as $L\to\infty$, an almost sure upper bound that matches with the right-hand side of (3.29). Finally, we prove in Proposition 10 that the family $\{\tau(\pi)/\log L : L>1\}$ is uniformly integrable. The almost sure upper bound of Lemma 16 combined with uniform integrability result of Proposition 10 yields the desired upper bound in (3.29).

Proposition 9. *Fix parameters $L>1$ and $\delta>0$. Also fix $h\in\mathcal{A}$ and assume that $\mathcal{H}_h$ is the true hypothesis. Under the non-stopping version of the policy $\pi_1^\star(L,\delta)$, for all $h'\in\mathcal{A}$ such that $h'\neq h$, the following relation holds almost surely:*

$$\lim_{n\to\infty}\frac{Z_{hh'}(n)}{n}=\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\nu^{\lambda_{h,\delta}}(\underline{d},\underline{i},h)\,D(P_1^{d_h}(\cdot|i_h)\|P_2^{d_h}(\cdot|i_h))+\nu^{\lambda_{h,\delta}}(\underline{d},\underline{i},h')\,D(P_2^{d_{h'}}(\cdot|i_{h'})\|P_1^{d_{h'}}(\cdot|i_{h'})). \tag{3.114}$$

Consequently, it follows that almost surely,

$$\begin{aligned}&\lim_{n\to\infty}\frac{M_h(n)}{n}\\&=\min_{h'\neq h}\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\nu^{\lambda_{h,\delta}}(\underline{d},\underline{i},h)\,D(P_1^{d_h}(\cdot|i_h)\|P_2^{d_h}(\cdot|i_h))+\nu^{\lambda_{h,\delta}}(\underline{d},\underline{i},h')\,D(P_2^{d_{h'}}(\cdot|i_{h'})\|P_1^{d_{h'}}(\cdot|i_{h'})).\end{aligned} \tag{3.115}$$

Proof of Proposition 9. The condition in (3.27) implies that for all $(\underline{d},\underline{i},a)\in\mathbb{S}\times\mathcal{A}$,

$$\begin{aligned}&\lim_{n\to\infty}P(A_n=a|\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},\{(\underline{d}(t),\underline{i}(t)):K\le t<n\})\\&=\lim_{n\to\infty}\frac{\eta}{K}+(1-\eta)\,\lambda_{\theta(n),\delta}(a|\underline{d},\underline{i})\\&=\frac{\eta}{K}+(1-\eta)\,\lambda_{h,\delta}(a|\underline{d},\underline{i}).\end{aligned} \tag{3.116}$$

Thus, we observe that because the arms are selected according to $\lambda_{\theta(n),\delta}(\cdot\mid\cdot)$ in the beginning, the non-stopping version of policy $\pi_1^\star(L,\delta)$ may not be regarded as an SRS policy (since $\theta(n)$ is, in general, a function of the entire history up to time n). However, $\theta(n)=h$ for all sufficiently

large values of n, and therefore the non-stopping version of policy $\pi_1^\star(L,\delta)$ eventually turns into an SRS policy. As an immediate consequence of this, we have the following almost sure convergences as $n \to \infty$:

$$\frac{N(n,\underline{d},\underline{i},a)}{N(n,\underline{d},\underline{i})} \longrightarrow \frac{\eta}{K} + (1-\eta)\,\lambda_{h,\delta}(a|\underline{d},\underline{i}), \tag{3.117}$$

$$\frac{N(n,\underline{d},\underline{i})}{n} \longrightarrow \mu^{\lambda_{h,\delta}}(\underline{d},\underline{i}). \tag{3.118}$$

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

$$\begin{aligned}
&\lim_{n\to\infty} \frac{Z_{hh'}(n)}{n} \\
&= \lim_{n\to\infty} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},h,j)}{n} \log \frac{P_1^{d_h}(j|i_h)}{P_2^{d_h}(j|i_h)} + \frac{N(n,\underline{d},\underline{i},h',j)}{n} \log \frac{P_2^{d_{h'}}(j|i_{h'})}{P_1^{d_{h'}}(j|i_{h'})} \\
&= \lim_{n\to\infty} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \left(\frac{N(n,\underline{d},\underline{i})}{n}\right) \left(\frac{N(n,\underline{d},\underline{i},h)}{N(n,\underline{d},\underline{i})}\right) \left(\frac{N(n,\underline{d},\underline{i},h,j)}{N(n,\underline{d},\underline{i},h)}\right) \log \frac{P_1^{d_h}(j|i_h)}{P_2^{d_h}(j|i_h)} \\
&+ \lim_{n\to\infty} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \left(\frac{N(n,\underline{d},\underline{i})}{n}\right) \left(\frac{N(n,\underline{d},\underline{i},h')}{N(n,\underline{d},\underline{i})}\right) \left(\frac{N(n,\underline{d},\underline{i},h',j)}{N(n,\underline{d},\underline{i},h')}\right) \log \frac{P_2^{d_{h'}}(j|i_{h'})}{P_1^{d_{h'}}(j|i_{h'})}.
\end{aligned} \tag{3.119}$$

Note that in each of the logarithmic terms in (3.119), when either the numerator or the denominator is equal to 0, the corresponding coefficient term is also equal to 0. Thus, we may assume without loss of generality that each term inside the summations in (3.119) is nonzero for all values of the summation indices. Under this assumption, it follows from the convergences in (3.108) that the logarithmic terms in (3.119) are bounded. Using the dominated convergence theorem to pass the limit inside the summation in each of the terms, and using the results in (3.106), (3.117) and (3.118), we arrive at the desired result. □

We now show that the stopping time of the policy $\pi_1^\star(L,\delta)$ grows with L.

Lemma 15. *Fix $h \in \mathcal{A}$ and $\delta > 0$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Then, under policy $\pi = \pi^\star(L,\delta)$, we have*

$$\liminf_{L\to\infty} \tau(\pi) = \infty \text{ almost surely.} \tag{3.120}$$

Proof of Lemma 15. Because we assume that the policy $\pi = \pi^\star(L,\delta)$ pulls arm 1 at time $t = 0$, arm 2 at time $t = 1$ and so on until arm K at time $t = K-1$, in order to prove the lemma, it

suffices to prove the following statement:

$$\text{for each } m \geq K, \quad \lim_{L\to\infty} P_h(\tau(\pi) \leq m) = 0. \tag{3.121}$$

Fix $m \geq K$, and note that

$$\begin{aligned}
&\limsup_{L\to\infty} P_h(\tau(\pi) \leq m) \\
&= \limsup_{L\to\infty} P_h\Big(\exists\ K \leq n \leq m \text{ and } \tilde{h} \in \mathcal{A} \text{ such that } M_{\tilde{h}}(n) > \log((K-1)L)\Big) \\
&\leq \limsup_{L\to\infty} \sum_{\tilde{h}\in\mathcal{A}} \sum_{n=K}^{m} P_h(M_{\tilde{h}}(n) > \log((K-1)L)) \\
&\leq \limsup_{L\to\infty} \frac{1}{\log((K-1)L)} \sum_{\tilde{h}\in\mathcal{A}} \sum_{n=K}^{m} E_h[M_{\tilde{h}}(n)],
\end{aligned} \tag{3.122}$$

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

We now show that for each $n \in \{K, \ldots, m\}$, the expectation term inside the summation in (3.122) is finite. This will then imply that the limit supremum on the right-hand side of (3.122) is equal to 0, thus proving the desired result. Note that almost surely,

$$M_{\tilde{h}}(n) = \min_{h'\neq\tilde{h}} Z_{\tilde{h}h'}(n) \leq Z_{\tilde{h}h'}(n) \text{ for all } h' \neq \tilde{h}. \tag{3.123}$$

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

$$\begin{aligned}
Z_{\tilde{h}h'}(n) &= \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} N(n,\underline{d},\underline{i},\tilde{h},j) \log \frac{P_1^{d_{\tilde{h}}}(j|i_{\tilde{h}})}{P_2^{d_{\tilde{h}}}(j|i_{\tilde{h}})} + N(n,\underline{d},\underline{i},h',j) \log \frac{P_2^{d_{h'}}(j|i_{h'})}{P_1^{d_{h'}}(j|i_{h'})} \\
&\leq n \max\left\{ \max\left\{ \log \frac{P_1^d(j|i)}{P_2^d(j|i)} : d \in \mathbb{N}, i, j \in \mathcal{S} \right\}, \max\left\{ \log \frac{P_2^d(j|i)}{P_1^d(j|i)} : d \in \mathbb{N}, i, j \in \mathcal{S} \right\} \right\}.
\end{aligned} \tag{3.124}$$

From to the convergences in (3.108), we note that the coefficient of n in (3.124) is finite. Thus, it follows that $E[M_{\tilde{h}}(n)] \leq E[Z_{\tilde{h}h'}(n)] \leq nC$ for all $h' \neq \tilde{h}$, where $C < \infty$ represents the constant multiplying n in (3.124). □

Going further, let $R_{\lambda_{h,\delta}}$ denote the right-hand side of (3.115).

Lemma 16. *Fix $h \in \mathcal{A}$ and $\delta > 0$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Then, under policy $\pi = \pi_1^\star(L, \delta)$, we have*

$$\limsup_{L\to\infty} \frac{\tau(\pi)}{\log L} \leq \frac{1}{R_{\lambda_{h,\delta}}} \quad \textit{almost surely.} \tag{3.125}$$

Proof of Lemma 16. Note that as a consequence of Proposition 9 and Lemma 15, we have

$$\lim_{L\to\infty} \frac{M_h(\tau(\pi))}{\tau(\pi)} = R_{\lambda_{h,\delta}} \quad \text{almost surely.} \tag{3.126}$$

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

$$\begin{aligned}
&Z_{hh'}(n) - Z_{hh'}(n-1) \\
&= \log \frac{P_h(A^n, \bar{X}^n)}{P_{h'}(A^n, \bar{X}^n)} - \log \frac{P_h(A^{n-1}, \bar{X}^{n-1})}{P_{h'}(A^{n-1}, \bar{X}^{n-1})} \\
&= \log \frac{P_h^{A_n}(\bar{X}_n | A^{n-1}, \bar{X}^{n-1})}{P_{h'}^{A_n}(\bar{X}_n | A^{n-1}, \bar{X}^{n-1})} \\
&= \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \sum_{j\in\mathcal{S}} \mathbb{I}_{\{\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},A_n=a,X_n^a=j\}} \log \frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)} \\
&= \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \left[\mathbb{I}_{\{\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},A_n=h,X_n^h=j\}} \log \frac{P_1^{d_h}(j|i_h)}{P_2^{d_h}(j|i_h)} + \mathbb{I}_{\{\underline{d}(n)=\underline{d},\underline{i}(n)=\underline{i},A_n=h',X_n^{h'}=j\}} \log \frac{P_2^{d_{h'}}(j|i_{h'})}{P_1^{d_{h'}}(j|i_{h'})} \right].
\end{aligned} \tag{3.127}$$

We now note that whenever either the numerator or the denominator of the logarithmic terms in (3.127) is equal to 0, then the corresponding indicator function is also equal to 0. This, together with the convergences in (3.108), implies that the right-hand side of (3.127) is bounded. This, together with the collection $\{Z_{hh'}(n) - Z_{hh'}(n-1) : 1 \leq n \leq K-1\}$ of finitely many terms, each of which is finite almost surely, establishes the almost sure boundedness of the increments $Z_{hh'}(n) - Z_{hh'}(n-1)$ for all $n \geq 1$ and all $h' \neq h$.

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

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

$$
\begin{aligned}
&\leq \limsup_{L\to\infty} \frac{Z_{hh''}(\tau(\pi))}{\log L} \\
&\overset{(a)}{=} \limsup_{L\to\infty} \frac{Z_{hh''}(\tau(\pi)-1)}{\log L} \\
&\leq \limsup_{L\to\infty} \frac{\log((K-1)L)}{\log L} \\
&= 1 \quad \text{almost surely,}
\end{aligned} \tag{3.128}
$$

where (a) above is due to the almost sure boundedness of the increments established earlier. Then, using (3.126) along with (3.128) yields

$$
\begin{aligned}
\limsup_{L\to\infty} \frac{\tau(\pi)}{\log L} &= \limsup_{L\to\infty} \left\{ \left(\frac{\tau(\pi)}{M_h(\tau(\pi))} \right) \left(\frac{M_h(\tau(\pi))}{\log L} \right) \right\} \\
&= \left(\lim_{L\to\infty} \frac{\tau(\pi)}{M_h(\tau(\pi))} \right) \left(\limsup_{L\to\infty} \frac{M_h(\tau(\pi))}{\log L} \right) \\
&\leq \frac{1}{R_{\lambda_{h,\delta}}} \quad \text{almost surely,}
\end{aligned} \tag{3.129}
$$

thus completing the proof of the lemma. □

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

$$
\limsup_{L\to\infty} \frac{\tau(\pi)}{\log L} \leq \frac{1+\delta}{R^*(P_1,P_2)} \quad \text{almost surely.} \tag{3.130}
$$

We now prove that the family $\{\tau(\pi_1^\star(L,\delta))/\log L : L > 1\}$ is uniformly integrable for all $\delta > 0$. This, along with the almost sure upper bound of (3.130) yields the desired upper bound of (3.29).

Proposition 10. *For each fixed $\delta > 0$, the family $\{\tau(\pi_1^\star(L,\delta))/\log L : L > 1\}$ is uniformly integrable.*

Proof of Proposition 10. Fix $h \in \mathcal{A}$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Then, in order to establish the desired uniform integrability, it suffices to show that

$$
\limsup_{L\to\infty} E_h\left[\exp\left(\frac{\tau(\pi)}{\log L} \right) \right] < \infty. \tag{3.131}
$$

Towards this, let us first define

$$D_{hh'} \coloneqq \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \nu^{\lambda_{h,\delta}}(\underline{d},\underline{i},a)\, D((P_h^a)^{d_a}(\cdot \mid i_a) \| (P_{h'}^a)^{d_a}(\cdot \mid i_a)). \tag{3.132}$$

Let

$$\tilde{n}(L) \coloneqq \frac{4\log((K-1)L)}{D_{hh'}} + K - 1, \tag{3.133}$$

and let

$$u(L) \coloneqq \exp\left(\frac{1+\tilde{n}(L)}{\log L}\right). \tag{3.134}$$

Recall the policy $\pi_h^\star(L,\delta)$ that stops only at declaration h. Also recall that $\tau(\pi_h^\star) \geq \tau(\pi)$ almost surely. Using this, we have

$$\begin{aligned}
&\limsup_{L\to\infty} E_h\left[\exp\left(\frac{\tau(\pi)}{\log L}\right)\right] \\
&= \limsup_{L\to\infty} \int_0^\infty P_h\left(\frac{\tau(\pi)}{\log L} > \log x\right) dx \\
&\leq \limsup_{L\to\infty} \int_0^\infty P_h\left(\tau(\pi_h^\star) \geq \lceil (\log x)(\log L)\rceil\right) dx \\
&\overset{(a)}{\leq} \limsup_{L\to\infty} \left\{u(L) + \int_{u(L)}^\infty P_h\left(\tau(\pi_h^\star) \geq \lceil (\log x)(\log L)\rceil\right) dx\right\} \\
&= \exp\left(\frac{4}{D_{hh'}}\right) + \limsup_{L\to\infty} \sum_{n\geq\tilde{n}(L)} \exp\left(\frac{n+1}{\log L}\right) P_h(M_h(n) < \log((K-1)L)),
\end{aligned} \tag{3.135}$$

where (a) above follows by upper bounding the probability term by 1 for all $x \leq u(L)$. In Lemma 17, we show that the probability term in (3.135) has an exponential upper bound. It then follows that this exponential upper bound results in the finiteness of the right-hand side of (3.135), thus completing the proof of the proposition. □

3.8.6 An Exponential Upper Bound for $P_h(M_h(n) < \log((K-1)L))$

We now demonstrate the stated exponential upper bound used in (3.135).

Lemma 17. *Fix $\delta > 0$ and $h \in \mathcal{A}$, and suppose that $\mathcal{H}_h$ is the true hypothesis. There exist*

constants $B > 0$ and $0 < \theta < \infty$ independent of L such for all $n \geq \tilde{n}(L)$,

$$P_h(M_h(n) < \log((K-1)L)) \leq Be^{-n\theta}. \tag{3.136}$$

Proof of Lemma 17. Since

$$\begin{aligned} P_h(M_h(n) < \log((K-1)L)) &= P_h\left(\min_{h' \neq h} Z_{hh'}(n) < \log((K-1)L)\right) \\ &\leq \sum_{h' \neq h} P_h\left(Z_{hh'}(n) < \log((K-1)L)\right); \end{aligned} \tag{3.137}$$

the last line above follows from the union bound. In order to prove the lemma, it suffices to show that each term inside the summation in (3.137) is exponentially bounded.

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

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

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

For $n \geq K$, let

$$\begin{aligned} \Delta Z_{hh'}(n) &\coloneqq Z_{hh'}(n) - Z_{hh'}(n-1) \\ &= \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \sum_{j \in \mathcal{S}} \mathbb{I}_{\{\underline{d}(n)=\underline{d}, \underline{i}(n)=\underline{i}, A_n=a, X_n^a=j\}} \log \frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)} \end{aligned} \tag{3.138}$$

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

Lemma 18. *For any $(\underline{d}, \underline{i}) \in \mathbb{S}$, $a \in \mathcal{A}$ and $0 < s < 1$, we have*

$$E_h\left[e^{s\Delta Z_{h'h}(n)} \middle| A_n = a, \underline{d}(n) = \underline{d}, \underline{i}(n) = \underline{i}\right] \leq 1 \quad \forall n, \tag{3.139}$$

with strict inequality in (3.139) if $(\underline{d}, \underline{i}) = (\underline{d}^*, \underline{i}^*)$ and $a = h$.

Proof of Lemma 18. Note that

$$\begin{aligned}
&E_h\left[e^{s\Delta Z_{h'h}(n)}\middle| A_n = h, \underline{d}(n) = \underline{d}, \underline{i}(n) = \underline{i}\right] \\
&= \sum_{j\in\mathcal{S}} \left(\frac{(P^a_{h'})^{d_a}(j|i_a)}{(P^a_h)^{d_a}(j|i_a)}\right)^s P_h(X^h_n = j | A_n = a, \underline{d}(n) = \underline{d}, \underline{i}(n) = \underline{i}) \\
&= \sum_{j\in\mathcal{S}} \left(\frac{(P^a_{h'})^{d_a}(j|i_a)}{(P^a_h)^{d_a}(j|i_a)}\right)^s (P^a_h)^{d_a}(j|i_a) \\
&= \sum_{j\in\mathcal{S}} ((P^a_h)^{d_a}(j|i_a))^{1-s} \ ((P^a_{h'})^{d_a}(j|i_a))^s \\
&\overset{(a)}{\leq} \left(\sum_{j\in\mathcal{S}} (P^a_h)^{d_a}(j|i_a)\right)^{1-s} \cdot \left(\sum_{j\in\mathcal{S}} (P^a_{h'})^{d_a}(j|i_a)\right)^s \\
&= 1, \qquad (3.140)
\end{aligned}$$

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

As an immediate consequence of Lemma 18, we have the following result.

Lemma 19. *For any $(\underline{d}, \underline{i}) \in \mathbb{S}$, $a \in \mathcal{A}$ and $0 < s < 1$, we have*

$$E_h\left[e^{s\Delta Z_{h'h}(n)} \mid \mathcal{F}_{n-1}\right] \mathbb{I}_{\{\underline{d}(n)=\underline{d}, \underline{i}(n)=\underline{i}\}} \leq 1 \quad \forall n \quad \text{almost surely}, \qquad (3.141)$$

with strict inequality in (3.139) if $(\underline{d}, \underline{i}) = (\underline{d}^*, \underline{i}^*)$ and $a = h$.

Proof of Lemma 19. We have, almost surely,

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

$$+ (1 - P(A_n = h \mid \underline{d}(n) = \underline{d}, \underline{i}(n) = \underline{i}, \mathcal{F}_{n-1}))$$

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

$$\overset{(c)}{\leq} \frac{\eta}{K} \left(E_h \left[e^{s\Delta Z_{h'h}(n)} \mid A_n = h, \underline{d}(n) = \underline{d}, \underline{i}(n) = \underline{i} \right] - 1 \right) + 1 \tag{3.142}$$

$$\overset{(d)}{\leq} 1, \tag{3.143}$$

where (a) above follows by noting that

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

(b) uses the result of Lemma 18, (c) follows from the fact that for any $n \geq K$, under the policy $\pi_1^\star(L, \delta)$,

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

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

Going further, let c denote the constant on the right-hand side of (3.142) when $(\underline{d}, \underline{i}) = (\underline{d}^*, \underline{i}^*)$. From the arguments above, we have $c < 1$. Then,

$$\begin{aligned} & E_h \left[e^{s\Delta Z_{h'h}(n)} \mid \mathcal{F}_{n-1} \right] \\ &= \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} E_h \left[e^{s\Delta Z_{h'h}(n)} \mid \mathcal{F}_{n-1} \right] \cdot \mathbb{I}_{\{\underline{d}(n) = \underline{d}, \underline{i}(n) = \underline{i}\}} \\ &= c\, \mathbb{I}_{\{\underline{d}(n) = \underline{d}^*, \underline{i}(n) = \underline{i}^*\}} + \sum_{(\underline{d},\underline{i}) \neq (\underline{d}^*, \underline{i}^*)} E_h \left[e^{s\Delta Z_{h'h}(n)} \mid \underline{d}(n) = \underline{d}, \underline{i}(n) = \underline{i}, \mathcal{F}_{n-1} \right] \cdot \mathbb{I}_{\{\underline{d}(n) = \underline{d}, \underline{i}(n) = \underline{i}\}} \\ &= \begin{cases} c, & \underline{d}(n) = \underline{d}^*, \underline{i}(n) = \underline{i}^*, \\ \leq 1, & \text{otherwise.} \end{cases} \end{aligned} \tag{3.144}$$

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

Lemma 20. *For $0 < s < 1$,*

$$E_h \left[e^{sZ_{h'h}(n)} \right] \leq B_1\, e^{-\theta_1 n}, \tag{3.145}$$

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

Proof of Lemma 20. We have

$$
\begin{aligned}
&E_h\left[e^{sZ_{h'h}(n)}\right] \\
&= E_h[e^{sZ_{h'h}(n-1)}\, E_h[e^{s\Delta Z_{h'h}(n)} \mid \mathcal{F}_{n-1}]] \\
&\overset{(a)}{\leq} E_h\left[c^{N(n,\underline{d}^*,\underline{i}^*)}\right] \\
&\overset{(b)}{=} E_h\left[c^{N(n,\underline{d}^*,\underline{i}^*)}\ ;\ N(n,\underline{d}^*,\underline{i}^*) > \frac{n\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*)}{2}\right] + E_h\left[c^{N(n,\underline{d}^*,\underline{i}^*)}\ ;\ N(n,\underline{d}^*,\underline{i}^*) \leq \frac{n\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*)}{2}\right] \\
&\leq c^{n\frac{\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*)}{2}} + P_h\left(N(n,\underline{d}^*,\underline{i}^*) \leq \frac{n\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*)}{2}\right). \qquad (3.146)
\end{aligned}
$$

In the above set of equations, (a) follows from by repeatedly applying (3.144), the notation $E[X;A]$ in (b) stands for $E[X\,\mathbb{I}_A]$, and the last line follows by noting that $c^{n\frac{\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*)}{2}} \leq 1$ almost surely. We now note that $\{N(n,\underline{d}^*,\underline{i}^*) - N(K,\underline{d}^*,\underline{i}^*) : n \geq K\}$ is a bounded martingale. Using the Azuma-Hoeffding inequality, we then have

$$
\begin{aligned}
&P_h\left(N(n,\underline{d}^*,\underline{i}^*) \leq \frac{n\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*)}{2}\right) \\
&= P_h\left(N(n,\underline{d}^*,\underline{i}^*) - N(K,\underline{d}^*,\underline{i}^*) \leq \frac{n\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*)}{2} - N(K,\underline{d}^*,\underline{i}^*)\right) \\
&\leq P_h\left(N(n,\underline{d}^*,\underline{i}^*) - N(K,\underline{d}^*,\underline{i}^*) \leq \frac{n\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*)}{2}\right) \\
&\leq \exp\left(-\frac{n(\mu^{\lambda_{h,\delta}}(\underline{d}^*,\underline{i}^*))^2}{8}\right). \qquad (3.147)
\end{aligned}
$$

Plugging (3.147) back in (3.146), and noting that c is a function of s, we arrive at (3.145). □

As a consequence of Lemma 20, we have the following result.

Lemma 21. *Fix an arbitrary $h \in \mathcal{A}$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Consider the non-stopping version of the policy $\pi = \pi_1^\star(L,\delta)$. There exist constants C_R and $\gamma > 0$ such that*

$$
P_h\left(\min_{h' \neq h} Z_{hh'}(n) < R\right) \leq C_R\, e^{-\gamma n}. \qquad (3.148)
$$

In (3.148), C_R is independent of h but γ depends on h.

Proof of Lemma 21. Observe that

$$
P_h\left(\min_{h' \neq h} Z_{hh'}(n) < R\right) = P_h\left(\max_{h' \neq h} Z_{h'h}(n) > -R\right)
$$

$$
\begin{aligned}
&\leq \sum_{h' \neq h} P_h \left(Z_{h'h}(n) > -R \right) \\
&= \sum_{h' \neq h} P_h \left(sZ_{h'h}(n) > -sR \right) \quad \forall\ 0 < s < 1 \\
&\overset{(a)}{\leq} \sum_{h' \neq h} e^{sR}\ E_h \left[e^{sZ_{h'h}(n)} \right] \\
&\overset{(b)}{\leq} e^{sR} \sum_{h' \neq h} B_1\ e^{-\theta n} \\
&\leq e^{sR} \cdot (K-1) \cdot \max_{h' \neq h} B_1\, e^{-\theta n} \\
&\leq C_R\ e^{-\gamma n}, \qquad (3.149)
\end{aligned}
$$

where $\max\limits_{h' \neq h} B_1\, e^{-\theta n} = e^{-\gamma}$ and $C_R = K e^{sR}$. In the above set of equations, (a) is due to Chernoff's bound for $0 < s < 1$, and (b) is due to Lemma 20. □

From (3.27), we know that under the non-stopping version of the policy $\pi_1^\star(L, \delta)$, the guess of the odd arm $\theta(n)$ eventually settles at h with probability 1 under the hypothesis $\mathcal{H}_h$. Indeed, we now show using Lemma 21 that something stronger holds. Towards this, fix $h \in \mathcal{A}$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Let

$$
T_h := \inf\{n : \theta(n') = h \text{ for all } n' \geq n\}. \qquad (3.150)
$$

We have the following result for T_h.

Lemma 22. *Fix an arbitrary $h \in \mathcal{A}$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Consider the non-stopping version of the policy $\pi_1^\star(L, \delta)$. There exist constants $C > 0$ and $b > 0$, both finite and possibly depending on h, such that*

$$
P_h \left(T_h > n \right) \leq C\, e^{-bn}. \qquad (3.151)
$$

Proof of Lemma 22. We have

$$
\begin{aligned}
P_h \left(T_h > n \right) &\leq P_h \left(\exists\ n' \geq n \text{ such that } \theta(n') \neq h \right) \\
&\leq \sum_{n' \geq n} P_h(\theta(n') \neq h) \\
&= \sum_{n' \geq n} P_h \left(\exists\ h' \neq h \text{ such that } \theta(n') = h' \right)
\end{aligned}
$$

$$\leq \sum_{n' \geq n} P_h \left(M_{h'}(n') > \max_{h'' \neq h'} M_{h''}(n) \right)$$
$$\leq \sum_{n' \geq n} P_h \left(M_h(n') - M_{h'}(n') < 0 \right). \tag{3.152}$$

We now note that, almost surely,

$$\begin{aligned} M_h(n') - M_{h'}(n') &= M_h(n') - \min_{h'' \neq h'} Z_{h'h''}(n') \\ &\geq M_h(n') - Z_{h'h}(n') \\ &= M_h(n') + Z_{hh'}(n') \\ &\geq 2 \min_{h' \neq h} Z_{hh'}(n'). \end{aligned} \tag{3.153}$$

Using (3.153) in (3.152), we get

$$P_h \left(T_h > n \right) \leq \sum_{n' \geq n} P_h \left(\min_{h' \neq h} Z_{hh'}(n) < 0 \right). \tag{3.154}$$

The result now follows from Lemma 21. □

We now use the results presented above to derive the desired exponential upper bound for each term of the summation in (3.137) to finish the proof of Lemma 17. Note that for any $\epsilon' > 0$, we have

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

$$+ P_h\Big((n-K+1)\,(D_{hh'}-2\epsilon') < \log((K-1)L)\Big). \quad (3.155)$$

We first choose ϵ' such that

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

In particular, it suffices to set $\epsilon' = D_{hh'}/4$. Let us fix this value of ϵ' for the rest of the proof, and note that this choice of ϵ' ensures that the third probability term in (3.155) is equal to 0. We now focus on the first probability term in (3.155), and note that each term inside the summation has strictly positive mean. Thus, from Chernoff's bounding technique [9, Lemma 2], we get that there exists $b(\epsilon')$ such that

$$P_h\Big(\sum_{k=K}^{n}(\Delta Z_{hh'}(k) - E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] + \epsilon') < 0\Big) \leq e^{-(n-K+1)\,b(\epsilon')}. \quad (3.156)$$

It thus remains to show that the second probability term in (3.155) is bounded above exponentially. To do so, we use the proof technique of Vaidhiyan et al. [4, pp. 4793-4794] and adapt it to our setting of restless arms.

Let

$$\begin{aligned}\tilde{C} &:= \inf_{(\underline{d},\underline{i})\in\mathbb{S},\ a\in\mathcal{A}} E_h[\Delta Z_{hh'}(n) \mid A_n = a, \underline{d}(n) = \underline{d}, \underline{i}(n) = \underline{i}] - D_{hh'} \\ &= \inf_{(\underline{d},\underline{i})\in\mathbb{S},\ a\in\mathcal{A}} D((P_h^a)^{d_a}(\cdot \mid i_a)\|(P_{h'}^a)^{d_a}(\cdot \mid i_a)) - D_{hh'}. \quad (3.157)\end{aligned}$$

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

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

here, $\epsilon' = D_{hh'}/4$ as chosen earlier. We may then write the second probability in (3.155) as follows:

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

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

$$\leq P_h\Big(\sum_{k=K}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \epsilon') < 0,\ T_h \leq n\epsilon'' \Big) + P_h\Big(T_h > n\epsilon'' \Big). \tag{3.158}$$

From Lemma 22, the second probability term in (3.158) is bounded above exponentially. The first probability term in (3.158) may be upper bounded as

$$\begin{aligned}
&P_h\Big(\sum_{k=K}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \epsilon') < 0,\ T_h \leq n\epsilon'' \Big) \\
&= P_h\Big(\sum_{k=K}^{\lfloor n\epsilon'' \rfloor} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \epsilon') \\
&\qquad\qquad + \sum_{k=\lfloor n\epsilon'' \rfloor+1}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \epsilon') < 0,\ T_h \leq n\epsilon'' \Big) \\
&\overset{(a)}{\leq} P_h\Big((\lfloor n\epsilon'' \rfloor - K + 1)(\tilde{C} + \epsilon') + \sum_{k=\lfloor n\epsilon'' \rfloor+1}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \epsilon') < 0,\ T_h \leq n\epsilon'' \Big) \\
&= P_h\Big((\lfloor n\epsilon'' \rfloor - K + 1)(\tilde{C} + \epsilon') + (n - \lfloor n\epsilon'' \rfloor)(\epsilon' - \tilde{\epsilon}) \\
&\qquad\qquad + \sum_{k=\lfloor n\epsilon'' \rfloor+1}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \tilde{\epsilon}) < 0,\ T_h \leq n\epsilon'' \Big) \\
&\overset{(b)}{\leq} P_h\Big(\lfloor n\epsilon'' \rfloor(\epsilon''\tilde{C} + \epsilon') - (K-1)\epsilon' + \sum_{k=\lfloor n\epsilon'' \rfloor+1}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \tilde{\epsilon}) < 0,\ T_h \leq n\epsilon'' \Big) \\
&\overset{(c)}{\leq} P_h\Big(\sum_{k=\lfloor n\epsilon'' \rfloor+1}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \tilde{\epsilon}) < 0,\ T_h \leq n\epsilon'' \Big) \\
&= \tilde{P}_h\Big(\sum_{k=\lfloor n\epsilon'' \rfloor+1}^{n} (E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}] - D_{hh'} + \tilde{\epsilon}) < 0 \Big),
\end{aligned} \tag{3.159}$$

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

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

$$\geq \lfloor n\epsilon'' \rfloor (\epsilon'' \tilde{C} + \epsilon') - (K-1)\epsilon' \tag{3.160}$$

since $\tilde{C} \leq 0$, (c) and the equality in (3.159) hold for all n such that $\lfloor n\epsilon'' \rfloor (\epsilon'' \tilde{C} + \epsilon') - (K-1)\epsilon' \geq 0$, and in (3.159), $\tilde{P}_h$ is a new probability measure under which at each time instant, an arm is selected according to the policy $\pi_1^\star(L, \delta)$ but assuming that the guess of the odd arm $\theta(k) = h$ for all k.

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

$$\begin{aligned}
&\tilde{E}_h[E_h[\Delta Z_{hh'}(k)|\mathcal{F}_{k-1}]] \\
&= \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \tilde{P}_h(\underline{d}(k) = \underline{d}, \underline{i}(k) = \underline{i}) \left(\frac{\eta}{K} + (1-\eta)\lambda_{h,\delta}(a|\underline{d},\underline{i}) \right) D((P_h^a)^{d_a}(\cdot \mid i_a) \| (P_{h'}^a)^{d_a}(\cdot \mid i_a)),
\end{aligned} \tag{3.161}$$

where $\tilde{E}_h$ in (3.161) denotes expectation under the measure $\tilde{P}_h$. We claim that under the measure $\tilde{P}_h$, the collection $\{(\underline{d}(k), \underline{i}(k)) : k \geq \lfloor n\epsilon'' \rfloor + 1\}$ is a Markov process. Indeed, for all $k \geq \lfloor n\epsilon'' \rfloor + 1$,

$$\begin{aligned}
&\tilde{P}_h(\underline{d}(k+1) = \underline{d}', \underline{i}(k+1) = \underline{i}' \mid (\underline{d}(t), \underline{i}(t)),\ \lfloor n\epsilon'' \rfloor + 1 \leq t \leq k) \\
&= \begin{cases}
\left(\frac{\eta}{K} + (1-\eta)\,\lambda_{h,\delta}(a|\underline{d}(k), \underline{i}(k))\right)\ (P_h^a)^{d_a(k)}(i_a'|i_a(k)), & \text{if } d_a' = 1,\ d_b' = d_b(k) + 1 \text{ for all } b \neq a, \\
& i_b' = i_b(k) \text{ for all } b \neq a, \\
0, & \text{otherwise.}
\end{cases}
\end{aligned} \tag{3.162}$$

Fix an integer $M >> K-1$ such that (3.54) holds, and let $\underline{d}' = (K, K-1, \ldots, 1)$ and $\underline{i}' = (i, \ldots, i)$ for some $i \in \mathcal{S}$. In what follows, we demonstrate that $(\underline{d}', \underline{i}')$ may be reached starting from any $(\underline{d}, \underline{i})$, with a strictly positive probability. Indeed, given any $(\underline{d}, \underline{i})$, assume that the Markov process $\{(\underline{d}(t), \underline{i}(t) : t \geq K)\}$ is in the state $(\underline{d}, \underline{i})$ at some time $T_0 \geq \lfloor n\epsilon'' \rfloor + 1$. Consider the following sequence of arm selections and observations: pull arm 1 at time $t = T_0 + 1$, arm 2 at time $t = T_0 + 2$ and so on until arm K at time $t = T_0 + K$. Thereafter, pull arm 1 at time $t = T_0 + M + 1$ and observe it in state i. Pull arm 2 at time $t = T_0 + M + 2$ and observe it in state i. Continuing this way, finally pull arm K at time $t = T_0 + M + K$ and observe it in state i. Notice that we do not specify the states of the arms as observed at times $T_0 + 1, \ldots, T_0 + K$. For computational purposes, let these states be $s_1, \ldots, s_K$ from arms $1, \ldots, K$ respectively.

Clearly, at time $t = T_0 + M + 2K + 1$, we have $\underline{d}(t) = \underline{d}'$, $\underline{i}(t) = \underline{i}'$, and

$$\begin{aligned}&\tilde{P}_h(\underline{d}(T_0 + M + 2K + 1) = \underline{d}', \underline{i}(T_0 + M + 2K + 1) = \underline{i}' \mid \underline{d}(T_0) = \underline{d}, \underline{i}(T_0) = \underline{i}) \\ &\geq \left(\frac{\eta}{K}\right)^{2K} \cdot \left[\prod_{a=1}^{K} (P_h^a)^M (i|s_a)\right]. \end{aligned} \tag{3.163}$$

Denoting the right-hand side of (3.163) by α, and noting that $\alpha > 0$ and independent of $(\underline{d}, \underline{i})$, we have

$$(\tilde{P}_h)^{M+2K+1}((\underline{d}'', \underline{i}'') \mid \underline{d}, \underline{i}) \geq \alpha \, \mathbb{I}_{\{(\underline{d}'', \underline{i}'') = (\underline{d}', \underline{i}')\}} \quad \text{for all } (\underline{d}, \underline{i}),\ (\underline{d}'', \underline{i}'') \in \mathbb{S}. \tag{3.164}$$

The condition in (3.164) is referred to as the "Doeblin's minorisation condition" [42, Eq. (5)]. Noting that (a) the Markov process $\{(\underline{d}(k), \underline{i}(k)) : k \geq \lfloor n\epsilon'' \rfloor + 1\}$ is ergodic under the measure $\tilde{P}_h$, with $\mu^{\lambda_{h,\delta}}$ as its unique stationary distribution, (b) (3.164) holds, and (c) the increment $\Delta Z_{hh'}(k)$ is almost surely bounded for each k as demonstrated in (3.127), we apply [42, Theorem 1] to deduce that the second probability term in (3.155) is bounded above exponentially. This establishes the lemma. □

3.9 Appendix

3.9.1 An Infinite-Dimensional Linear Programming Problem

In order to better appreciate the usefulness of taking into account the arm delays and last observed states of *all* the arms in deriving the lower bound, we present below a proof sketch of a possibly weaker lower bound in which we first fix an arm a and consider only its delay and last observed state in the subsequent calculations. Fix arm $a \in \mathcal{A}$. Given an integer $d \geq 1$, $i, j \in \mathcal{S}$ and a policy π, let

$$N(\tau(\pi), d, i, a, j) \coloneqq \sum_{t=K}^{\tau(\pi)} \mathbb{I}_{\{d_a(t)=d, i_a(t)=i, A_t=a, X_t^a=j\}}. \tag{3.165}$$

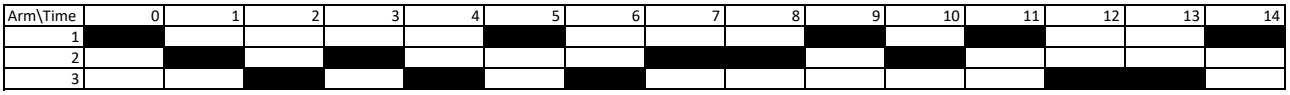


Figure 3.1: A schematic representation of arm selections over time for $K = 3$ arms. In this schematic, an arm selected at any given time is indicated by a black box. Note that arm 1 is selected at time $t = 0$, arm 2 at time $t = 1$ and arm 3 at time $t = 2$. Thereafter, for $t \geq 3$, arm 1 is selected at certain time instants and is not selected at certain other time instants. Whenever arm 1 is not selected, *some* other arm is selected, as a consequence of which the delay of arm 1 increases, and it is this fact that must be captured as a constraint on the delays of arm 1. Similar constraints apply for each of the other arms.

Recall that $\tau(\pi)$ denotes the stopping time of policy π. Following the earlier approaches of [4, 5, 6, 36], and using the data processing inequality, one arrives at[1]

$$\sum_{a=1}^{K} \sum_{d=1}^{\infty} \sum_{i \in \mathcal{S}} E_h[N(\tau(\pi), d, i, a)] \, D((P_h^a)^d(\cdot|i) \| (P_{h'}^a)^d(\cdot|i)), \tag{3.166}$$

where $N(\tau(\pi), d, i, a)$ in (3.166) is simply the summation over all $j \in \mathcal{S}$ of the right-hand side of (3.165).

From the exposition in Section 3.2, we know that at any given time $t \geq K$, the vector $\underline{d}(t)$ must satisfy the following constraint: exactly one component of $\underline{d}(t)$ is equal to 1, and all the other components are strictly greater than 1. Let us now express this constraint mathematically. Recall the assumption that the policy π selects, without loss of generality, arm 1 at time $t = 0$, arm 2 at time $t = 1$ and so on until arm K at time $t = K - 1$. From time $t = K$ onwards, arm a may or may not be selected at all time instants, and whenever it is not selected, *some* arm $b \neq a$ is selected. It is this observation (that some arm is selected at every time instant until the stopping time of the policy) that must be modelled as a constraint mathematically. Figure 3.1 depicts the selection of arms at various time instants for the case when $K = 3$.

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

$$(a-1) + \sum_{i \in \mathcal{S}} \sum_{d=1}^{\infty} d \, N(\tau(\pi), d, i, a) + 1 = \tau(\pi) + 1; \tag{3.167}$$

in (3.167), the term $(a-1)$ on the left-hand side denotes the number of time instants that have

[1]For the gentle reader interested in the details, this can be obtained by following the chain of equalities leading up to (3.82) in Section 3.8.2, with the inner summation over $(\underline{d}, \underline{i})$ now replaced by a summation over $(d, i) \in \{1, 2, \ldots\} \times \mathcal{S}$.

passed before arm a is selected for the first time. The second term on the left-hand side of (3.167) denotes the total number of time instants that have passed, starting from time $t = K$, until the final selection time instant of arm a. The last term on the left-hand side of (3.167) counts the final selection instant of arm a. Thus, the total value of the left-hand side of (3.167) is equal to the total number of time instants that have passed from $t = 0$ to $t = \tau(\pi)$ (both inclusive), which is precisely the quantity on the right-hand side of (3.167). Applying $E_h[\cdot]$ to both sides of (3.167), and using the monotone convergence theorem, we arrive at the following relation after some rearrangement:

$$\sum_{i \in \mathcal{S}} \sum_{d=1}^{\infty} d \, \frac{E_h[N(\tau(\pi), d, i, a)]}{E_h[\tau(\pi)]} + \frac{a-1}{E_h[\tau(\pi)]} = 1. \tag{3.168}$$

In fact, it is easy to see that (3.167), and therefore (3.168), holds for every arm, whether or not the arm is selected at time $t = \tau(\pi)$. Mimicking the steps in Section 3.8.2, and using (3.166) in place of (3.82) in Section 3.8.2 along with the constraint in (3.168), we arrive at the following relation in place of (3.87):

$$\begin{aligned} d(\epsilon, 1-\epsilon) \leq \sup_{\kappa} \min_{h' \neq h} \Bigg\{ & E_h \left[\sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)} \right] \\ & + \Big(E_h[\tau(\pi) - K + 1] \Big) \cdot \sum_{a=1}^{K} \sum_{d=1}^{\infty} \sum_{i \in \mathcal{S}} \kappa(d, i, a) \, D((P_h^a)^d(\cdot|i) \| (P_{h'}^a)^d(\cdot|i)) \Bigg\}, \end{aligned} \tag{3.169}$$

where $d(\epsilon, 1-\epsilon)$ is the relative entropy between a Bernoulli distribution with parameter ϵ and a Bernoulli distribution with parameter $1-\epsilon$, and the supremum in (3.169) is over all probability distributions κ on $\{1, 2, \ldots\} \times \mathcal{S} \times \mathcal{A}$ that satisfy the constraint

$$\sum_{i \in \mathcal{S}} \sum_{d=1}^{\infty} d \, \kappa(d, i, a) = 1 \quad \text{for all } a \in \mathcal{A}. \tag{3.170}$$

The constraint in (3.170) may be obtained from (3.168) by letting $E_h[\tau(\pi)] \to \infty$ (which is the same as $\epsilon \downarrow 0$) and replacing the fractional term on the left-hand side of (3.168) by $\kappa(d, i, a)$; here, $\kappa(d, i, a)$ represents the long-term joint probability of observing arm a to have a delay d and last observed state i, and subsequently selecting arm a.

Dividing both sides of (3.169) by $d(\epsilon, 1-\epsilon)$, and using the fact that $d(\epsilon, 1-\epsilon)/\log(1/\epsilon) \to 1$

as $\epsilon \downarrow 0$, we arrive at

$$\liminf_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} \geq \frac{1}{R_1^*(P_1, P_2)}, \tag{3.171}$$

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

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

subject to

$$\begin{aligned}
&\sum_{i \in \mathcal{S}} \sum_{d=1}^{\infty} d\, \kappa(d, i, a) = 1 \quad \text{for all } a \in \mathcal{A}, \\
&\sum_{d=1}^{\infty} \sum_{i \in \mathcal{S}} \sum_{a=1}^{K} \kappa(d, i, a) = 1, \\
&\kappa(d, i, a) \geq 0 \quad \text{for all } d \in \{1, 2, \ldots\},\ i \in \mathcal{S},\ a \in \mathcal{A}.
\end{aligned} \tag{3.172}$$

Notice that (3.172) constitutes an infinite-dimensional linear programming problem with linear constraints. It is not clear if there exists κ that (a) satisfies the constraints in (3.172) and (b) attains the supremum in the expression for $R_1^*(P_1, P_2)$. Also, it is not clear if the constraints in (3.172) constitute the tightest set of constraints. From Proposition 7, we must of course have $R_1^*(P_1, P_2) \geq R^*(P_1, P_2)$.

We end with a remark that by taking into account the delays and the last observed states of all the arms in deriving the lower bound, as done in Section 3.8.2, the constraint in (3.167) is automatically captured since any vector $\underline{d} = (d_1, \ldots, d_K)$ of arm delays belongs, by definition, to the subset $\mathbb{S}$ which obeys the constraint in (3.167). Thus, the viewpoint of controlled Markov processes greatly simplifies the analysis of the lower bound. The key insight of this chapter is that our 'lift' approach of considering the arm delays and the last observed states of all the arms jointly, instead of dealing with the delays and last observed states of each arm separately, makes the problem amenable to analysis.

3.9.2 Restriction to SRS Class Suffices

An important step in the derivation of the lower bound (3.11) presented in Section 3.8.2 is the replacement of the supremum over the set $\Pi(\epsilon)$ appearing in (3.87) to the set Π_{SRS} of all SRS policies (compare the right hand side of (3.87) with that of (3.91)). Here, $\Pi(\epsilon)$, which the set of all policies whose probability of error at stoppage is at most ϵ, may potentially include non-SRS policies too. The aforementioned step in the proof of the lower bound is possible thanks to the following theorem which is an analogue of [40, Theorem 8.8.2] for countable state space

controlled Markov processes. We omit the proof of the theorem as it follows straightforwardly from the proof of [40, Theorem 8.8.2].

Recall that for each $\pi^\lambda \in \Pi_{\mathsf{SRS}}$, the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is, in fact a Markov process. Furthermore, when the trembling hand parameter $\eta > 0$, this Markov process is ergodic (Lemma 9).

Theorem 3. *1. For each $\pi^\lambda \in \Pi_{\mathsf{SRS}}$, $(\underline{d}, \underline{i}) \in \mathbb{S}$ and $a \in \mathcal{A}$, let*

$$\nu^\lambda(\underline{d}, \underline{i}, a) = \mu^\lambda(\underline{d}, \underline{i}) \left(\frac{\eta}{K} + (1-\eta)\ \lambda(a \mid \underline{d}, \underline{i}) \right), \tag{3.173}$$

where $\eta > 0$ is the trembling hand parameter and μ^λ is the unique stationary distribution of the Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ under the SRS policy π^λ. Then, ν^λ is a feasible solution to (3.88)-(3.90).

2. Let ν be any feasible solution to (3.88)-(3.90). Then, for each $(\underline{d}, \underline{i}) \in \mathbb{S}$, $\sum_{a=1}^{K} \nu(\underline{d}, \underline{i}, a) > 0$. Let λ^* be such that

$$\frac{\eta}{K} + (1-\eta)\ \lambda^*(a \mid \underline{d}, \underline{i}) := \frac{\nu(\underline{d}, \underline{i}, a)}{\sum_{a=1}^{K} \nu(\underline{d}, \underline{i}, a)}, \quad (\underline{d}, \underline{i}) \in \mathbb{S},\ a \in \mathcal{A}.$$

Then, ν^{λ^*} is a feasible solution to (3.88)-(3.90) and

$$\nu^{\lambda^*}(\underline{d}, \underline{i}, a) = \nu(\underline{d}, \underline{i}, a) \quad \text{for all } (\underline{d}, \underline{i}) \in \mathbb{S} \text{ and } a \in \mathcal{A}.$$

3.10 Summary

We make several concluding remarks in summary.

1. From (3.30), when the trembling hand parameter $\eta > 0$, we see that

$$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} = \frac{1}{R^*_\eta(P_1, P_2)}. \tag{3.174}$$

   We have thus provided an answer to (3.8) on the minimum growth rate of the expected time to identify the odd arm location as $\epsilon \downarrow 0$.

2. The asymptotically optimal $\lambda(\cdot|\cdot)$ in the restless case may depend on the history unlike that in the prior works [4, 5, 6, 36] where $\lambda(\cdot)$ did not depend on history, even in the

rested Markov case. At first glance, this is surprising for the rested Markov case, but in retrospect, these features are apparent from an examination of the optimisation problem (3.12) in these special cases.

3. Computability of $R^*_\eta(P_1, P_2)$ may be an issue, and one must usually resort to Q-learning for restless Markov arms [37] to arrive at good policies. The fact that $D(P_k^{d_a}(\cdot|i_a)||P_l^{d_a}(\cdot|i_a))$, $k, l \in \{1, 2\}$, converges as $d_a \to \infty$ could enable restriction of the countable state space $\mathbb{S}$ to a finite set, and could lead to good approximations.

4. When the trembling hand parameter $\eta > 0$, the ergodicity of the Markov process $(\underline{d}(t), \underline{i}(t))$ under any SRS policy ensures that time averages approach the ensemble averages. This is crucial to show achievability. Note also the use of uniqueness of the stationary distribution to show the converse. The trembling hand model may be viewed as a *regularisation* that gives stability of the aforementioned Markov process for free. If the trembling hand parameter η were 0, one could deliberately add some regularisation parameterised by η, and let this parameter $\eta \downarrow 0$. $R^*_0(P_1, P_2)$ governs the lower bound, whereas $\lim_{\eta \downarrow 0} R^*_\eta(P_1, P_2)$ governs the upper bound. The resulting lower and upper bounds on the growth rate may have a gap.

Chapter 4

Restless Arms with TPMs Unknown

4.1 Preamble

In the previous chapter, we analysed the setting of restless arms in the case when the TPMs of the arms are known beforehand. In this chapter, we extend the results of the previous chapter to the more difficult case when the TPMs are unknown. That is, given a multi-armed bandit with $K \geq 3$ arms and an arms configuration $C = (h, P_1, P_2)$ in which h is the index of the odd arm, the TPM of arm h is P_1, and the TPM of each of the remaining arms is $P_2 \neq P_1$, we wish to characterise

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

for the setting of restless arms when P_1 and P_2 are unknown and must be learnt along the way.

As in the previous chapter, we assume that the decision entity has a *trembling hand*. This means that for some fixed $\eta > 0$, the decision entity samples the intended arm with probability $1 - \eta$, but with probability η, the decision entity samples a uniformly randomly chosen arm. Here, η is known as the trembling hand parameter. See Chapter 2 and the previous chapter for a description of how trembling arises in visual neuroscience experiments. It is likely that in such visual search experiments, the human subject scans multiple images at once before narrowing down the search to the oddball image [43]. While the trembling hand model of the previous chapter does not model such nuances, it broadly captures the search dynamics in a way that makes the problem amenable to analysis.

As described in the previous chapter, when the arms are restless, it is necessary for the decision entity to maintain a record of (a) the time elapsed since each arm was previously sampled (called the arm's *delay*), and (b) the state of each arm as observed at its previous sampling instant (called the arm's *last observed state*). Further, as demonstrated in Section 3.2

of the previous chapter, the delays and the last observed states of the arms collectively form a *controlled Markov process* and, in turn, lead to a Markov decision problem (MDP) whose state space is countably infinite and action space is the set of arms. Also, the transition probabilities of this MDP are stationary across time and are functions of the odd arm index, P_1 and P_2. However, the objective is not to maximise rewards (or minimise regret), as is typical in MDPs, but to find the odd arm index quickly and accurately.

4.1.1 Certainty Equivalence and Identifiability

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

Sufficient conditions that lead to system identifiability have been proposed in the literature. In an important paper [44], Mandl demonstrated that for a finite-state MDP, when the parameter estimate at each time t is chosen so as to minimise a "contrast" function computed using all the observations and actions up to time t, the parameter estimates converge to the true parameter value as $t \to \infty$ [44, Theorem 6]. In particular, when the contrast function is the negative log-likelihood, the resulting parameter estimates are the maximum likelihood (ML) estimates. Mandl demonstrated the convergence of the ML estimates to the true parameter value under an additional condition (known as *Mandl's identifiability condition* [44, Eq. (35)]) on the MDP transition probabilities. However, it is not clear if Mandl's identifiability condition is sufficient for identifiability in countable-state MDPs. In [45], the authors consider the same problem as Mandl's, but for countable-state MDPs when Mandl's identifiability condition is relaxed. The authors of [45] show that under some regularity assumptions on the MDP transition probabilities, certainty equivalence based on ML estimation renders the system identifiable.

In this chapter, we use certainty equivalence with ML estimation as in [45] to learn the true arms configuration. Due to the presence of arm delays in the likelihood function, closed-form expressions for the ML estimates of the TPMs are not available. Nevertheless, we show that

the system is identifiable under a mild regularity assumption on the TPMs.

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

The principle of certainty equivalence seems to have been rigorously explored first in the context of linear systems by Åström and Wittenmark [46] where it has been referred to as a *self-tuning regulator*. The paper by Mandl [44] applies the principle of certainty equivalence to the problem of adaptively controlling a finite-state controlled Markov process (equivalently, a finite-state MDP). Mandl demonstrated that under a regularity condition (known as Mandl's identifiability condition [44, Eq. (35)]) on the MDP transition probabilities, certainty equivalence based on ML estimation renders the system identifiable. Doshi and Shreve [47] consider a problem similar to Mandl's and show that under a slightly weaker condition than Mandl's, identifiability holds for a scheme based on certainty equivalence with modified maximum likelihood estimates (estimates that nearly maximise the log-likelihood).

It is not clear if either Mandl's identifiability condition or its weaker version in [47] is sufficient for identifiability in infinite-state MDPs. Also, as remarked in [48], Mandl's condition may be too restrictive in practical applications. The authors of [48] provide an example of a linear, real-valued Markovian system for which Mandl's identifiability condition fails to hold. For this example, the authors of [48] demonstrate that the parameter estimates may not converge, and even if they do, the convergence is not necessarily to the true parameter value. This suggests that in order to show system identifiability in infinite-state MDPs, it may be necessary to impose additional regularity conditions on the MDP transition probabilities.

One such set of such regularity conditions that ensures identifiability in countable-state MDPs when Mandl's condition is relaxed may be found in [45]. We show that the regularity conditions of [45] are satisfied for the problem studied in this chapter under a mild assumption on the unknown TPMs of the odd arm and the non-odd arm Markov processes. For a detailed survey of results in stochastic adaptive control of Markov chains, see [49].

4.1.3 A Brief Overview of Our Contributions

Below, we highlight our contributions and bring out the challenges that we need to overcome in the analysis of the setting of restless arms when the TPMs of the arms are not known beforehand.

1. We derive an asymptotic lower bound on the growth rate of the expected time required to find the odd arm index subject to an upper bound on the error probability, where the asymptotics is as the error probability vanishes. Specifically, given an arms configuration

$C = (h, P_1, P_2)$ in which h is the odd arm index, P_1 is the TPM of arm h, and $P_2 \neq P_1$ is the common TPM of each non-odd arm, we show that the lower bound is $1/R^*(h, P_1, P_2)$, where $R^*(h, P_1, P_2)$ is an arms configuration dependent (or problem instance-dependent) constant and is the value of a max-min optimisation problem. See Section 4.3, equation (4.12) for the exact mathematical expression for $R^*(h, P_1, P_2)$. The max-min expression in (4.12) consists of an outer supremum over all conditional probability distributions on the arms conditioned on the arm delays and the last observed states, and an inner infimum over all arms configurations in which the odd arm index is any $h' \neq h$.

2. Given an arms configuration $C = (h, P_1, P_2)$, the question of whether there exists an optimal conditional distribution that attains the outer supremum in the expression for $R^*(h, P_1, P_2)$ is still under study. Notwithstanding this, we look at conditional distributions that attain the supremum to within a factor of $1/(1 + \delta)$ for any $\delta > 0$. We refer to such conditional distributions as δ-optimal solutions for the arms configuration $C = (h, P_1, P_2)$, and note that there may be multiple such solutions in general for each arms configuration.

   It is worth noting here that in the prior works [6, 36], the outer supremum in the expression for the constant appearing in the lower bound is over all *unconditional* probability distributions on the arms which are simpler objects to deal with than conditional probability distributions. Further, in these works, this outer supremum is attained by a unique (unconditional) probability distribution on the arms.

3. An important property of the optimal solutions to the lower bounds in [6, 36] is that they are continuous functions of the arms configuration. It is this continuity property that is used to show that the certainty equivalence-based arm sampling policies of these works (using ML estimation) render the system identifiable. This, in turn, is used to show that these policies achieve the respective lower bounds asymptotically. Furthermore, a close look at [6, 36] reveals that the ML estimates in these works have closed-form expressions. This readily helps in showing the convergence of the ML estimates to their true values by using either the law of large numbers (for the case of iid observations from the arms in [6, 36]) or the ergodic theorem (for the case of Markov observations from the arms analysed in Chapter 2).

   In the context of this chapter, we note that neither the existence of unique optimal solutions to $R^*(h, P_1, P_2)$ nor the above mentioned continuity property is available. Furthermore, the presence of arm delays in the likelihood function poses difficulty in obtaining

closed-form expressions for the ML estimates of the TPMs of the odd arm and the non-odd arm Markov processes As a result, showing that the ML estimates converge to their true values is a challenging task.

4. Notwithstanding the difficulties highlighted in the previous point, we demonstrate that under two key assumptions, certainty equivalence based on ML estimation renders the system identifiable. The first of these assumptions is on the existence of a continuous selection of δ-optimal solutions (the analogue of continuity of optimal solutions in the prior works). The second assumption, analogous to that appearing in [48], is a mild regularity on the TPMs of the odd arm and the non-odd arm Markov processes that requires the non-zero entries of the powers of the TPMs to be larger than a certain constant $\bar{\varepsilon}^* \in (0, 1)$.

5. Given an error probability $\epsilon > 0$ and $\delta > 0$, we construct a policy based on the principle of certainty equivalence with ML estimation and using the δ-optimal solutions. Under the assumptions mentioned in the previous point, we demonstrate that the policy stops in finite time almost surely and that the error probability of the policy at stoppage is upper bounded by ϵ. Further, we show that as $\epsilon \downarrow 0$, the policy's expected time to find the odd arm index satisfies an upper bound that is away from the lower bound only by a multiplicative factor of $(1+\delta)^2$. Our achievability analysis relies crucially on (a) resolving the identifiability problem for the countable-state MDP arising from the arm delays and the last observed states, and (b) showing that the policy eventually samples the arms according to the δ-optimal solution for the underlying arms configuration. Finally, by letting $\delta \downarrow 0$, we show that our policy meets the lower bound.

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

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

However, unlike the policy we propose in Section 4.4, repeated sampling of the arms may not ensure that given $\delta > 0$ and an underlying arms configuration C, the arms are eventually sampled according to the δ-optimal solution for C, a condition that is crucial in order to meet the constant in the lower bound. Moreover, it is not clear if the desired error probability criterion can be met. While policies that sample the arms repeatedly have been proposed and shown to perform well for the problem of minimising regret [28], it is not clear if such policies perform well for optimal stopping problems such as that studied in this chapter.

4.1.4 Chapter Organisation

The rest of this chapter is organised as follows. In Section 4.2, we set up the notations and provide some preliminaries on MDPs that will be used in the remainder of the chapter. In Section 4.3, we present the asymptotic lower bound on the growth rate of the expected time to identify the odd arm index. In Section 4.4, we state the two assumptions, present a policy based on certainty equivalence with ML estimation, and demonstrate that it achieves the lower bound asymptotically. We state the main result of this chapter in Section 4.5 and provide some concluding remarks in Section 4.7. The proofs of all the results are contained in Section 4.6.

4.2 Notations and Preliminaries

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

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

where A_t and B_t satisfy the probabilistic relation

$$P(A_t = a|B_t = b) = \frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{a=b\}} \tag{4.1}$$

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

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

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

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

$$\Pi(\epsilon) \coloneqq \left\{\pi : P^{\pi}\Big(\theta(\tau(\pi)) \neq h \;\Big|\; C = (h, P_1, P_2)\Big) \leq \epsilon \quad \forall\, C = (h, P_1, P_2),\ h \in \mathcal{A},\ P_2 \neq P_1\right\}. \tag{4.2}$$

In (4.2) and throughout the chapter, $P^{\pi}(\cdot|C)$ denotes probabilities computed under the policy π and the arms configuration C. Similarly, $E^{\pi}[\cdot|C]$ will be used to denote expectations.

Note the requirement of policies in (4.2) to work for all arms configurations. A careful reader may raise the question that the error probability criterion may not hold for arms configurations $C = (h, P_1, P_2)$ in which the entries of P_1 and P_2 are arbitrarily close to each another, with some entries of P_1 possibly matching with those of P_2. However, the key here is to note that

for such arms configurations, a policy $\pi \in \Pi(\epsilon)$ might have to wait longer before it stops and declares the odd arm index with an error probability $\leq \epsilon$. Therefore, "closer" P_1 and P_2 are, the harder they are to distinguish, which results in larger stopping times of the policies.

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

$$\inf_{\pi \in \Pi(\epsilon)} E^{\pi}[\tau(\pi)|C] = \Theta(\log(1/\epsilon)). \tag{4.3}$$

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

4.2.1 Arm Delays and Last Observed States

Due to the restless nature of the arms, the decision entity has to keep a record of (a) the time elapsed since an arm was previously selected (the arm's *delay*), and (b) the state of the arm at its previous selection time instant (the arm's *last observed state*). As noted in Section 3.2 of Chapter 3, arm delays and the last observed states are striking features of the setting of restless arms, and are superfluous when each arm yields independent and identically distributed (iid) observations or when each arm yields Markov observations and the arms are rested.

Following the notations in Chapter 3 , let $d_a(t)$ and $i_a(t)$ respectively denote the delay and the last observed state of arm $a \in \mathcal{A}$ at time t. Let $\underline{d}(t) = (d_a(t) : a \in \mathcal{A})$ and $\underline{i}(t) = (i_a(t) : a \in \mathcal{A})$ denote the vectors of arm delays and last observed states at time t. The arm delays and the last observed states make sense when each arm is sampled at least once, and shall therefore be defined for $t \geq K$ keeping in mind that $A_0 = 1, \ldots, A_{K-1} = K$ under every policy. For $t = K$, we set $\underline{d}(K) = (K, K-1, \ldots, 1)$. This means that with reference to time $t = K$, arm 1 was sampled K time instants earlier (i.e., at $t = 0$), arm 2 was sampled $K - 1$ time instants earlier (i.e., at $t = 1$) and so on. The rule for updating $(\underline{d}(t), \underline{i}(t))$, based on the value of A_t, is straightforward and is as follows: if $A_t = a$, then

$$d_{\tilde{a}}(t+1) = \begin{cases} d_{\tilde{a}}(t) + 1, & \tilde{a} \neq a, \\ 1, & \tilde{a} = a, \end{cases} \qquad i_{\tilde{a}}(t+1) = \begin{cases} i_{\tilde{a}}(t), & \tilde{a} \neq a, \\ \bar{X}_t, & \tilde{a} = a, \end{cases} \tag{4.4}$$

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

Therefore, it follows that under any policy π, the decision entity first samples each of the K arms in such a way that $A_0 = 1$, $A_1 = 2$ and so on until $A_{K-1} = K$. Then, for all $K \leq t < \tau(\pi)$, based on (a) the history $\{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}$ of arm delays and last observed states, and (b) the history $\{B_s : K \leq s < t\}$ of all the intended arm samples, the decision maker chooses to sample the arm B_t. Notice that this is equivalent to sampling B_t based the history $(B^{t-1}, A^{t-1}, \bar{X}^{t-1})$. Due to the trembling hand, the decision entity observes that arm A_t is instead pulled at time t. Subsequently, the decision entity observes the state $\bar{X}_t$ of arm A_t, and updates $(\underline{d}(t), \underline{i}(t))$ to $(\underline{d}(t+1), \underline{i}(t+1))$ based on the update rule in (4.4). At time $t = \tau(\pi)$, the decision entity announces its estimate $\theta(\tau(\pi))$ of the odd arm index.

4.2.2 Controlled Markov Process and the Associated Markov Decision Problem

From Chapter 3, we know that $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is a *controlled Markov process* with $(\underline{d}(t), \underline{i}(t))$ regarded as the state at time t and B_t regarded as the control at time t. That is, we are in the setting of a Markov decision problem (MDP) whose state space is $\mathbb{S}$, action space is $\mathcal{A}$, and the transition probabilities under an arms configuration C are as follows: for all $(\underline{d}', \underline{i}'), (\underline{d}, \underline{i}) \in \mathbb{S}$ and $b \in \mathcal{A}$,

$$\begin{aligned} &P^{\pi}(\underline{d}(t+1) = \underline{d}', \underline{i}(t+1) = \underline{i}' \mid \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, B_t = b, C) \\ &= \sum_{a=1}^{K} \left(\frac{\eta}{K} + (1-\eta)\mathbb{I}_{\{a=b\}} \right) (P_C^a)^{d_a}(i'_a | i_a)\, \mathbb{I}_{\{d'_a = 1 \text{ and } d'_{\tilde{a}} = d_{\tilde{a}} + 1 \text{ for all } \tilde{a} \neq a\}}\, \mathbb{I}_{\{i'_{\tilde{a}} = i_{\tilde{a}} \text{ for all } \tilde{a} \neq a\}}. \end{aligned} \tag{4.5}$$

In (4.5), P_C^a is the TPM of arm a under the arms configuration C. For instance, if $C = (h, P_1, P_2)$, then

$$P_C^a = \begin{cases} P_1, & a = h, \\ P_2, & a \neq h. \end{cases} \tag{4.6}$$

Also, d'_a and i'_a denote the component corresponding to arm a in the vectors $\underline{d}'$ and $\underline{i}'$ respectively. Similarly, $d_{\tilde{a}}$ and $i_{\tilde{a}}$ denote the component corresponding to arm $\tilde{a}$ in the vectors $\underline{d}$ and $\underline{i}$ respectively. Notice that the transition probabilities in (4.5) are stationary across time and parameterised by the arms configuration. We write $Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b)$ as a short hand

representation of the transition probabilities in (4.5).

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

$$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^{\pi}[\tau(\pi)|C]}{\log(1/\epsilon)}. \tag{4.7}$$

4.2.3 SRS Policies and State-Action Occupancy Measures

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

$$\lambda = \bigotimes_{(\underline{d},\underline{i}) \in \mathbb{S}} \lambda(\cdot|\underline{d},\underline{i}), \tag{4.8}$$

with $\lambda(\cdot|\underline{d},\underline{i})$ being a probability measure on $\mathcal{A}$ for all $(\underline{d},\underline{i}) \in \mathbb{S}$, such that B_t is sampled according to $\lambda(\cdot|\underline{d}(t),\underline{i}(t))$ for all $t \geq K$. Let such an SRS policy be denoted more explicitly as π^{λ}, and let Π_{SRS} be the space of all SRS policies. Clearly, $\{(\underline{d}(t),\underline{i}(t)) : t \geq K\}$ is a Markov process under every SRS policy. Further, Lemma 9 of Chapter 3 shows that this Markov process is ergodic when the trembling hand parameter $\eta > 0$, and therefore possess a unique stationary distribution. Let $\mu^{\lambda} = \{\mu^{\lambda}(\underline{d},\underline{i}) : (\underline{d},\underline{i}) \in \mathbb{S}\}$ be the stationary distribution under π^{λ}. Also, for $(\underline{d},\underline{i}) \in \mathbb{S}$ and $a \in \mathcal{A}$, let

$$\nu^{\lambda}(\underline{d},\underline{i},a) \coloneqq \mu^{\lambda}(\underline{d},\underline{i})\left(\frac{\eta}{K} + (1-\eta)\,\lambda(a|\underline{d},\underline{i})\right) \tag{4.9}$$

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

4.3 Converse: Lower Bound

In this section, we present a lower bound for (4.7) when. Given two probability distributions μ and ν on $\mathcal{S}$, the Kullback-Leibler (KL) divergence (also called the *relative entropy*) between μ and ν is defined as

$$D(\mu\|\nu) \coloneqq \sum_{i \in \mathcal{S}} \mu(i) \log \frac{\mu(i)}{\phi(i)}, \tag{4.10}$$

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

[1] Writing $\mathcal{P}(\mathcal{A})$ to denote the space of all probability distributions on $\mathcal{A}$, it follows that (4.8) is an element of the product space $\bigotimes_{(\underline{d},\underline{i}) \in \mathbb{S}} \mathcal{P}(\mathcal{A})$.

Proposition 11. *Under the arms configuration $C = (h, P_1, P_2)$,*

$$\liminf_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^{\pi}[\tau(\pi)|C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h, P_1, P_2)}, \tag{4.11}$$

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

$$R^*(h, P_1, P_2) := \sup_{\pi^\lambda \in \Pi_{SRS}} \inf_{\substack{C'=(h', P_1', P_2'): \\ h' \neq h,\ P_1' \neq P_2'}} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu^\lambda(\underline{d}, \underline{i}, a)\ k_{CC'}(\underline{d}, \underline{i}, a), \tag{4.12}$$

with

$$\begin{aligned} k_{CC'}(\underline{d}, \underline{i}, a) &:= D((P_C^a)^{d_a}(\cdot|i_a) \| (P_{C'}^a)^{d_a}(\cdot|i_a)) \\ &= \begin{cases} D(P_1^{d_a}(\cdot|i_a) \| (P_2')^{d_a}(\cdot|i_a)), & a = h, \\ D(P_2^{d_a}(\cdot|i_a) \| (P_1')^{d_a}(\cdot|i_a)), & a = h', \\ D(P_2^{d_a}(\cdot|i_a) \| (P_2')^{d_a}(\cdot|i_a)), & a \neq h, h'. \end{cases} \end{aligned} \tag{4.13}$$

The infimum in (4.12) is over all alternative odd arm configurations $C' = (h', P_1', P_2')$ satisfying (a) $h' \neq h$, and (b) $P_1' \neq P_2'$.

Proof. See Section 4.6.1. □

Notice that closer the TPMs P_1 and P_2 are (in terms of relative entropy), the smaller is the value of $R^*(h, P_1, P_2)$, resulting in a larger value of the lower bound (4.11). The key ingredients in the proof of the lower bound are (a) a data processing inequality for the setting of restless arms based on a change of measure argument presented in [32], (b) a Wald-type lemma for Markov processes, (c) a recognition of the fact that for any $(\underline{d}, \underline{i}) \in \mathbb{S}$, the long-term fraction of exits from $(\underline{d}, \underline{i})$ matches the long-term fraction of entries to $(\underline{d}, \underline{i})$, and (d) restriction of the supremum in (4.12) to the class of SRS policies, which is possible thanks to Theorem 3 of Chapter 3.

4.3.1 Simplifying $R^*(h, P_1, P_2)$

Note that

$$\begin{aligned} &R^*(h, P_1, P_2) \\ &= \sup_{\pi^\lambda \in \Pi_{\mathsf{SRS}}} \inf_{\substack{C'=(h', P_1', P_2'): \\ h' \neq h,\ P_1' \neq P_2'}} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \Big[\nu^\lambda(\underline{d}, \underline{i}, h)\, D(P_1^{d_h}(\cdot|i_h) \| (P_2')^{d_h}(\cdot|i_h)) \end{aligned}$$

$$+ \nu^{\lambda}(\underline{d}, \underline{i}, h')\, D(P_2^{d_{h'}}(\cdot|i_{h'}) \| (P_1')^{d_{h'}}(\cdot|i_{h'})) \\ + \sum_{a \neq h, h'} \nu^{\lambda}(\underline{d}, \underline{i}, a)\, D(P_2^{d_a}(\cdot|i_a) \| (P_2')^{d_a}(\cdot|i_a)) \Bigg]. \quad (4.14)$$

Because P_1' appears only in the second term within the square brackets, it follows that the infimum over all P_1' of this term is equal to zero and may be achieved by setting $P_1' = P_2$. Therefore,

$$R^*(h, P_1, P_2) \\ = \sup_{\pi^{\lambda} \in \Pi_{\mathsf{SRS}}} \inf_{\substack{h', P_2': \\ h' \neq h,\ P_2' \neq P_2}} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \Bigg[\nu^{\lambda}(\underline{d}, \underline{i}, h)\, D(P_1^{d_h}(\cdot|i_h) \| (P_2')^{d_h}(\cdot|i_h)) \\ + \sum_{a \neq h, h'} \nu^{\lambda}(\underline{d}, \underline{i}, a)\, D(P_2^{d_a}(\cdot|i_a) \| (P_2')^{d_a}(\cdot|i_a)) \Bigg]. \quad (4.15)$$

It is worth comparing (4.15) with (2.89), the analogue of (4.15) for the setting of rested arms, which is reproduced below for ease of comparison:

$$R^*_{\mathsf{rested}}(h, P_1, P_2) = \sup_{\lambda \in \mathcal{P}(\mathcal{A})} \inf_{\substack{h', P_2': \\ h' \neq h,\ P_2' \neq P_2}} \Big[\lambda(h)\, D(P_1 \| P_2 | \mu_1) + (1 - \lambda(h) - \lambda(h'))\, D(P_2 \| P_1 | \mu_2) \Big]. \quad (4.16)$$

In the above equation, $\mathcal{P}(\mathcal{A})$ denotes the space of all probability distributions on $\mathcal{A}$, and for $i = 1, 2$, μ_i denotes the stationary distribution of the transition probability matrix P_i. The absence of arm delays in (4.16) makes it possible to simplify this term further by first computing the infimum over all $h' \neq h$ and subsequently computing the optimal value of P_2' (the one that attains the infimum over all $P_2' \neq P_2$) using the method of Lagrange multipliers; see Chapter 2 for the detailed steps.

The presence of ergodic state-action occupancy measures inside the summation in (4.15) does not allow for simplifying the infimum over $P_2' \neq P_2$ first, as was possible in the setting of rested arms. Further, because of the presence of arm delays in the relative entropy terms in (4.15), it may not be possible to obtain a closed-form expression for the choice of P_2' that attains the inner infimum in (4.15).

4.3.2 Near-Optimal Solutions to the Supremum

It is not clear if there exists an optimal SRS policy π^{λ} that attains the supremum in (4.12). However, from (4.12), by the definition of supremum, we know that for each $\delta > 0$, there exists

$\lambda = \lambda_{h,P_1,P_2,\delta}(\cdot|\cdot) \in \bigotimes_{(\underline{d},\underline{i})\in\mathbb{S}} \mathcal{P}(\mathcal{A})$ such that

$$\inf_{\substack{C'=(h',P_1',P_2'):\\ h'\neq h,\ P_1'\neq P_2'}} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \nu^{\lambda}(\underline{d},\underline{i},a)\ k_{CC'}(\underline{d},\underline{i},a) \geq \frac{R^*(h,P_1,P_2)}{1+\delta}. \tag{4.17}$$

Call $\lambda_{h,P_1,P_2,\delta}$ a *δ-optimal solution* for the arms configuration $C = (h, P_1, P_2)$. More generally, let $\lambda_{h,P,Q,\delta}$ denote a δ-optimal solution for $C = (h, P, Q)$. Notice that one or more λ may satisfy (4.17), thus implying that multiple δ-optimal solutions may exist for each arms configuration.

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

4.4 Achievability

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

4.4.1 Two Key Assumptions

Given $\delta > 0$, (4.17) suggests that in order to approach the constant $R^*(h, P_1, P_2)$ in the lower bound to within a factor of $1/(1+\delta)$, the arms must eventually be sampled according to $\lambda_{h,P_1,P_2,\delta}$ or one of the δ-optimal solutions for $C = (h, P_1, P_2)$. Because the unknown underlying arms configuration may be any one among the uncountably infinite collection $\{C = (h, P, Q) : h \in \mathcal{A}, P \neq Q\}$ of all possible arms configurations, a feasible option is to eventually sample the arms according to $\lambda_{h,P,Q,\delta}$ which is "close" to $\lambda_{h,P_1,P_2,\delta}$ when (P, Q) close to (P_1, P_2). We show this works under some regularity conditions.

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

The paper [35] considers a similar assumption as above, but for a more general sequential hypothesis testing problem in multi-armed bandits (see [35, Assumption A]). Also, the analogue of Assumption 1 for the maximisers, instead of δ-optimal solutions, holds in the settings of the prior works [6, 36] as a consequence of Berge's maximum theorem [50]. Henceforth, for each $\delta > 0$, fix a selection $\{\lambda_{h,P,Q,\delta} : h \in \mathcal{A}, P \neq Q\}$ of δ-optimal solutions satisfying Assumption 1.

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

$$\mathscr{P}(\bar{\varepsilon}^*) \coloneqq \{P \in \mathscr{P}(\mathcal{S}) : \ P \text{ is ergodic}, \ \forall\, d \geq 1, \ i,j \in \mathcal{S}, \ P^d(j|i) > 0 \implies P^d(j|i) \geq \bar{\varepsilon}^*\}. \tag{4.18}$$

Eq. (4.18) defines the class of all ergodic $P \in \mathscr{P}(\mathcal{S})$ such that each non-zero entry of P^d is lower bounded by $\bar{\varepsilon}^*$ uniformly in d. Clearly, every ergodic P belongs to $\mathscr{P}(\bar{\varepsilon}^*)$ for some P-dependent $\bar{\varepsilon}^*$. To see this, fix an arbitrary ergodic P, and let $\mu = (\mu(j) : j \in \mathcal{S})$ be the unique stationary distribution for P. From [41, Theorem 4.9], we have $P^d(j|i) \to \mu(j) > 0$ as $d \to \infty$ for all $i,j \in \mathcal{S}$. Let $\mu_{\mathsf{min}} = \min_j \mu(j)$. Then, there exists D such that $\forall\, d \geq D$, each non-zero entry of P^d is lower bounded by $\mu_{\mathsf{min}}/2$. Further, let

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

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

$$\mathcal{C}(\bar{\varepsilon}^*) \coloneqq \{(P,Q) : \ P(\cdot|i) \text{ is mutually absolutely continuous with } Q(\cdot|i) \text{ for all } i \in \mathcal{S}, \\ P \in \mathscr{P}(\bar{\varepsilon}^*), \ Q \in \mathscr{P}(\bar{\varepsilon}^*)\}. \tag{4.19}$$

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

Some remarks are in order. An arms configuration $C = (h,P,Q)$ satisfying Assumption 2 only increases the difficulty of identifying the odd arm index h. If $P, Q \in \mathscr{P}(\bar{\varepsilon}^*)$ for some $\bar{\varepsilon}^* > 0$, then they are harder to "distinguish" from one another. To see this, note that for any ergodic P, Q, we know from [41, Proposition 2.4] that that there exists $M = M(P,Q)$ such that all the entries of P^d and Q^d are strictly positive $\forall\, d \geq M$, thus implying that $\forall\, d \geq M$ and $i \in \mathcal{S}$,

$$D(P^d(\cdot|i)\|Q^d(\cdot|i)) < \infty, \quad D(Q^d(\cdot|i)\|P^d(\cdot|i)) < \infty. \tag{4.20}$$

For $d < M$, it may be the case that one or both of the relative entropy terms in (4.20) equals $+\infty$ and discrimination becomes easier. However, when $P, Q \in \mathscr{P}(\bar{\varepsilon}^*)$, it follows that $\forall\, d \geq 1$, each row of P^d is mutually absolutely continuous with the corresponding row of Q^d. Furthermore, $\forall\, d \geq 1$ and $i,j \in \mathcal{S}$ such that $P^d(j|i) > 0$, $Q^d(j|i) > 0$, the relation

$$\bar{\varepsilon}^* \leq \frac{P^d(j|i)}{Q^d(j|i)} \leq \frac{1}{\bar{\varepsilon}^*} \tag{4.21}$$

holds, thus implying that $\forall d \geq 1$, each of the relative entropy terms in (4.20) is at most $\log(1/\bar{\varepsilon}^*)$. Therefore, $P, Q \in \mathscr{P}(\bar{\varepsilon}^*)$ cannot have an arbitrarily large separation (in terms of relative entropy) and are harder to distinguish from one another.

It is worth noting here that Assumption 2 is equivalent to [48, Assumption I] which, in the context of this chapter, states that there exists $\bar{\epsilon} > 0$ such that for all $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$, $b \in \mathcal{A}$ and $C = (h, P_1, P_2)$,

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

Indeed, we note from (4.5) that if $Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) > 0$, then it must be true that $d'_{a_0} = 1$, $d'_{\tilde{a}} = d_{\tilde{a}} + 1$ for all $\tilde{a} \neq a_0$, and $i'_{\tilde{a}} = i_{\tilde{a}}$ for all $\tilde{a} \neq a_0$ for some $a_0 \in \mathcal{A}$. Then, in this case,

$$Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) = \left(\frac{\eta}{K} + (1-\eta)\mathbb{I}_{\{a_0 = b\}}\right) \; (P_C^{a_0})^{d_{a_0}}(i'_{a_0} | i_{a_0}). \tag{4.22}$$

Because the term within brackets in (4.22) is $\geq \frac{\eta}{K} > 0$, it follows that any lower bound on $Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b)$ implies a lower bound on $(P_C^{a_0})^{d_{a_0}}(i'_{a_0} | i_{a_0})$ and vice-versa. This establishes the equivalence between Assumption 2 and [48, Assumption I].

4.4.2 Test Statistic

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

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

$$\bar{f}(B^n, A^n, \bar{X}^n | \mathcal{H}_h) = \int\limits_{\mathscr{P}(\mathcal{S}) \times \mathscr{P}(\mathcal{S})} f(B^n, A^n, \bar{X}^n | C = (h, P, Q)) \; D(P) \; D(Q) \; dP \; dQ. \tag{4.23}$$

Further, let $\hat{f}(B^n, A^n, \bar{X}^n | \mathcal{H}_h)$ denote the maximum likelihood of all the arm samples and

observations up to time n under the hypothesis $\mathcal{H}_h$, i.e.,

$$\hat{f}(B^n, A^n, \bar{X}^n | \mathcal{H}_h) = \sup_{P,Q} f(B^n, A^n, \bar{X}^n | C = (h, P, Q)). \tag{4.24}$$

The supremum in (4.24) is over the compact set $\mathscr{P}(\mathcal{S}) \times \mathscr{P}(\mathcal{S})$, and is attained because the likelihood is a continuous function of the TPMs. Let $(\hat{P}_{h,1}(n), \hat{P}_{h,2}(n))$ attain the supremum in (4.24). These are the ML estimates of the TPMs under the hypothesis $\mathcal{H}_h$ whose closed-form expressions are not available. Notice that we do not impose the constraint $P \neq Q$ while computing the supremum in (4.24). We show later that under an arms configuration $C = (h, P_1, P_2)$, the convergences $\hat{P}_{h,1}(n) \to P_1$, $\hat{P}_{h,2}(n) \to P_2$ hold as $n \to \infty$ almost surely. Because $P_1 \neq P_2$ by the definition of arms configuration, the constraint $\hat{P}_{h,1}(n) \neq \hat{P}_{h,2}(n)$ is almost surely automatically satisfied for all n sufficiently large.

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

$$M_{hh'}(n) := \log \frac{\bar{f}(B^n, A^n, \bar{X}^n | \mathcal{H}_h)}{\hat{f}(B^n, A^n, \bar{X}^n | \mathcal{H}_{h'})}. \tag{4.25}$$

Before presenting the exact mathematical expression for (4.25), we introduce a few notations. Given a policy π and an arms configuration C, let $Z_C^\pi(n)$ denote the log-likelihood of all the arm samples and the observations up to time n under the policy π and the arms configuration C. That is,

$$\begin{aligned}
&Z_C^\pi(n) \\
&= \log f(B^n, A^n, \bar{X}^n | C) \\
&= \sum_{a=1}^{K} \log P^\pi(X_{a-1}^a | C) + \sum_{t=K}^{n} \log P^\pi(B_t, A_t, \bar{X}_t \mid B^{t-1}, A^{t-1}, \bar{X}^{t-1}, C) \\
&= \sum_{a=1}^{K} \log P^\pi(X_{a-1}^a | C) + \sum_{t=K}^{n} \log P_C^\pi(B_t, A_t \mid B^{t-1}, A^{t-1}, \bar{X}^{t-1}, C) \\
&\qquad\qquad + \sum_{t=K}^{n} \log P^\pi(\bar{X}_t \mid B^t, A^t, \bar{X}^{t-1}, C) \\
&= \sum_{a=1}^{K} \log P^\pi(X_{a-1}^a | C) + \sum_{t=K}^{n} \log P^\pi(B_t, A_t \mid B^{t-1}, A^{t-1}, \bar{X}^{t-1}, C) + \sum_{t=K}^{n} \log P^\pi(\bar{X}_t \mid A^t, \bar{X}^{t-1}, C).
\end{aligned} \tag{4.26}$$

In writing (4.26), we use the fact that $\bar{X}_t$ is conditionally independent of B^t, conditioned on

the actually sampled arm A_t. The last summation term in (4.26) may be written as

$$\begin{aligned}
&\sum_{t=K}^{n} \log P^{\pi}(\bar{X}_t \mid A^t, \bar{X}^{t-1}, C) \\
&= \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \sum_{j\in\mathcal{S}} \sum_{t=K}^{n} \mathbb{I}_{\{\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a,X_t^a=j\}} \ \log P^{\pi}(\bar{X}_t \mid A^t, \bar{X}^{t-1}, C) \\
&\overset{(a)}{=} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \sum_{j\in\mathcal{S}} \sum_{t=K}^{n} \mathbb{I}_{\{\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a,X_t^a=j\}} \ \log (P_C^a)^{d_a}(j|i_a) \\
&= \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \sum_{j\in\mathcal{S}} N(n,\underline{d},\underline{i},a,j) \ \log (P_C^a)^{d_a}(j|i_a), \qquad (4.27)
\end{aligned}$$

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

$$P^{\pi}(\bar{X}_t \mid A^t, \bar{X}^{t-1}, C) = P^{\pi}(X_t^a = j \mid X_{t-d_a}^a = i_a, C) = (P_C^a)^{d_a}(j|i_a). \qquad (4.28)$$

Also, the term $N(n,\underline{d},\underline{i},a,j)$ in (4.27) is defined as

$$N(n,\underline{d},\underline{i},a,j) := \sum_{t=K}^{n} \mathbb{I}_{\{\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a,X_t^a=j\}}, \qquad (4.29)$$

and denotes the number of times up to time n the controlled Markov process $\{(\underline{d}(t),\underline{i}(t)) : t \geq K\}$ is observed to be in the state $(\underline{d},\underline{i})$, arm a is sampled subsequently, and the state $j \in \mathcal{S}$ is observed on arm a. Plugging (4.27) into (4.26), we get

$$\begin{aligned}
Z_C^{\pi}(n) = \sum_{a=1}^{K} \log P^{\pi}(X_{a-1}^a|C) + \sum_{t=K}^{n} \log P^{\pi}(B_t, A_t \mid B^{t-1}, A^{t-1}, \bar{X}^{t-1}, C) \\
+ \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \sum_{j\in\mathcal{S}} N(n,\underline{d},\underline{i},a,j) \ \log (P_C^a)^{d_a}(j|i_a). \qquad (4.30)
\end{aligned}$$

Under an arms configuration $C = (h, P, Q)$, (4.30) may be written more explicitly as

$$\begin{aligned}
Z_C^{\pi}(n) &= \sum_{t=K}^{n} \log P^{\pi}(B_t, A_t \mid B^{t-1}, A^{t-1}, \bar{X}^{t-1}, C) \qquad (4.31) \\
&+ \log P^{\pi}(X_{h-1}^h|C) + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} N(n,\underline{d},\underline{i},h,j) \ \log \ P^{d_h}(j|i_h) \qquad (4.32)
\end{aligned}$$

$$+\sum_{a\neq h}\log P^{\pi}(X_{a-1}^{a}|C)+\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a\neq h}\sum_{j\in\mathcal{S}}N(n,\underline{d},\underline{i},a,j)\ \log\ Q^{d_a}(j|i_a). \tag{4.33}$$

Suppose that $X_0^a \sim \phi\ \forall\ a \in \mathcal{A}$, where ϕ is a probability distribution ϕ on $\mathcal{S}$ that puts a strictly positive mass on each element of $\mathcal{S}$ (e.g., $\phi(i) = \frac{1}{|\mathcal{S}|}$ for all $i \in \mathcal{S}$) and is known to the decision entity beforehand; see Section 4.7 for a discussion on ϕ. Then, we have

$$Z_C^{\pi}(n) = \sum_{t=K}^{n}\log P^{\pi}(B_t, A_t \mid B^{t-1}, A^{t-1}, \bar{X}^{t-1}, C) \tag{4.34}$$

$$+\log\left(\sum_{i\in\mathcal{S}}\ \phi(i)\,P^{h-1}(X_{h-1}^{h}|i)\right)+\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}N(n,\underline{d},\underline{i},h,j)\ \log\ P^{d_h}(j|i_h) \tag{4.35}$$

$$+\sum_{a\neq h}\log\left(\sum_{i\in\mathcal{S}}\ \phi(i)\,Q^{a-1}(X_{a-1}^{a}|i)\right)+\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a\neq h}\sum_{j\in\mathcal{S}}N(n,\underline{d},\underline{i},a,j)\ \log\ Q^{d_a}(j|i_a). \tag{4.36}$$

With the above notations in place, (4.25) may be written as

$$M_{hh'}(n) = T_1(n) + T_2(n) + T_3(n) + T_4(n), \tag{4.37}$$

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

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

$$T_1(n) = \log\mathbb{E}\Bigg[\exp\Bigg(\log\Bigg(\sum_{i\in\mathcal{S}}\ \phi(i)\,P^{h-1}(X_{h-1}^{h}|i)\Bigg) + \sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}N(n,\underline{d},\underline{i},h,j)\ \log P^{d_h}(j|i_h)\Bigg)\Bigg], \tag{4.38}$$

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

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

$$T_2(n) = \log\mathbb{E}\Bigg[\exp\Bigg(\sum_{a\neq h}\log\Bigg(\sum_{i\in\mathcal{S}}\ \phi(i)\,Q^{a-1}(X_{a-1}^{a}|i)\Bigg) + \sum_{a\neq h}\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}N(n,\underline{d},\underline{i},a,j)\ \log Q^{d_a}(j|i_a)\Bigg)\Bigg], \tag{4.39}$$

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

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

$$T_3(n) = \log \frac{1}{\sum\limits_{i \in \mathcal{S}} \phi(i)\ \hat{P}_{h',1}(n)(X_{h'-1}^{h'}|i)} + \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h', j)\ \log \frac{1}{(\hat{P}_{h',1}^{n})^{d_{h'}}(j|i_{h'})}. \tag{4.40}$$

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

$$T_4(n) = \sum_{a \neq h'} \log \frac{1}{\sum\limits_{i \in \mathcal{S}} \phi(i)\ \hat{P}_{h',2}^{n}(X_{a-1}^{a}|i)} + \sum_{a \neq h'} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, a, j)\ \log \frac{1}{(\hat{P}_{h',2}^{n})^{d_a}(j|i_a)}. \tag{4.41}$$

In writing the sum $T_1(n) + T_2(n)$ in (4.37), we use the fact that P and Q are sampled independently according to the prior D. Notice that the term (4.34) does not appear in the expression for $M_{hh'}(n)$. This is because π is oblivious to the knowledge of the underlying arms configuration C, as a result of which the probability of selecting arm B_t based on the history $(B^{t-1}, A^{t-1}, \bar{X}^{t-1})$ is independent of C. Therefore, (4.34) appears in both the numerator and the denominator of (4.25) and cancels out. We shall refer to $M_{hh'}(n)$ as the *modified GLR test statistic* of hypothesis $\mathcal{H}_h$ with respect to $\mathcal{H}_{h'}$ at time n. Let $M_h(n) = \min\limits_{h' \neq h} M_{hh'}(n)$ denote the modified GLR test statistic of hypothesis $\mathcal{H}_h$ with respect to its nearest alternative hypothesis.

4.4.3 Policy Based on Certainty Equivalence

Fix $L > 1, \delta > 0$. Our policy, which we denote by $\pi_2^\star(L, \delta)$, is as below with L and δ as parameters.

Policy $\pi_2^\star(L, \delta)$:

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

(1) Compute $\theta(n) \in \arg\max\limits_{h \in \mathcal{A}} \min\limits_{h' \neq h} M_{hh'}(n)$. Resolve ties, if any, uniformly at random.

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

(3) If $M_{\theta(n)}(n) < \log((K-1)L)$, sample arm B_{n+1} according to $\lambda_{\theta(n),\hat{P}_{\theta(n),1}(n),\hat{P}_{\theta(n),2}(n),\delta}(\cdot \mid \underline{d}(n), \underline{i}(n))$.

(4) Update $n \leftarrow n+1$ and go back to (1).

In item (1) above, $\theta(n)$ denotes the guess of the odd arm at time n. In item (2), we check if $M_{\theta(n)}(n)$ has exceeded a certain fixed threshold ($\geq \log((K-1)L)$). If this is the case, then the policy is confident that the true odd arm index is $\theta(n)$, in which case the policy terminates

at time n and outputs $\theta(n)$ as the odd arm index. If the condition in item (2) fails, the policy supposes that $C_n = (\theta(n), \hat{P}_{\theta(n),1}(n), \hat{P}_{\theta(n),2}(n))$ is the true arms configuration, and samples arm B_{n+1} according to the δ-optimal solution for C_n, thus following the principle of certainty equivalence. Observe that the policy does not rely on the knowledge of the constant $\bar{\varepsilon}^*$ from Assumption 2.

Remark 6. *1. The presence of arm delays in (4.34)-(4.36), the expression for the log-likelihood function, makes obtaining the closed-form expressions for $(\hat{P}_{h,1}(n), \hat{P}_{h,2}(n))$ and the exact computation of $M_{hh'}(n)$ difficult. This difficulty can be circumvented by repeatedly sampling the Markov process of each arm and using the successive observations from each of the arms to estimate the TPMs. Because the Markov process of each arm is ergodic, these TPM estimates will converge to their true values asymptotically. However, this might be suboptimal in not achieving the lower bound (4.11). We shall see that $\pi_2^\star(L, \delta)$ samples the arms eventually according to $\lambda_{h,P_1,P_2,\delta}$, and therefore approaches the lower bound asymptotically as $L \to \infty$ and $\delta \downarrow 0$.*

2. Moreover, when sampling the arms repeatedly, it is not clear if the desired error probability criterion can be met. We shall see that the policy $\pi_2^\star(L, \delta)$ meets the desired error probability criterion for a suitable choice of L.

4.4.4 Results on the Performance of the Policy

We now present the results on the performance of the above described policy. The proofs are given in Section 4.6. Let Assumptions 1 and 2 hold.

Proposition 12. *Under the arms configuration $C = (h, P_1, P_2)$,*

$$\hat{P}_{h,1}(n) \longrightarrow P_1, \quad \hat{P}_{h,2}(n) \longrightarrow P_2 \quad \text{as} \quad n \to \infty \quad \text{almost surely.} \tag{4.42}$$

Proof. See Section 4.6.4. □

The proof of Proposition 12 is based on verifying that the assumptions of [45] hold in the context of this chapter. The result then simply follows from [45, Theorem 4.3]. It is instructive to note here that the proof of [45, Theorem 4.3] is based on a notion of "$\{\varepsilon_i\}$-randomisation" which, for controlled Markov processes, ensures a strictly positive probability of choosing each control at each time instant. The trembling hand model (4.1) of this chapter ensures that the probability of sampling an arm at any given time is $\geq \frac{\eta}{K} > 0$, thus alleviating the need to consider $\{\varepsilon_i\}$-randomisations.

An immediate consequence of Proposition 12 is the following: suppose $C = (h, P_1, P_2)$ is the underlying arms configuration. Then, for any $h' \in \mathcal{A}$ such that $h' \neq h$,

$$\hat{P}_{h',1}(n) \longrightarrow P_2, \quad \hat{P}_{h',2}(n) \longrightarrow P \quad \text{as} \quad n \to \infty \quad \text{almost surely}, \tag{4.43}$$

where P in (4.43) is a transition probability matrix that is a function of P_1 and P_2, but whose closed-form expression is not available. It is worth noting here that when the arms are rested, the (i, j)th entry of P is a convex combination of the (i, j)th entries of P_1 and P_2 whose closed-form expression given by (2.93) of Chapter 2.

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

Proposition 13. *Fix $L > 1$ and $\delta > 0$. Let $C = (h, P_1, P_2)$ be the underlying arms configuration. Under $\pi^{\star}_{\mathsf{ns}}(L, \delta)$, the non-stopping version of the policy $\pi^{\star}_2(L, \delta)$, $\forall\ h' \neq h$, we have*

$$\liminf_{n\to\infty} \frac{M_{hh'}(n)}{n} > 0 \quad \textit{almost surely.} \tag{4.44}$$

Proof. The proof uses the convergences in (4.42). See Section 4.6.5 for the details. □

An immediate consequence of Proposition 13 is that, almost surely, $\liminf_{n\to\infty} M_h(n) > 0$. For each $h' \in \mathcal{A}$, let $\pi^{\star}_{h'}(L, \delta)$ denote a version of the policy $\pi^{\star}_2(L, \delta)$ that waits until the event $M_{h'}(n) \geq \log((K-1)L)$ occurs, at which point it stops and always declares h' as the index of the odd arm. Under the arms configuration $C = (h, P_1, P_2)$, it then follows that the stopping time of the policy $\pi^{\star}_2(L, \delta)$ may be upper bounded by that of $\pi^{\star}_h(L, \delta)$, i.e., $\tau(\pi^{\star}_2(L, \delta)) \leq \tau(\pi^{\star}_h(L, \delta))$ almost surely under $C = (h, P_1, P_2)$. As a consequence, under $C = (h, P_1, P_2)$, the following set of inequalities hold almost surely:

$$\begin{aligned}
\tau(\pi^{\star}_2(L, \delta)) &\leq \tau(\pi^{\star}_h(L, \delta)) \\
&= \inf\{n \geq 1 : M_h(n) \geq \log((K-1)L)\} \\
&\leq \inf\left\{n \geq 1 : M_{hh'}(n') \geq \log((K-1)L) \text{ for all } n' \geq n \text{ and for all } h' \neq h\right\} \\
&< \infty,
\end{aligned} \tag{4.45}$$

where the last line above is due to Proposition 13. This establishes that the policy $\pi^{\star}_2(L, \delta)$ stops in finite time almost surely.

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

$$\mathsf{GLR}_{hh'}(n) := \log \frac{\hat{f}(B^n, A^n, \bar{X}^n | \mathcal{H}_h)}{\hat{f}(B^n, A^n, \bar{X}^n | \mathcal{H}_{h'})} \tag{4.46}$$

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

$$\begin{aligned}
\limsup_{n\to\infty} M_{h'}(n) &= \limsup_{n\to\infty} \min_{a\neq h'} M_{h'a}(n) \\
&\leq \limsup_{n\to\infty} M_{h'h}(n) \\
&\leq \limsup_{n\to\infty} \mathsf{GLR}_{h'h}(n) \\
&= \limsup_{n\to\infty} -\mathsf{GLR}_{hh'}(n) \\
&= -\liminf_{n\to\infty} \mathsf{GLR}_{hh'}(n) \\
&\leq -\liminf_{n\to\infty} M_{hh'}(n) \\
&\leq -\liminf_{n\to\infty} \min_{a\neq h} M_{h,a}(n) \\
&= -\liminf_{n\to\infty} M_h(n) \\
&< 0, \qquad (4.47)
\end{aligned}$$

where (4.47) follows from Proposition 13. Under the arms configuration $C = (h, P_1, P_2)$ and under the policy $\pi^{\star}_{\mathsf{ns}}(L, \delta)$, almost surely,

$$\theta(n) = \arg\max_{h\in\mathcal{A}} M_h(n) = h \quad \forall\ n \text{ sufficiently large}, \tag{4.48}$$

which together with (4.42) implies that

$$\lambda_{\theta(n),\hat{P}_{\theta(n),1}(n),\hat{P}_{\theta(n),2}(n),\delta} \longrightarrow \lambda_{h,P_1,P_2,\delta} \quad \text{as } n \to \infty, \tag{4.49}$$

where the convergence in (4.49) is with respect to the product topology on the space $\bigotimes_{(\underline{d},\underline{i})\in\mathbb{S}} \mathcal{P}(\mathcal{A})$, i.e.,

$$\lambda_{\theta(n),\hat{P}_{\theta(n),1}(n),\hat{P}_{\theta(n),2}(n),\delta}(\cdot|\underline{d},\underline{i}) \to \lambda_{h,P_1,P_2,\delta}(\cdot|\underline{d},\underline{i}) \quad \text{as } n \to \infty \quad \forall\ (\underline{d},\underline{i}) \in \mathbb{S}. \tag{4.50}$$

The property in (4.49) ensures that under the non-stopping policy $\pi^{\star}_{\mathsf{ns}}(L,\delta)$, as more observations are collected from the arms and as the ML estimates of the TPMs approach their true values defined by the underlying arms configuration, the arms are eventually sampled according to the δ-optimal selection for the underlying arms configuration. As we shall see shortly, it is this property of $\pi^{\star}_{\mathsf{ns}}(L,\delta)$ that plays a key role in showing the asymptotic optimality of the policy $\pi^{\star}_{2}(L,\delta)$.

Let us return to the policy $\pi^{\star}_{2}(L,\delta)$. The next result shows that any error probability $\epsilon > 0$ can be met by setting the parameter $L = 1/\epsilon$.

Proposition 14. *Fix an error probability $\epsilon > 0$. If $L = 1/\epsilon$, then $\pi^{\star}_{2}(L,\delta) \in \Pi(\epsilon)$ for all $\delta > 0$.*

Proof. The proof uses the fact that the policy stops in finite time almost surely. The details may be found in Section 4.6.6. □

An important step in the proof of Proposition 14 is to upper bound the likelihood function under the arms configuration $C = (h, P_1, P_2)$ by the maximum likelihood under the hypothesis $\mathcal{H}_h$. Such an exercise of upper bounding the likelihood function is not possible if, for instance, the ML estimates in the expression for the maximum likelihood are replaced by the estimates of the TPMs computed by repeatedly sampling each of the arms.

Remark 7. *In the proof of Proposition 14 presented in Appendix 4.6.6, it is worth noting that in the series of steps leading to (4.133), the integral in (b) over the set $\mathcal{R}_{h'}(n)$ with respect to the average likelihood function $\bar{f}(\cdot|\mathcal{H}_{h'})$ is simply equal to the probability of the set $\mathcal{R}_{h'}(n)$ under the probability measure $\bar{P}^{\pi}(\cdot|\mathcal{H}_{h'})$ generated by $\bar{f}(\cdot|\mathcal{H}_{h'})$. In contrast, the integral over $\mathcal{R}_{h'}(n)$ with respect to the maximum likelihood function $\hat{f}(\cdot|\mathcal{H}_{h'})$ cannot be replaced by a corresponding probability term. This is because the maximum likelihood function is not a valid density, whereas the average likelihood function is a valid density. It is for this reason that we modify the classical GLR test statistic by replacing the maximum likelihood in the numerator of the classical GLR test statistic with the average likelihood.*

The next result improves upon Proposition 13 and shows that the modified GLR test statistic has the correct drift under the non-stopping version of the policy $\pi^{\star}_{2}(L,\delta)$.

Proposition 15. *Suppose that $C = (h, P_1, P_2)$ is the underlying arms configuration. Fix $L > 1$ and $\delta > 0$. For all $h' \in \mathcal{A}$ such that $h' \neq h$, under the non-stopping version of the policy $\pi^{\star}_{2}(L,\delta)$,*

$$\liminf_{n\to\infty} \frac{M_{hh'}(n)}{n} \geq \frac{R^{*}(h,P_1,P_2)}{(1+\delta)^2} \quad \textit{almost surely.} \tag{4.51}$$

Proof. It is in proving this proposition that we make use of Assumption 1. The proof uses the fact that almost surely, $\theta(n) = h$ for all sufficiently large n under the arms configuration $C = (h, P_1, P_2)$. As a result, $\hat{P}_{\theta(n),1}(n) = \hat{P}_{h,1}(n)$, $\hat{P}_{\theta(n),2}(n) = \hat{P}_{h,2}(n)$ for all sufficiently large n. The convergences in (4.42) imply that the pair $(\hat{P}_{h,1}(n), \hat{P}_{h,2}(n))$ lies inside a neighbourhood (chosen based on the value of δ) of (P_1, P_2) for all n sufficiently large. These together with Assumption 1 and (4.49) imply that $\lambda_{\theta(n),\hat{P}_{\theta(n),1}(n),\hat{P}_{\theta(n),2}(n),\delta}$ is close to $\lambda_{h,P_1,P_2,\delta}$ for all n sufficiently large. Using this, we arrive at (4.51). For the details, see Section 4.6.7. □

Next, we show that the stopping time of the policy $\pi_2^\star(L, \delta)$ blows up as $L \to \infty$.

Proposition 16. *For each $\delta > 0$,*

$$\liminf_{L\to\infty} \tau(\pi_2^\star(L, \delta)) = \infty \quad \textit{almost surely.} \tag{4.52}$$

Proof. See Section 4.6.8. □

Combining the results of Proposition 15 and Proposition 16, it follows that for all $\delta > 0$,

$$\liminf_{L\to\infty} \frac{M_{hh'}(\tau(\pi_2^\star(L, \delta)))}{\tau(\pi_2^\star(L, \delta))} \geq \frac{R^*(h, P_1, P_2)}{(1+\delta)^2} \quad \text{almost surely.} \tag{4.53}$$

The following result shows that the stopping time of the policy $\pi_2^\star(L, \delta)$ satisfies an almost sure upper bound that is arbitrarily close to the lower bound in (4.11).

Proposition 17. *Suppose that $C = (h, P_1, P_2)$ is the underlying arms configuration. For all $\delta > 0$, the stopping time of policy $\pi_2^\star(L, \delta)$ satisfies*

$$\limsup_{L\to\infty} \frac{\tau(\pi_2^\star(L, \delta))}{\log L} \leq \frac{(1+\delta)^2}{R^*(h, P_1, P_2)} \quad \textit{almost surely.} \tag{4.54}$$

Proof of Proposition 17. By the definition of $\tau(\pi_2^\star(L, \delta))$, we know that under the arms configuration $C = (h, P_1, P_2)$,

$$M_h(\tau(\pi_2^\star(L, \delta))) \geq \log((K-1)L), \quad M_h(\tau(\pi_2^\star(L, \delta)) - 1) < \log((K-1)L).$$

We then have, almost surely,

$$\begin{aligned} 1 &= \limsup_{L\to\infty} \frac{\log((K-1)L)}{\log L} \\ &\geq \limsup_{L\to\infty} \frac{M_h(\tau(\pi_2^\star(L, \delta)) - 1)}{\log L} \end{aligned}$$

$$= \limsup_{L\to\infty} \frac{M_h(\tau(\pi_2^\star(L,\delta)) - 1)}{\tau(\pi_2^\star(L,\delta)) - 1} \cdot \frac{\tau(\pi_2^\star(L,\delta)) - 1}{\log L}$$
$$\geq \frac{R^*(h, P_1, P_2)}{(1+\delta)^2} \cdot \limsup_{L\to\infty} \frac{\tau(\pi_2^\star(L,\delta))}{\log L}, \tag{4.55}$$

where (4.55) follows from (4.53). The desired result follows by rearranging (4.55). □

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

Proposition 18. *Fix $\delta > 0$. Under the arms configuration $C = (h, P_1, P_2)$, the expected stopping time of the policy $\pi = \pi_2^\star(L,\delta)$ satisfies*

$$\limsup_{L\to\infty} \frac{E^\pi[\tau(\pi)|C]}{\log L} \leq \frac{(1+\delta)^2}{R^*(h, P_1, P_2)}. \tag{4.56}$$

Proof. In the proof, which we present in Section 4.6.9, we show that for each $\delta > 0$, the family $\{\tau(\pi_2^\star(L,\delta))/\log L : L > 1\}$ is uniformly integrable. Combining the almost sure upper bound in (4.54) with uniform integrability yields the desired upper bound in (4.56) for the expected stopping time of the policy $\pi_2^\star(L,\delta)$. □

4.5 Main Result

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

Theorem 4. *Consider a multi-armed bandit with $K \geq 3$ arms in which each arm is a time homogeneous and ergodic Markov process on the finite state space $\mathcal{S}$. Fix an arms configuration $C = (h, P_1, P_2)$; here, h is the odd arm, P_1 is the TPM of the odd arm h, and $P_2 \neq P_1$ is the TPM of each non-odd arm. Fix $\eta \in (0, 1]$, and suppose that a decision entity that wishes to identify the odd arm has a trembling hand with parameter η. When neither P_1 nor P_2 is known beforehand to the decision entity, the expected time required to identify the odd arm satisfies the asymptotic lower bound*

$$\liminf_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{E^\pi[\tau(\pi)|C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h, P_1, P_2)}. \tag{4.57}$$

Further, under Assumption 1 and Assumption 2, the policy $\pi_2^\star(L,\delta)$ satisfies the asymptotic upper bound

$$\limsup_{\delta\downarrow 0} \limsup_{L\to\infty} \frac{E^{\pi_2^\star(L,\delta)}[\tau(\pi_2^\star(L,\delta))|C]}{\log L} \leq \frac{1}{R^*(h, P_1, P_2)}. \tag{4.58}$$

Under these assumptions, we therefore have

$$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^{\pi}[\tau(\pi)|C]}{\log(1/\epsilon)} = \lim_{\delta \downarrow 0} \lim_{L \to \infty} \frac{E^{\pi_2^\star(L,\delta)}[\tau(\pi_2^\star(L,\delta))|C]}{\log L} = \frac{1}{R^*(h, P_1, P_2)}. \tag{4.59}$$

Proof. The asymptotic lower bound in (4.57) follows from Proposition 11. Recall Assumption 1 and Assumption 2. Taking limits as $\delta \downarrow 0$ in (4.56), we arrive at (4.58). From Proposition 14, we know that given any error tolerance parameter $\epsilon > 0$, by setting $L = 1/\epsilon$, we have $\pi_2^\star(L, \delta) \in \Pi(\epsilon)$ for all $\delta > 0$. Therefore, it follows that for all $\epsilon, \delta > 0$,

$$\inf_{\pi \in \Pi(\epsilon)} \frac{E^{\pi}[\tau(\pi)|C]}{\log\left(1/\epsilon\right)} \leq \frac{E^{\pi_2^\star(L,\delta)}[\tau(\pi_2^\star(L,\delta))|C]}{\log L}. \tag{4.60}$$

Fixing $\delta > 0$ and letting $\epsilon \downarrow 0$ (which is identical to letting $L \to \infty$) in (4.60), and using the upper bound in (4.56), we get

$$\limsup_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^{\pi}[\tau(\pi)|C]}{\log(1/\epsilon)} \leq \limsup_{L \to \infty} \frac{E^{\pi_2^\star(L,\delta)}[\tau(\pi_2^\star(L,\delta))|C]}{\log L} \leq \frac{(1+\delta)^2}{R^*(h, P_1, P_2)}. \tag{4.61}$$

Letting $\delta \downarrow 0$ in (4.61) and noting that the leftmost term in (4.61) does not depend on δ, we get

$$\limsup_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^{\pi}[\tau(\pi)|C]}{\log(1/\epsilon)} \leq \limsup_{\delta \downarrow 0} \limsup_{L \to \infty} \frac{E^{\pi_2^\star(L,\delta)}[\tau(\pi_2^\star(L,\delta))|C]}{\log L} \leq \frac{1}{R^*(h, P_1, P_2)}. \tag{4.62}$$

Combining the result in (4.62) with the lower bound in (4.11), we get

$$\begin{aligned}
\frac{1}{R^*(h, P_1, P_2)} &\leq \liminf_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^{\pi}[\tau(\pi)|C]}{\log(1/\epsilon)} \\
&\leq \limsup_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^{\pi}[\tau(\pi)|C]}{\log(1/\epsilon)} \\
&\leq \limsup_{\delta \downarrow 0} \limsup_{L \to \infty} \frac{E^{\pi_2^\star(L,\delta)}[\tau(\pi_2^\star(L,\delta))|C]}{\log L} \\
&\leq \frac{1}{R^*(h, P_1, P_2)}.
\end{aligned} \tag{4.63}$$

Thus, it follows that the limit infimum and the limit suprema in the chain of inequalities leading to (4.63) are indeed limits, thereby yielding (4.59). This completes the proof of the theorem. □

Thus, under the arms configuration $C = (h, P_1, P_2)$, the asymptotic growth rate for the

expected time required to identify the odd arm is $1/R^*(h, P_1, P_2)$. While the asymptotic lower bound holds for all arms configurations, the asymptotic upper bound is only established under a continuous selection assumption for those arms configurations whose TPMs satisfy a regularity condition. The continuous selection assumption and the regularity condition are used to prove identifiability. The trembling hand model (4.1) ensures that at any given time, each arm is selected with a strictly positive probability under every policy, and plays a key role in deriving the lower and the upper bounds.

4.6 Proofs

4.6.1 Proof of Proposition 11

Given a policy π and two arm configurations $C = (h, P_1, P_2)$ and $C' = (h', P_1', P_2')$ such that $h \neq h'$, let $Z^{\pi}_{CC'}(n) := Z^{\pi}_{C}(n) - Z^{\pi}_{C'}(n)$ denote the log-likelihood ratio, under the policy π, of all the arm samples and observations up to time n under the arms configuration C with respect to that under C'. Note that (4.34), being independent of the arms configuration and common to the expressions for $Z^{\pi}_{C}(n)$ and $Z^{\pi}_{C'}(n)$, cancels out when writing the expression for $Z^{\pi}_{C}(n) - Z^{\pi}_{C'}(n)$. It then follows from (4.30) that

$$Z^{\pi}_{CC'}(n) = \sum_{a=1}^{K} \log \frac{P^{\pi}_{C}(X^a_{a-1})}{P^{\pi}_{C'}(X^a_{a-1})} + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \sum_{j\in\mathcal{S}} N(n, \underline{d}, \underline{i}, a, j) \log \frac{(P^a_C)^{d_a}(j|i_a)}{(P^a_{C'})^{d_a}(j|i_a)}. \tag{4.64}$$

We organise the proof of Proposition 11 as follows. Given an error probability $\epsilon > 0$, we first obtain a lower bound for $E^{\pi}[Z^{\pi}_{CC'}(\tau(\pi))|C]$ for all $\pi \in \Pi(\epsilon)$ using a change of measure argument of Kaufmann et al. [32]. Following this, we obtain an upper bound for $E^{\pi}[Z^{\pi}_{CC'}(\tau(\pi))|C]$ in terms of $E^{\pi}[\tau(\pi)|C]$ using a Wald-type lemma for the setting of restless Markov arms. Combining the upper and the lower bounds, and letting $\epsilon \downarrow 0$, we arrive at the desired result. The ergodicity property established in Lemma 9 of Chapter 3 for SRS policies plays a crucial role in deriving the final lower bound (4.11).

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

As a first step towards deriving the lower bound, we use a result of Kaufmann et al. [32] to obtain a lower bound for $E^{\pi}[Z^{\pi}_{CC'}(\tau(\pi))|C]$ in terms of the error probability parameter ϵ. This is based on a generalisation of [32, Lemma 18], a change of measure argument for iid observations from the arms, to the setting of restless arms with Markov observations. We present this generalisation in the following lemma.

Lemma 23. *Let $(\Omega, \mathcal{F})$ be a measurable space. Let the filtration $\{\mathcal{F}_t : t \geq 0\}$ be defined as follows: $\mathcal{F}_0 = \sigma(\Omega, \emptyset)$ and $\mathcal{F}_t = \sigma(B^t, A^t, \bar{X}^t)$ for all $t \geq 1$. Fix $\pi \in \Pi(\epsilon)$, and let $\tau(\pi)$ be the stopping time of policy π. Let $\mathcal{F}_{\tau(\pi)}$ be the σ-algebra*

$$\mathcal{F}_{\tau(\pi)} = \{E \in \mathcal{F} : E \cap \{\tau(\pi) = t\} \in \mathcal{F}_t \text{ for all } t \geq 0\}. \tag{4.65}$$

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

$$P^\pi(E|C') = E^\pi[1_E \exp(-Z^\pi_{CC'}(\tau(\pi)))|C] \tag{4.66}$$

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

Proof of Lemma 23. We prove the lemma by demonstrating, through mathematical induction, that the relation

$$E^\pi[g(B^t, A^t, \bar{X}^t)|C'] = E^\pi[g(B^t, A^t, \bar{X}^t) \exp(-Z^\pi_{CC'}(t))|C] \tag{4.67}$$

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

$$\begin{aligned}
&E^\pi[g(B_0, A_0, \bar{X}_0)|C'] \\
&= \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}} g(b,a,i)\ P^\pi(B_0 = b, A_0 = a, \bar{X}_0 = i|C') \\
&= \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}} g(b,a,i)\ P^\pi(B_0 = b|C')\ P^\pi(A_0 = a|B_0 = b, C')\ P^\pi(\bar{X}_0 = i|B_0 = b, A_0 = a, C') \\
&\overset{(a)}{=} \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}} g(b,a,i)\ P^\pi(B_0 = b|C)\ P^\pi(A_0 = a|B_0 = b, C)\ P^\pi(\bar{X}_0 = i|A_0 = a, C') \\
&= \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}} g(b,a,i)\ P^\pi(B_0 = b|C)\ P^\pi(A_0 = a|B_0 = b, C)\ P^\pi(X_0^a = i|C'),
\end{aligned} \tag{4.68}$$

where (a) follows by noting that

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

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

$$
\begin{aligned}
&E^\pi[g(B_0, A_0, \bar{X}_0)|C'] \\
&= \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}} g(b,a,i)\ P^\pi(B_0 = b|C)\ P^\pi(A_0 = a|B_0 = b, C)\ \phi(i) && (4.69)\\
&= \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}} g(b,a,i)\ P^\pi(B_0 = b|C)\ P^\pi(A_0 = a|B_0 = b, C)\ P^\pi(X_0^a = i|A_0 = a, C) && (4.70)\\
&= \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}} g(b,a,i)\ P^\pi(B_0 = b|C)\ P^\pi(A_0 = a|B_0 = b, C)\ P^\pi(X_0^a = i|A_0 = a, B_0 = b, C). && (4.71)
\end{aligned}
$$

Noting that

$$Z^\pi_{CC'}(0) = \log \frac{P^\pi(B_0, A_0, \bar{X}_0|C)}{P^\pi(B_0, A_0, \bar{X}_0|C')} = 0, \tag{4.72}$$

and combining (4.71) with (4.72), we get

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

This proves (4.67) for $t = 0$.

Assume now that (4.67) is true for some $t > 0$. We shall demonstrate that (4.67) holds for $t + 1$. By the law of iterated expectations,

$$E^\pi[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})|C'] = E^\pi[E^\pi[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})|\mathcal{F}_t, C']|C']. \tag{4.73}$$

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

$$E^\pi[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})|\mathcal{F}_t, C'] = E^\pi[E^\pi[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})|\mathcal{F}_t, C']\ \exp(-Z^\pi_{CC'}(t))|C]. \tag{4.74}$$

We now note that

$$
\begin{aligned}
&E^\pi[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})|\mathcal{F}_t, C']\ \exp(-Z^\pi_{CC'}(t)) \\
&\overset{(a)}{=} E^\pi[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})\ \exp(-Z^\pi_{CC'}(t))|\mathcal{F}_t, C']
\end{aligned}
$$

$$= \sum_{b=1}^{K} \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} \Big[g(B^t, A^t, \bar{X}^t, b, a, i) \cdot P^\pi(B_{t+1} = b | B^t, A^t, \bar{X}^t, C') \\ \cdot P^\pi(A_{t+1} = a | B^{t+1} = b, B^t, A^t, \bar{X}^t, C') \\ \cdot P^\pi(\bar{X}_{t+1} = i | B^{t+1} = b, A_{t+1} = a, B^t, A^t, \bar{X}^t, C') \cdot \exp(-Z^\pi_{CC'}(t)) \Big]$$

$$\overset{(b)}{=} \sum_{b=1}^{K} \sum_{a=1}^{K} \sum_{i \in \mathcal{S}} \Big[g(B^t, A^t, \bar{X}^t, b, a, i) \cdot P^\pi(B_{t+1} = b | B^t, A^t, \bar{X}^t, C) \\ \cdot P^\pi(A_{t+1} = a | B^{t+1} = b, B^t, A^t, \bar{X}^t, C) \\ \cdot P^\pi(\bar{X}_{t+1} = i | A_{t+1} = a, A^t, \bar{X}^t, C') \cdot \exp(-Z^\pi_{CC'}(t)) \Big], \qquad (4.75)$$

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

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

Also, we have

$$\sum_{i \in \mathcal{S}} P^\pi(\bar{X}_{t+1} = i | A_{t+1} = a, A^t, \bar{X}^t, C') \, \exp(-Z^\pi_{CC'}(t)) \\ = \sum_{i \in \mathcal{S}} \frac{P^\pi(\bar{X}_{t+1} = i | A_{t+1} = a, A^t, \bar{X}^t, C')}{P^\pi(\bar{X}_{t+1} = i | A_{t+1} = a, A^t, \bar{X}^t, C)} \, \exp(-Z^\pi_{CC'}(t)) \, P^\pi(\bar{X}_{t+1} = i | A_{t+1} = a, A^t, \bar{X}^t, C) \\ = \sum_{i \in \mathcal{S}} \exp(-Z^\pi_{CC'}(t+1, a, i)) \, P^\pi(\bar{X}_{t+1} = i | A_{t+1} = a, A^t, \bar{X}^t, C), \qquad (4.76)$$

since

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

Substituting (4.76) in (4.75) and simplifying, we get

$$
\begin{aligned}
&E^{\pi}[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})|\mathcal{F}_t, C'] \exp(-Z^{\pi}_{CC'}(t)) \\
&= \sum_{b=1}^{K}\sum_{a=1}^{K}\sum_{i\in\mathcal{S}} \Big[g(B^t, A^t, \bar{X}^t, b, a, i) \cdot P^{\pi}(B_{t+1} = b|B^t, A^t, \bar{X}^t, C) \\
&\qquad \cdot P^{\pi}(A_{t+1} = a|B_{t+1} = b, B^t, A^t, \bar{X}^t, C) \\
&\qquad \cdot P^{\pi}(\bar{X}_{t+1} = i|B_{t+1} = b, A_{t+1} = a, B^t, A^t, \bar{X}^t, C) \cdot \exp(-Z^{\pi}_{CC'}(t+1, a, i)) \Big] \qquad (4.77) \\
&= E^{\pi}[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1}) \ \exp(-Z^{\pi}_{CC'}(t+1))|\mathcal{F}_t, C]. \qquad (4.78)
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
&E^{\pi}[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1})|\mathcal{F}_t, C'] \exp(-Z^{\pi}_{CC'}(t)) \\
&= E^{\pi}[g(B^{t+1}, A^{t+1}, \bar{X}^{t+1}) \exp(-Z^{\pi}_{CC'}(t+1))|\mathcal{F}_t, C]. \qquad (4.79)
\end{aligned}
$$

Applying $E^{\pi}[\cdot|C]$ to both sides of (4.79), plugging it into the right hand side of (4.74), and using the law of iterated expectations, we arrive at the desired relation for $t+1$. This proves (4.67) for all $t \geq 0$.

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

$$
\begin{aligned}
P^{\pi}(E|C') &= E^{\pi}[1_E|C'] \\
&= E^{\pi}\left[\sum_{t\geq 0} 1_{E\cap\{\tau(\pi)=t\}}\middle| C'\right] \\
&\overset{(a)}{=} \sum_{t\geq 0} E^{\pi}\left[1_{E\cap\{\tau(\pi)=t\}}\middle| C'\right] \\
&\overset{(b)}{=} \sum_{t\geq 0} E^{\pi}\left[1_{E\cap\{\tau(\pi)=t\}} \exp(-Z^{\pi}_{CC'}(t))\middle| C\right] \\
&= \sum_{t\geq 0} E^{\pi}\left[1_{E\cap\{\tau(\pi)=t\}} \exp(-Z^{\pi}_{CC'}(\tau(\pi)))\middle| C\right] \\
&= E^{\pi}\left[1_E \exp(-Z^{\pi}_{CC'}(\tau(\pi)))\middle| C\right], \qquad (4.80)
\end{aligned}
$$

where (a) is due to monotone convergence theorem, and (b) above follows from (4.67) and the fact that $E \cap \{\tau(\pi) = t\} \in \mathcal{F}_t$ for all $t \geq 0$ since $E \in \mathcal{F}_{\tau(\pi)}$. This completes the proof of the lemma. □

Lemma 23 in conjunction with [32, Lemma 19] implies that when $C = (h, P_1, P_2)$ is the underlying arms configuration, the following inequality holds for all $\pi \in \Pi(\epsilon)$ and $C' = (h', P_1', P_2')$ such that $h' \neq h$:

$$E^\pi[Z^\pi_{CC'}(\tau(\pi))|C] \geq \sup_{E \in \mathcal{F}_{\tau(\pi)}} d(P^\pi(E|C), P^\pi(E|C')), \tag{4.81}$$

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

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

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

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

As noted in [32], the mapping $x \mapsto d(x,y)$ is monotone increasing for $x < y$, and the mapping $y \mapsto d(x,y)$ is monotone decreasing for any fixed x. Using these facts in (4.81), we get

$$\begin{aligned} E^\pi[Z^\pi_{CC'}(\tau(\pi))|C] &\geq d(P^\pi(\theta(\tau(\pi)) = h|C), P^\pi(\theta(\tau(\pi)) = h|C')) \\ &\geq d(\epsilon, 1-\epsilon) \end{aligned} \tag{4.82}$$

for all $\pi \in \Pi(\epsilon)$ and for all $C' = (h', P_1', P_2')$ such that $h' \neq h$, from which it follows that

$$\inf_{\substack{C'=(h',P_1',P_2'):\\ h'\neq h,\ P_1'\neq P_2'}} E^\pi[Z^\pi_{CC'}(\tau(\pi))|C] \geq d(\epsilon, 1-\epsilon). \tag{4.83}$$

4.6.3 An Upper Bound for $E^\pi[Z^\pi_{CC'}(\tau(\pi))|C]$ in Terms of $E^\pi[\tau(\pi)|C]$

We now obtain an upper bound for the left-hand side of (4.83). From (4.64), we have

$$\begin{aligned} &E^\pi[Z^\pi_{CC'}(\tau(\pi))|C] \\ &= E^\pi\left[\sum_{a=1}^K \log\frac{P^\pi_C(X^a_{a-1})}{P^\pi_{C'}(X^a_{a-1})}\middle|C\right] + E^\pi\left[\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^K\sum_{j\in\mathcal{S}} N(\tau(\pi),\underline{d},\underline{i},a,j)\log\frac{(P^a_C)^{d_a}(j|i_a)}{(P^a_{C'})^{d_a}(j|i_a)}\middle|C\right]. \end{aligned} \tag{4.84}$$

To simplify the second expectation term on the right-hand side of (4.84), we use the following result.

Lemma 24. *For every* $(\underline{d}, \underline{i}) \in \mathbb{S}$, $a \in \mathcal{A}$ *and* $j \in \mathcal{S}$,

$$\begin{aligned} &E^\pi[E^\pi[N(\tau(\pi), \underline{d}, \underline{i}, a, j)|X_{a-1}^a, C]|\tau(\pi), C] \\ &= E^\pi[E^\pi[N(\tau(\pi), \underline{d}, \underline{i}, a)|X_{a-1}^a, C]|\tau(\pi), C]\,(P_C^a)^{d_a}(j|i_a). \end{aligned} \tag{4.85}$$

Proof of Lemma 24. Substituting $n = \tau(\pi)$ in (4.29), we have

$$\begin{aligned} &E^\pi[E^\pi[N(\tau(\pi), \underline{d}, \underline{i}, a, j)|X_{a-1}^a, C]|\tau(\pi), C] \\ &= E^\pi\left[E^\pi\left[\sum_{t=K}^{\tau(\pi)} 1_{\{\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a, X_t^a=j\}}\middle| X_{a-1}^a, C\right]\middle|\tau(\pi), C\right] \\ &= E^\pi\left[\sum_{t=K}^{\tau(\pi)} P^\pi(\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a, X_t^a=j|X_{a-1}^a, C)\middle|\tau(\pi), C\right]. \end{aligned} \tag{4.86}$$

For each t in the range of the summation in (4.86), the conditional probability term for t may be expressed as

$$\begin{aligned} &P^\pi(\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a, X_t^a=j|X_{a-1}^a, C) \\ &= P^\pi(\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a|X_{a-1}^a, C)\cdot P^\pi(X_t^a=j|A_t=a, \underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, X_{a-1}^a, C) \\ &= P^\pi(\underline{d}(t)=\underline{d}, \underline{i}(t)=\underline{i}, A_t=a|X_{a-1}^a, C)\cdot (P_C^a)^{d_a}(j|i_a). \end{aligned} \tag{4.87}$$

Plugging (4.87) back in (4.86) and simplifying, we arrive at the desired relation in (4.85). □

Using Lemma 24, the second expectation term on the right-hand side of (4.84) can be simplified as follows.

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

$$\stackrel{(a)}{=} E^{\pi}\Bigg[\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^{K}\sum_{j\in\mathcal{S}} E^{\pi}[E^{\pi}[N(\tau(\pi),\underline{d},\underline{i},a)|X_{a-1}^{a},C]|\tau(\pi),C] \cdot (P_C^a)^{d_a}(j|i)\cdot\log\frac{(P_C^a)^{d_a}(j|i_a)}{(P_{C'}^a)^{d_a}(j|i_a)}\Bigg|C\Bigg]$$

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

$$= \sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^{K} E^{\pi}[N(\tau(\pi),\underline{d},\underline{i},a)|C]\cdot D((P_C^a)^{d_a}(\cdot|i_a)\|(P_{C'}^a)^{d_a}(\cdot|i_a)), \tag{4.88}$$

where in the above set of equations, (a) follows from Lemma 24,

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

for all $n \geq K$, $(\underline{d},\underline{i}) \in \mathbb{S}$ and $a \in \mathcal{A}$, and (4.88) is due to monotone convergence theorem and the fact that

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

Plugging (4.88) back in (4.84), we get

$$E^{\pi}[Z_{CC'}^{\pi}(\tau(\pi))|C]$$
$$= E^{\pi}\Bigg[\sum_{a=1}^{K}\log\frac{P_C^{\pi}(X_{a-1}^{a})}{P_{C'}^{\pi}(X_{a-1}^{a})}\Bigg] + \sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^{K} E^{\pi}[N(\tau(\pi),\underline{d},\underline{i},a)|C]\cdot D((P_C^a)^{d_a}(\cdot|i_a)\|(P_{C'}^a)^{d_a}(\cdot|i_a)). \tag{4.89}$$

Noting that

$$\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^{K} E^{\pi}[N(\tau(\pi),\underline{d},\underline{i},a)|C] \stackrel{(a)}{=} E^{\pi}\Bigg[\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^{K} N(\tau(\pi),\underline{d},\underline{i},a)\Bigg|C\Bigg]$$
$$= E^{\pi}\Bigg[\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{a=1}^{K}\sum_{t=K}^{\tau(\pi)} 1_{\{\underline{d}(t)=\underline{d},\underline{i}(t)=\underline{i},A_t=a\}}\Bigg|C\Bigg]$$
$$= E^{\pi}\Bigg[\sum_{t=K}^{\tau(\pi)} 1\Bigg|C\Bigg] \tag{4.90}$$

$$= E^{\pi}[\tau(\pi) - K + 1|C], \tag{4.91}$$

where (a) above is due to monotone convergence theorem, we write (4.89) as

$$\begin{aligned}
&E^{\pi}[Z^{\pi}_{CC'}(\tau(\pi))|C] \\
&= E^{\pi}\left[\sum_{a=1}^{K} \log \frac{P^{\pi}_{C}(X^{a}_{a-1})}{P^{\pi}_{C'}(X^{a}_{a-1})}\right] \\
&+ \left(E^{\pi}[\tau(\pi) - K + 1|C]\right) \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \frac{E^{\pi}[N(\tau(\pi),(\underline{d},\underline{i}),a)|C]}{E^{\pi}[\tau(\pi) - K + 1|C]} \cdot D((P^{a}_{C})^{d_a}(\cdot|i_a)\|(P^{a}_{C'})^{d_a}(\cdot|i_a)).
\end{aligned} \tag{4.92}$$

Combining (4.83) and (4.92), and noting that (4.92) holds for all $C' = (h', P'_1, P'_2)$ such that $h' \neq h$, we get

$$\begin{aligned}
d(\epsilon, 1-\epsilon) &\leq \inf_{\substack{C'=(h',P'_1,P'_2):\\ h'\neq h,\ P'_1\neq P'_2}} \left\{ E^{\pi}\left[\sum_{a=1}^{K} \log \frac{P^{\pi}_{C}(X^{a}_{a-1})}{P^{\pi}_{C'}(X^{a}_{a-1})}\right] \right. \\
&\left. + \left(E^{\pi}[\tau(\pi) - K + 1|C]\right) \cdot \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \frac{E^{\pi}[N(\tau(\pi),\underline{d},\underline{i},a)|C]}{E^{\pi}[\tau(\pi) - K + 1|C]} \cdot D((P^{a}_{C})^{d_a}(\cdot|i_a)\|(P^{a}_{C'})^{d_a}(\cdot|i_a)) \right\} \\
&\leq \sup_{\nu} \inf_{\substack{C'=(h',P'_1,P'_2):\\ h'\neq h,\ P'_1\neq P'_2}} \left\{ E^{\pi}\left[\sum_{a=1}^{K} \log \frac{P^{\pi}_{C}(X^{a}_{a-1})}{P^{\pi}_{C'}(X^{a}_{a-1})}\right] \right. \\
&\qquad \left. + \left(E^{\pi}[\tau(\pi) - K + 1|C]\right) \cdot \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a)\, k_{CC'}(\underline{d},\underline{i},a) \right\},
\end{aligned} \tag{4.93}$$

where the supremum in (4.93) is over all state-action occupancy measures satisfying

$$\sum_{a=1}^{K} \nu(\underline{d}',\underline{i}',a) = \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a)\, Q(\underline{d}',\underline{i}'|\underline{d},\underline{i},a) \quad \text{for all } (\underline{d}',\underline{i}') \in \mathbb{S}, \tag{4.94}$$

$$\sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \nu(\underline{d},\underline{i},a) = 1, \tag{4.95}$$

$$\nu(\underline{d},\underline{i},a) \geq 0 \quad \text{for all } (\underline{d},\underline{i},a) \in \mathbb{S}\times\mathcal{A}. \tag{4.96}$$

Recall that Q in (4.94) denotes the transition probability matrix given by (4.5). The left-hand side of (4.94) represents the long-term probability of leaving the state $(\underline{d},\underline{i})$, while the right-

hand side of (4.95) represents the long-term probability of entering into the state $(\underline{d}, \underline{i})$. Thus, (4.94) is the *global balance equation* for the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$. Equations (4.95) and (4.96) together imply that ν is a probability measure on $\mathbb{S} \times \mathcal{A}$.

As outlined in Section 4.2, the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$, together with the sequence $\{B_t : t \geq 0\}$ of intended arm selections (or equivalently the sequence $\{A_t : t \geq 0\}$ of actual arm selections), defines a Markov decision problem (MDP) with state space $\mathbb{S}$ and action space $\mathcal{A}$. From Lemma 9 of Chapter 3, we know that $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is an ergodic Markov process under every SRS policy. This suffices to apply Theorem 3 of Chapter 3 to deduce a one-one correspondence between feasible solutions to (4.94)-(4.96) and policies in Π_{SRS}. In other words, Theorem 3 of Chapter 3 implies that for any given ν satisfying (4.94)-(4.96), we can find an SRS policy $\pi^\lambda \in \Pi_{\mathsf{SRS}}$ such that $\nu^\lambda(\underline{d}, \underline{i}, a) = \nu(\underline{d}, \underline{i}, a)$ for all $(\underline{d}, \underline{i}, a) \in \mathbb{S} \times \mathcal{A}$. (Recall that under the SRS policy π^λ, the stationary distribution of the Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is μ^λ. The associated ergodic state occupancy measure, ν^λ, is then defined according to (4.9).)

Using Theorem 3 of Chapter 3, we may replace the supremum in (4.93) by a supremum over all SRS policies. Doing so leads us to the relation

$$\begin{aligned} & d(\epsilon, 1-\epsilon) \\ & \leq \sup_{\pi^\lambda \in \Pi_{\mathsf{SRS}}} \inf_{\substack{C'=(h', P_1', P_2'): \\ h' \neq h,\ P_1' \neq P_2'}} \left\{ E^\pi \left[\sum_{a=1}^{K} \log \frac{P_C^\pi(X_{a-1}^a)}{P_{C'}^\pi(X_{a-1}^a)} \right] \right. \\ & \qquad \left. + \left(E^\pi[\tau(\pi) - K + 1 | C] \right) \cdot \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu^\lambda(\underline{d}, \underline{i}, a)\, k_{CC'}(\underline{d}, \underline{i}, a) \right\}. \end{aligned} \tag{4.97}$$

for all $\pi \in \Pi(\epsilon)$. Observe that the constant term multiplying $E^\pi[\tau(\pi) - K + 1|C]$ in (4.97) is finite; further, it is not a function of either ϵ or of $\pi \in \Pi(\epsilon)$. The finiteness of this constant follows from the following observation: denote by μ_C^a the stationary distribution of the transition probability matrix P_C^a (i.e., $\mu_C^a = \mu_1$ for $a = h$ and $\mu_C^a = \mu_2$ for all $a \neq h$). An application of the ergodic theorem to the Markov process of arm a yields

$$D((P_C^a)^{d_a}(\cdot|i_a) \| (P_{C'}^a)^{d_a}(\cdot|i_a)) \longrightarrow D(\mu_C^a \| \mu_{C'}^a) < \infty \quad \text{as } d_a \to \infty. \tag{4.98}$$

Since every convergent sequence is bounded, we may write $D((P_C^a)^{d_a}(\cdot|i_a) \| (P_{C'}^a)^{d_a}(\cdot|i_a)) \leq M$ for all $(\underline{d}, \underline{i}, a) \in \mathbb{S} \times \mathcal{A}$, where $0 < M < \infty$. Using (4.95), it follows that the constant term multiplying $E^\pi[\tau(\pi) - K + 1|C]$ in (4.97) is bounded above by M. We also note that the first term inside the braces in (4.97) does not depend on ϵ. Since $d(\epsilon, 1-\epsilon) \to d(0,1) = +\infty$ as $\epsilon \downarrow 0$, the boundedness of the constant multiplying $E^\pi[\tau(\pi) - K + 1|C]$ implies that $\epsilon \downarrow 0$ is equivalent

to $E^{\pi}[\tau(\pi)|C] \to \infty$ for all $\pi \in \Pi(\epsilon)$. Letting $\epsilon \downarrow 0$, and using $d(\epsilon, 1-\epsilon)/\log(1/\epsilon) \longrightarrow 1$ as $\epsilon \downarrow 0$, we arrive at the lower bound in (4.11). This completes the proof of the proposition.

4.6.4 Proof of Proposition 12

We first state the analogues of [45, Assumptions A1-A5 and A6.1] as applicable to the context of this chapter and then verify each of the assumptions. Throughout this section, we let $C_0 = (h, P_1, P_2)$ denote the underlying arms configuration.

- The analogue of [45, Assumption A1] is that for all arms configuration C, $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$, and $b \in \mathcal{A}$, the transition probabilities $Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b)$ (defined in (4.5)) are continuous in b and C.

- The analogue of [45, Assumption A2] simply states that the actual transition probabilities are those corresponding to the underlying arms configuration C_0.

- The analogue of [45, Assumption A3] is that if $Q_{C_0}(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b) = 0$ for some $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$ and $b \in \mathcal{A}$, then $Q_C(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b') = 0$ for all C and $b' \in \mathcal{A}$. Further, if $Q_{C_0}(\underline{d}', \underline{i}' | \underline{d}, \underline{i}, b) > 0$ for some $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$ and $b \in \mathcal{A}$, then there exists $\bar{\epsilon} > 0$ independent of $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$ and $b \in \mathcal{A}$ such that for all C, the relation

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

  holds.

- The analogue of [45, Assumption A4] is that for all $b \in \mathcal{A}$ and arms configuration C, the controlled Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ with the control process $\{B_t : t \geq 0\}$ given by $B_t = b$ for all t, and transition probabilities given by $Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b)$, is a positive recurrent Markov process whose state space $\mathbb{S}$ is a single communicating class.

- Instead of stating the analogue of [45, Assumption A5], we state the analogue of its equivalent form [51, Condition C2]; for a proof of this equivalence, we refer the reader to [51] and the references therein. The analogue of [51, Condition C2] as applicable to the context of this chapter is that there is a finite set K, an integer $d \geq 1$, and a number $\rho > 0$ such that under every SRS policy π^{λ}, the probability of the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ starting from any state $(\underline{d}, \underline{i}) \in \mathbb{S}$ and hitting the set K after d time steps is lower bounded by ρ.

- The analogue of [45, Assumption A6.1] is that for any two arm configurations $C \neq C'$, there exists $(\underline{d}, \underline{i}) \in \mathbb{S}$ (possibly depending on C and C') such that for every $b \in \mathcal{A}$,

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

We now proceed to verify each of the above stated assumptions. From (4.5), it is clear that the transition probabilities $Q_C(\underline{d}', \underline{i}'|\underline{d}, \underline{i}, b)$ are continuous in C and b. This verifies the analogue of [45, Assumption A1].

Next, we note that if $Q_{C_0}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) = 0$ for some $(\underline{d}', \underline{i}'), (\underline{d}, \underline{i}) \in \mathbb{S}$ and $b \in \mathcal{A}$, then it follows from (4.5) that one of the following conditions must hold:

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

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

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

- $\mathbb{I}_{\{d'_a=1 \text{ and } d'_{\tilde{a}}=d_{\tilde{a}}+1 \text{ for all } \tilde{a}\neq a\}} \cdot \mathbb{I}_{\{i'_{\tilde{a}}=i_{\tilde{a}} \text{ for all } \tilde{a}\neq a\}} = 1$ for some $a \in \mathcal{A}$. In this case, $(\underline{d}', \underline{i}')$ is a valid state that can be reached in one step from the state $(\underline{d}, \underline{i})$. From (4.5), it follows that

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

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

Now, suppose that $Q_{C_0}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) > 0$ for some $(\underline{d}', \underline{i}'), (\underline{d}, \underline{i}) \in \mathbb{S}$ and $b \in \mathcal{A}$. From (4.22), this implies that there exists $a \in \mathcal{A}$ such that $P^{d_a}_{C_0}(i'_a|i_a) > 0$. From Assumption 2, we then have $P^{d_a}_{C_0}(i'_a|i_a) > \bar{\varepsilon}^*$, which together with (4.22) gives us that $Q_{C_0}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) \geq \frac{\eta}{K} \cdot \bar{\varepsilon}^*$. Setting $\bar{\epsilon} = \frac{\eta}{K} \cdot \bar{\varepsilon}^*$, it follows that

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

This verifies the analogue of [45, Assumption A3].

Next, from Lemma 9 of Chapter 3 , we know that under the trembling hand model and under every SRS policy π^λ, the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is ergodic, i.e., irreducible, aperiodic and positive recurrent. This verifies the analogue of [45, Assumption A4].

To verify the analogue of [51, Condition C2], let $\underline{d}' = (K, \ldots, 1)$ and $\underline{i}' = (i, \ldots, i)$, where $i \in \mathcal{S}$ is a fixed state. We now show that given any SRS policy π^λ, the probability of reaching

the state $(\underline{d}', \underline{i}')$ starting from any state $(\underline{d}, \underline{i})$ is lower bounded by a strictly positive constant, say $\rho > 0$, that is independent of $(\underline{d}, \underline{i})$ and the underlying arms configuration C_0. Note that the Markov process of each arm evolves on the common, finite state space $\mathcal{S}$. Therefore, from [38, Proposition 1.7], we have that there exists M sufficiently large such that

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

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

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

$$\begin{aligned}
&P^{\pi^\lambda}(\underline{d}(T_0 + M + K + 1) = \underline{d}', \underline{i}(T_0 + M + K + 1) = \underline{i}' \mid \underline{d}(T_0) = \underline{d}, \underline{i}(T_0) = \underline{i}, C_0) \\
&\geq \prod_{t=T_0+1}^{T_0+K} P^{\pi^\lambda}(A_t|\underline{d}(t), \underline{i}(t), C_0) \cdot \prod_{t=T_0+M+1}^{T_0+M+K+1} P^{\pi^\lambda}(A_t|\underline{d}(t), \underline{i}(t), C_0) \cdot \prod_{a=1}^{K} (P_{C_0}^a)^M(i|s_a) \\
&\overset{(a)}{\geq} \prod_{t=T_0+1}^{T_0+K} \frac{\eta}{K} \cdot \prod_{t=T_0+M+1}^{T_0+M+K+1} \frac{\eta}{K} \cdot \prod_{a=1}^{K} (P_{C_0}^a)^M(i|s_a) \\
&= \left(\frac{\eta}{K}\right)^{2K} \prod_{a=1}^{K} (P_{C_0}^a)^M(i|s_a),
\end{aligned} \tag{4.99}$$

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

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

From Assumption 2 we know that $P_1, P_2 \in \mathscr{P}(\bar{\varepsilon}^*)$ for some $\bar{\varepsilon}^* > 0$. This implies that

$(P^a_{C_0})^M(i|s_a) > \bar{\varepsilon}^*$ for all $a \in \mathcal{A}$. Using this in (4.99), and setting $\rho = \left(\frac{\eta}{K}\right)^{2K} (\bar{\varepsilon}^*)^K$, we have

$$P^{\pi^\lambda}(\underline{d}(T_0+M+K+1) = \underline{d}', \underline{i}(T_0+M+K+1) = \underline{i}' \mid \underline{d}(T_0) = \underline{d}, \underline{i}(T_0) = \underline{i}, C_0) \geq \rho \quad \forall\, (\underline{d}, \underline{i}) \in \mathbb{S}. \tag{4.100}$$

This verifies the analogue of [51, Condition C2].

Lastly, in order to verify the analogue of [45, Assumption A6.1], we show that a condition stronger than [45, Assumption A6.1], one that is the equivalent of Mandl's identifiability condition for countable-state controlled Markov processes, holds in the context of this chapter. This is demonstrated in the following lemma.

Lemma 25. *For all arm configurations $C = (h, P_1, P_2)$ and $C' = (h', P_1', P_2')$ such that $C \neq C'$ (i.e., C and C' differ in at least one component), there exists $(\underline{d}, \underline{i}) \in \mathbb{S}$, possibly depending on C and C', such that*

$$Q_C(\cdot \mid \underline{d}, \underline{i}, b) \neq Q_{C'}(\cdot \mid \underline{d}, \underline{i}, b) \quad \textit{for all } b \in \mathcal{A}. \tag{4.101}$$

Proof of Lemma 25. We present the proof of the lemma under various cases.

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

We first consider the case when C and C' differ only in their first components. Recall that the transition probability matrices P_1 and P_2 satisfy the condition $P_2 \neq P_1$. This means that there exist $i^*, j^* \in \mathcal{S}$ such that $P_1(j^*|i^*) \neq P_2(j^*|i^*)$. Fix $a^* = h$. Let $(\underline{d}, \underline{i}) \in \mathbb{S}$ be such that $d_{a^*} = 1$ and $i_{a^*} = i^*$. Also, let $(\underline{d}', \underline{i}')$ be such that $d'_{a^*} = 1$, $d'_a = d_a + 1$ for all $a \neq a^*$, $i'_{a^*} = j^*$ and $i'_a = i_a$ for all $a \neq a^*$. Then, for the above choices of $(\underline{d}, \underline{i})$ and $(\underline{d}', \underline{i}')$, it follows from (4.5) that for all $b \in \mathcal{A}$,

$$\begin{aligned} Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1-\eta)\, \mathbb{I}_{\{b=a^*\}}\right) P_1(j^*|i^*), \\ Q_{C'}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1-\eta)\, \mathbb{I}_{\{b=a^*\}}\right) P_2(j^*|i^*). \end{aligned} \tag{4.102}$$

Because $P_1(j^*|i^*) \neq P_2(j^*|i^*)$, it follows that $Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) \neq Q_{C'}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b)$ for all $b \in \mathcal{A}$. This establishes (4.101).

4.6.4.2 Case 2: $C = (h, P_1, P_2)$, $C' = (h, P_1', P_2)$, where $P_1 \neq P_1'$

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

$$\begin{aligned} Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=a^*\}}\right) P_1(j^*|i^*), \\ Q_{C'}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=a^*\}}\right) P_1'(j^*|i^*) \end{aligned} \tag{4.103}$$

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

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

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

$$\begin{aligned} Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=a^*\}}\right) P_2(j^*|i^*), \\ Q_{C'}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=a^*\}}\right) P_2'(j^*|i^*) \end{aligned} \tag{4.104}$$

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

4.6.4.4 Case 4: $C = (h, P_1, P_2)$, $C' = (h', P_1', P_2)$, where $h' \neq h$, $P_1 \neq P_1'$

In this case, because C' is a valid arms configuration, it follows that $P_1' \neq P_2$. The proof for this case then follows along the lines of that for Case 1, with $a^* = h'$ and the following modifications:

$$\begin{aligned} Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=a^*\}}\right) P_2(j^*|i^*), \\ Q_{C'}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=a^*\}}\right) P_1'(j^*|i^*) \end{aligned} \tag{4.105}$$

for all $b \in \mathcal{A}$, where $i^*, j^* \in \mathcal{S}$ are chosen such that $P_1'(j^*|i^*) \neq P_2(j^*|i^*)$.

4.6.4.5 Case 5: $C = (h, P_1, P_2)$, $C' = (h', P_1, P_2')$, where $h' \neq h$, $P_2 \neq P_2'$

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

$$\begin{aligned} Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=a^*\}}\right) P_2(j^*|i^*), \\ Q_{C'}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=a^*\}}\right) P_2'(j^*|i^*) \end{aligned} \tag{4.106}$$

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

4.6.4.6 Case 6: $C = (h, P_1, P_2)$, $C' = (h, P_1', P_2')$, where $P_1 \neq P_1'$, $P_2 \neq P_2'$

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

$$\begin{aligned} Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=a^*\}}\right) P_1(j^*|i^*), \\ Q_{C'}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=a^*\}}\right) P_1'(j^*|i^*) \end{aligned} \tag{4.107}$$

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

4.6.4.7 Case 7: $C = (h, P_1, P_2)$, $C' = (h', P_1', P_2')$, where $h' \neq h$, $P_1 \neq P_1'$, $P_2 \neq P_2'$

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

$$\begin{aligned} Q_C(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=a^*\}}\right) P_1(j^*|i^*), \\ Q_{C'}(\underline{d}', \underline{i}' \mid \underline{d}, \underline{i}, b) &= \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=a^*\}}\right) P_1'(j^*|i^*) \end{aligned} \tag{4.108}$$

for all $b \in \mathcal{A}$, where $i^*, j^* \in \mathcal{S}$ are chosen such that $P_1(j^*|i^*) \neq P_1'(j^*|i^*)$. This completes the proof of the lemma and also the verification of the analogue of [45, Assumption A6.1]. □

Finally, with the analogues of [45, Assumptions A1-A5 and A6.1] being verified, we apply [45, Theorem 4.3] (which simply states that under [45, Assumptions A1-A5 and A6.1], the ML estimates converge to their true values almost surely) to deduce that under the arms configuration C_0,

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

thus proving Proposition 12. We note here that in addition to [45, Assumptions A1-A5 and A6.1], the proof of [45, Theorem 4.3] uses the notion of "$\{\varepsilon_i\}$-randomisation" which, for controlled Markov processes, ensures that the probability of selecting any control at any given time is strictly positive. The trembling hand model (4.1) considered here guarantees that the probability of sampling any arm at any given time is $\geq \frac{\eta}{K} > 0$, thus alleviating the need to consider $\{\varepsilon_i\}$ randomisation in our work.

4.6.5 Proof of Proposition 13

We first note the following important points.

- From the exposition in Section 3.8.3, we note that

$$\liminf_{n\to\infty} \frac{N(n,\underline{d},\underline{i},a,j)}{n} > 0, \quad \liminf_{n\to\infty} \frac{N(n,\underline{d},\underline{i},a)}{n} > 0 \quad \text{almost surely} \tag{4.109}$$

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

$$\frac{N(n,\underline{d},\underline{i},a,j)}{N(n,\underline{d},\underline{i},a)} \longrightarrow (P_C^a)^{d_a}(j \mid i_a) \quad \text{as } n \to \infty. \tag{4.110}$$

- Let $\mathbb{S} = \{(\underline{d}^{(1)},\underline{i}^{(1)}),(\underline{d}^{(2)},\underline{i}^{(2)}),\ldots\}$ denote an enumeration of the countably infinite set $\mathbb{S}$. For each $l \in \{1,2,\ldots\}$ and $a \in \mathcal{A}$, the sequence

$$\left\{\frac{N(n,\underline{d}^{(l)},\underline{i}^{(l)},a)}{n}\right\}_{n\geq K} \tag{4.111}$$

  is bounded almost surely. This implies that there exists a null set $B \subset \Omega$ such that for every $\omega \notin B$ and $l \in \{1,2,\ldots\}$, there exists a subsequence, say $\{n_k(\omega,\underline{d}^{(l)},\underline{i}^{(l)}) : k \in \{1,2,\ldots\}\}$, along which (4.111) converges. Using Tychonoff's theorem [52, Theorem 37.3], we get that for all $l \in \{1,2,\ldots\}$, (4.111) converges along the 'diagonal' subsequence $\{n_k(\omega,\underline{d}^{(k)},\underline{i}^{(k)}) : k \in \{1,2,\ldots\}\}$. Fixing attention to this subsequence and relabeling it as $\{n_k(\omega) : k \geq 1\}$, we have

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

  where the limit $\alpha(\underline{d},\underline{i},a,\omega) > 0$ for all $(\underline{d},\underline{i}) \in \mathbb{S}$, $a \in \mathcal{A}$, and $\omega \notin B$, thanks to (4.109).

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

$$\frac{N(n_k(\omega),\underline{d},\underline{i},a,j)(\omega)}{n_k(\omega)} \longrightarrow \alpha(\underline{d},\underline{i},a,\omega)\cdot(P_C^a)^{d_a}(j \mid i_a) \quad \text{as} \quad k \to \infty. \tag{4.112}$$

The following lemma shows that the convergence in (4.112) is, in fact, uniform in $(\underline{d},\underline{i})$.

Lemma 26. *Let $\mathscr{S} = \{s_1, s_2, \ldots\}$ be a countable set, and let $f_k, f : \mathscr{S} \to [0,1]$, $k \geq 1$, be such that*

$$0 < D = \sum_{i=1}^{\infty} f_k(s_i) < \infty \quad \forall\ k \geq 1, \quad 0 < \sum_{i=1}^{\infty} f(s_i) < \infty.$$

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

Proof of Lemma 26. Without loss of generality, suppose that $\sum_{i=1}^{\infty} f_k(s_i) = 1$ for all $k \geq 1$ and $\sum_{i=1}^{\infty} f(s_i) = 1$. Fix $\varepsilon > 0$, and choose $N = N(\varepsilon)$ sufficiently large so that

$$\sum_{i=1}^{N} f(s_i) \geq 1 - \varepsilon. \tag{4.113}$$

Then, on the finite set $\{s_1, \ldots, s_N\}$, we have have that $f_k \to f$ uniformly. This implies that there exists M sufficiently large such that for all $k \geq M$ and for all $i \in \{1, 2, \ldots, N\}$, we have

$$|f_k(s_i) - f(s_i)| \leq \frac{\varepsilon}{N}. \tag{4.114}$$

From (4.113) and (4.114), it follows that for all $k \geq M$,

$$\begin{aligned}
\sum_{i=1}^{N} f_k(s_i) &= \sum_{i=1}^{N} f(s_i) + \sum_{i=1}^{N} (f_k(s_i) - f(s_i)) \\
&\geq \sum_{i=1}^{N} f(s_i) + \sum_{i=1}^{N} |f_k(s_i) - f(s_i)| \\
&\geq 1 - \varepsilon - \varepsilon \\
&= 1 - 2\varepsilon.
\end{aligned} \tag{4.115}$$

This implies that for all $k \geq M$ and for all $i > N$,

$$0 \leq f_k(s_i) \leq 2\varepsilon, \quad 0 \leq f(s_i) \leq \varepsilon, \tag{4.116}$$

as a result of which it follows that $|f_k(s_i) - f(s_i)| \leq \varepsilon$ for all $k \geq M$ and for all $i > N$. Combining this with (4.113) yields the desired result. □

We now begin the proof of Proposition 13. Fix an arbitrary null set $B \subset \Omega$. Then, for every

$\omega \notin B$, there exists a subsequence $\{n_k(\omega) : k \geq 1\}$ such that

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

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

$$\lim_{k\to\infty} \frac{M_{hh'}(n_k(\omega))(\omega)}{n_k(\omega)} = \lim_{k\to\infty} \frac{T_1(n_k(\omega)) + T_2(n_k(\omega)) + T_3(n_k(\omega)) + T_4(n_k(\omega))}{n_k(\omega)}. \tag{4.117}$$

We shall demonstrate that the right hand side of (4.117) is strictly positive for all $\omega \notin B$. Fix an arbitrary $\varepsilon \in (0, 1)$.

- From (4.38), we have

$$\begin{aligned}
&\frac{T_1(n_k(\omega))}{n_k(\omega)} \\
&= \frac{1}{n_k(\omega)} \log \mathbb{E}\Bigg[\exp\Bigg\{ \log\Bigg(\sum_{i\in\mathcal{S}} \phi(i)\, P^{h-1}(X_{h-1}^h|i) \Bigg) \\
&\qquad\qquad + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \log P^{d_h}(j|i_h) \Bigg\}\Bigg] \\
&= \frac{1}{n_k(\omega)} \log \mathbb{E}\Bigg[\exp\Bigg\{ \log\Bigg(\sum_{i\in\mathcal{S}} \phi(i)\, P^{h-1}(X_{h-1}^h|i) \Bigg) \\
&\qquad\qquad + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \log \frac{P^{d_h}(j|i_h)}{P_1^{d_h}(j|i_h)} \Bigg\}\Bigg] \\
&\qquad\qquad + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \log P_1^{d_h}(j|i_h).
\end{aligned} \tag{4.118}$$

  Noting that the expectation in the first term of (4.118) is of a non-negative (exponential) function, we may lower bound this expectation term by

$$\mathbb{E}\Bigg[\exp\Bigg\{ \log\Bigg(\sum_{i\in\mathcal{S}} \phi(i)\, P^{h-1}(X_{h-1}^h|i) \Bigg) + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \log \frac{P^{d_h}(j|i_h)}{P_1^{d_h}(j|i_h)} \Bigg\} \cdot \mathbb{I}\Big(P \in \mathscr{P}(\bar{\varepsilon}^*) \Big) \Bigg]$$

$$\geq \exp\left\{ \log(\phi_{\mathsf{min}} \cdot \bar{\varepsilon}^*) + (\log \bar{\varepsilon}^*) \cdot \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \right\} \cdot \mathbb{P}(P \in \mathscr{P}(\bar{\varepsilon}^*))$$

$$\geq \exp\left\{ \log \phi_{\mathsf{min}} + \log \bar{\varepsilon}^* + \log \bar{\varepsilon}^* \right\} \cdot \mathbb{P}(P \in \mathscr{P}(\bar{\varepsilon}^*)). \tag{4.119}$$

In (4.119), $\mathbb{P}(\cdot) = \mathbb{E}[\mathbb{I}(\cdot)]$, where the first inequality follows by noting that on the set $\{P \in \mathscr{P}(\bar{\varepsilon}^*)\}$, we have[1]

$$P^d(j|i) > 0,\ P_1^d(j|i) > 0 \implies \frac{P^d(j|i)}{P_1^d(j|i)} \geq \bar{\varepsilon}^* \quad \text{for all } d \geq 1,\ i, j \in \mathcal{S}.$$

Also, in the first inequality above, $\phi_{\mathsf{min}} = \min_{i \in \mathcal{S}} \phi(i) > 0$. The second inequality above follows by using

$$\sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \leq 1.$$

Applying logarithm to both sides of (4.119), we get

$$\frac{T_1(n_k(\omega))}{n_k(\omega)} \geq \frac{\log \phi_{\mathsf{min}} + 2 \log \bar{\varepsilon}^* + \log \mathbb{P}(P \in \mathscr{P}(\bar{\varepsilon}^*))}{n_k(\omega)} + \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \log P_1^{d_h}(j|i_h). \tag{4.120}$$

We now claim that $\mathbb{P}(P \in \mathscr{P}(\bar{\varepsilon}^*)) > 0$, and therefore the limiting value of the first term in (4.120) is equal to zero. To verify the claim, note that[2]

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

is an open set (with respect to the relative topology on $\mathscr{P}(\mathcal{S})$ induced by the topology arising from the Euclidean metric on $\mathbb{R}^{|\mathcal{S}|(|\mathcal{S}|-1)}$). Also, $D(\mathscr{B}(\bar{\varepsilon}^*)) > 0$[3]. Therefore, we have

$$\mathbb{P}(P \in \mathscr{P}(\bar{\varepsilon}^*)) \geq \mathbb{P}(P \in \mathscr{B}(\bar{\varepsilon}^*))$$

[1]Note that when $P \sim D$, where D is the prior on $\mathscr{P}(\mathcal{S})$, we have $D(P^d(j|i) > 0 \text{ for all } d \geq 1,\ i, j \in \mathcal{S}) = 1$.

[2]In order for the set $\mathscr{B}(\bar{\varepsilon}^*)$ to be non-empty, it is important that $\bar{\varepsilon}^* < \frac{1}{|\mathcal{S}|}$. If this is not true, $\bar{\varepsilon}^*$ may be replaced by $\frac{\bar{\varepsilon}^*}{|\mathcal{S}|}$ without altering the proof.

[3]Recall that the prior D is induced by sampling each row of $P \in \mathscr{P}(\mathcal{S})$ independently according to $\text{Dir}(\mathbf{1})$. Because $\text{Dir}(\mathbf{1})$ is simply the uniform probability distribution on the probability simplex $\mathscr{P}(\mathcal{S})$, every open set has a strictly positive probability under D.

$$= D(\mathscr{B}(\bar{\varepsilon}^*))$$
$$> 0. \tag{4.121}$$

Thus, there exists $K_{11} = K_{11}(\varepsilon, \omega)$ such that for all $k \geq K_{11}$, we have

$$\frac{\log \phi_{\mathsf{min}} + 2\log \bar{\varepsilon}^* + \log \mathbb{P}(P \in \mathscr{P}(\bar{\varepsilon}^*))}{n_k(\omega)} \geq -\varepsilon. \tag{4.122}$$

We now turn our attention to the second term in (4.120). From the convergence in (4.112), we know that given any $\varepsilon' > 0$, there exists $K_{12} = K_{12}(\varepsilon', \omega)$ (the dependence of K_{12} on $(\underline{d}, \underline{i}) \in \mathbb{S}$ is removed thanks to Lemma 26) such that under the arms configuration $C = (h, P_1, P_2)$,

$$\frac{N(n_k(\omega), \underline{d}, \underline{i}, h, j, \omega)}{n_k(\omega)} \leq \alpha(\underline{d}, \underline{i}, h, \omega) \cdot P_1^{d_h}(j|i_h) \cdot (1 + \varepsilon') \tag{4.123}$$

for all $k \geq K_{12}$. Combining (4.122) and (4.123), we get that for all $k \geq K_1 = \max\{K_{11}, K_{12}\}$,

$$\frac{T_1(n_k(\omega))}{n_k(\omega)} \geq (1 + \varepsilon') \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \alpha(\underline{d}, \underline{i}, h, \omega)\ P_1^{d_h}(j|i_h)\ \log P_1^{d_h}(j|i_h) - \varepsilon. \tag{4.124}$$

Choose ε' so that

$$\frac{T_1(n_k(\omega))}{n_k(\omega)} \geq \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \alpha(\underline{d}, \underline{i}, h, \omega)\ P_1^{d_h}(j|i_h)\ \log P_1^{d_h}(j|i_h) - 2\,\varepsilon. \tag{4.125}$$

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

$$\frac{T_2(n_k(\omega))}{n_k(\omega)} \geq \sum_{a \neq h} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \alpha(\underline{d}, \underline{i}, a, \omega)\ P_2^{d_a}(j|i_a)\ \log P_2^{d_a}(j|i_a) - 2\,\varepsilon. \tag{4.126}$$

- We now handle the term $\frac{T_3(n_k(\omega))}{n_k(\omega)}$. Note that under the arms configuration $C = (h, P_1, P_2)$, it follows from (4.43) that there exists $K_3 = K_3(\varepsilon, \omega)$ such that

$$\frac{T_3(n_k(\omega))}{n_k(\omega)} \geq \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \alpha(\underline{d}, \underline{i}, h', j, \omega)\ P_2^{d_{h'}}(j|i_{h'})\ \log \frac{1}{P_2^{d_{h'}}(j|i_{h'})} - \varepsilon \tag{4.127}$$

for all $k \geq K_3$ (in arriving at (4.127), the first term in (4.40), being non-negative, is lower bounded by 0). Also, there exists $K_4 = K_4(\varepsilon, \omega)$ such that

$$\frac{T_4(n_k(\omega))}{n_k(\omega)} \geq \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \alpha(\underline{d}, \underline{i}, h, \omega)\ P_1^{d_h}(j|i_h)\ \log \frac{1}{P^{d_h}(j|i_h)} + \sum_{a \neq h,h'} \alpha(\underline{d}, \underline{i}, a, \omega)\ P_2^{d_a}(j|i_a)\ \log \frac{1}{P^{d_a}(j|i_a)} - \varepsilon. \tag{4.128}$$

for all $k \geq K_4$.

Combining the inequalities in (4.125)-(4.128), we see that for all $\varepsilon > 0$ and for all $\omega \notin B$,

$$\frac{M_{hh'}(n_k(\omega))}{n_k(\omega)} \geq \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \alpha(\underline{d}, \underline{i}, h, \omega) \cdot D(P_1^{d_h}(\cdot|i_h) \| P^{d_h}(\cdot|i_h)) + \sum_{a \neq h,h'} \alpha(\underline{d}, \underline{i}, a, \omega) \cdot D(P_2^{d_a}(\cdot|i_a) \| P^{d_a}(\cdot|i_a)) - 6\varepsilon \tag{4.129}$$

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

$$\begin{aligned} &\liminf_{n \to \infty} \frac{M_{hh'}(n)(\omega)}{n} \\ &= \lim_{k \to \infty} \frac{M_{hh'}(n_k(\omega))}{n_k(\omega)} \\ &\geq \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \alpha(\underline{d}, \underline{i}, h, \omega) \cdot D(P_1^{d_h}(\cdot|i_h) \| P^{d_h}(\cdot|i_h)) + \sum_{a \neq h,h'} \alpha(\underline{d}, \underline{i}, a, \omega) \cdot D(P_2^{d_a}(\cdot|i_a) \| P^{d_a}(\cdot|i_a)) - 6\varepsilon. \end{aligned} \tag{4.130}$$

Letting $\varepsilon \downarrow 0$ in (4.130), we get

$$\begin{aligned} &\liminf_{n \to \infty} \frac{M_{hh'}(n)(\omega)}{n} \\ &\geq \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \alpha(\underline{d}, \underline{i}, h, \omega) \cdot D(P_1^{d_h}(\cdot|i_h) \| P^{d_h}(\cdot|i_h)) + \sum_{a \neq h,h'} \alpha(\underline{d}, \underline{i}, a, \omega) \cdot D(P_2^{d_a}(\cdot|i_a) \| P^{d_a}(\cdot|i_a)) \end{aligned} \tag{4.131}$$

for all $\omega \notin B$. Because $\alpha(\underline{d}, \underline{i}, a, \omega) > 0$ for all $(\underline{d}, \underline{i}) \in \mathbb{S}$, $a \in \mathcal{A}$, and $\omega \notin B$, the right hand side of (4.131) is strictly positive. This establishes (4.44).

4.6.6 Proof of Proposition 14

The policy $\pi_2^\star(L, \delta)$ commits error if one of the following events is true:

1. The policy never stops in finite time.

2. The policy stops in finite time and, when $C = (h, P_1, P_2)$ is the actual (underlying) arms configuration, declares $h' \neq h$ as the true index of the odd arm.

The event in item 1 above has zero probability thanks to Proposition 13. Thus, the probability of error of policy $\pi = \pi_2^\star(L, \delta)$, which we denote by P_e^π, may be evaluated as follows: suppose $C = (h, P_1, P_2)$ is the underlying arms configuration. Then,

$$P_e^\pi = P^\pi(\theta(\tau(\pi)) \neq h | C) = P^\pi\left(\exists\ n \text{ and } \ h' \neq h \text{ such that } \theta(\tau(\pi)) = h' \text{ and } \tau(\pi) = n \middle| C\right). \tag{4.132}$$

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

$$\begin{aligned}
P_e^\pi &= P^\pi\left(\bigcup_{h' \neq h} \bigcup_{n=0}^{\infty} \mathcal{R}_{h'}(n) \middle| C\right) \\
&= \sum_{h' \neq h} \sum_{n=0}^{\infty} P^\pi(\tau(\pi) = n, \theta(n) = h' | C) \\
&= \sum_{h' \neq h} \sum_{n=0}^{\infty} \int\limits_{\mathcal{R}_{h'}(n)} dP^\pi(\omega | C) \\
&= \sum_{h' \neq h} \sum_{n=0}^{\infty} \int\limits_{\mathcal{R}_{h'}(n)} f(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) | C)\ d(B^n(\omega), A^n(\omega), \bar{X}^n(\omega)) \\
&\overset{(a)}{\leq} \sum_{h' \neq h} \sum_{n=0}^{\infty} \int\limits_{\mathcal{R}_{h'}(n)} \hat{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) | H_h)\ d(B^n(\omega), A^n(\omega), \bar{X}^n(\omega)) \\
&= \sum_{h' \neq h} \sum_{n=0}^{\infty} \left\{ \int\limits_{\mathcal{R}_{h'}(n)} \frac{\hat{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) | H_h)}{\bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) | H_{h'})} \right. \\
&\qquad \left. \cdot\, \bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) | H_{h'})\ d(B^n(\omega), A^n(\omega), \bar{X}^n(\omega)) \right\} \\
&= \sum_{h' \neq h} \sum_{n=0}^{\infty} \int\limits_{\mathcal{R}_{h'}(n)} e^{-M_{h'h}(n)(\omega)}\ \bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) | H_{h'})\ d(B^n(\omega), A^n(\omega), \bar{X}^n(\omega))
\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{\leq} \sum_{h' \neq h} \sum_{n=0}^{\infty} \int\limits_{\mathcal{R}_{h'}(n)} \frac{1}{(K-1)L}\, \bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) | \mathcal{H}_{h'})\, d(B^n(\omega), A^n(\omega), \bar{X}^n(\omega)) \\
&= \sum_{h' \neq h} \frac{1}{(K-1)L}\, \bar{P}^{\pi}\left(\bigcup_{n=0}^{\infty} \mathcal{R}_{h'}(n) \middle| \mathcal{H}_{h'} \right) \\
&\leq \frac{1}{L}, \qquad (4.133)
\end{aligned}
$$

where (a) follows from the definition of the maximum likelihood $\hat{f}$, (b) follows from noting that when $\pi_2^\star(L, \delta)$ stops at time n and outputs h' as the odd arm, we must have $M_{h'h}(n) \geq \log((K-1)L)$ almost surely, and $\bar{P}^\pi$ in (4.133) denotes the probability measure under the average likelihood $\bar{f}$. Setting $L = 1/\epsilon$ yields the desired result.

4.6.7 Proof of Proposition 15

The convergences in (4.50) imply that for each $a \in \mathcal{A}$ and $(\underline{d}, \underline{i}) \in \mathbb{S}$, almost surely,

$$
\begin{aligned}
P(A_n = a \mid \underline{d}(n) = \underline{d},\ \underline{i}(n) = \underline{i},\ \mathcal{F}_{n-1}) &= \frac{\eta}{K} + (1-\eta)\, \lambda_{\theta(n), \hat{P}_{\theta(n),1}(n), \hat{P}_{\theta(n),2}(n), \delta}(a \mid \underline{d}, \underline{i}) \\
&\longrightarrow \frac{\eta}{K} + (1-\eta)\, \lambda_{h, P_1, P_2, \delta}(a \mid \underline{d}, \underline{i}) \quad \text{as } n \to \infty. \qquad (4.134)
\end{aligned}
$$

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

Lemma 27. *For each $(\underline{d}', \underline{i}') \in \mathbb{S}$, the mapping $\lambda \mapsto \mu^\lambda(\underline{d}', \underline{i}')$ is continuous.*

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

$$
\begin{aligned}
&P^{\pi^\lambda}(\underline{d}(t+1) = \underline{d}',\ \underline{i}(t+1) = \underline{i}' \mid \underline{d}(t) = \underline{d},\ \underline{i}(t) = \underline{i},\ C) \\
&= \begin{cases}
\left(\frac{\eta}{K} + (1-\eta)\, \lambda(a | \underline{d}, \underline{i}) \right) (P_C^a)^{d_a}(i'_a | i_a), & \text{if } d'_a = 1 \text{ and } d'_{\tilde{a}} = d_{\tilde{a}} + 1 \text{ for all } \tilde{a} \neq a, \\
& i'_{\tilde{a}} = i_{\tilde{a}} \text{ for all } \tilde{a} \neq a, \\
0, & \text{otherwise.}
\end{cases} \qquad (4.135)
\end{aligned}
$$

It is clear that the one-step transition probabilities in (4.135) are continuous in λ for all $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$. As a consequence, it follows that for each $t \geq 1$, the t-step transition probabilities derived from the one-step transition probabilities in (4.135) are continuous in λ

for all $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$. Denoting the t-step transition probability from the state $(\underline{d}, \underline{i})$ to the state $(\underline{d}', \underline{i}')$ under π^λ by $P^t_{(\underline{d},\underline{i}),\ (\underline{d}',\underline{i}')}(\lambda)$, we invoke [45, Assumption A5′], an assumption that is equivalent to [45, Assumption A5], to deduce that for each $(\underline{d}', \underline{i}') \in \mathbb{S}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} P^t_{(\underline{d},\underline{i}),\ (\underline{d}',\underline{i}')}(\lambda) = \mu^\lambda(\underline{d}', \underline{i}') \quad \text{uniformly in } (\underline{d}, \underline{i}) \in \mathbb{S} \text{ and } \lambda. \tag{4.136}$$

Combining (a) the continuity of $P^t_{(\underline{d},\underline{i}),\ (\underline{d}',\underline{i}')}(\lambda)$ in λ for all $t \geq 1$ and $(\underline{d}, \underline{i}), (\underline{d}', \underline{i}') \in \mathbb{S}$, and (b) the uniform convergence in (4.136), we get that $\mu^\lambda(\underline{d}', \underline{i}')$ is continuous in λ for each $(\underline{d}', \underline{i}') \in \mathbb{S}$. This establishes the lemma. □

Lemma 27 implies that for every $(\underline{d}, \underline{i}) \in \mathbb{S}$, there exists an open neighbourhood $O = O(\underline{d}, \underline{i})$ around (P_1, P_2) such that for all $(P, Q) \in O$,

$$\mu^{\lambda_{h,P,Q,\delta}}(\underline{d}, \underline{i}) \geq \frac{\mu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d}, \underline{i})}{1+\delta}. \tag{4.137}$$

Also, Lemma 27 together with Proposition 12 implies that for each $(\underline{d}, \underline{i}) \in \mathbb{S}$, under the arms configuration $C = (h, P_1, P_2)$,

$$\mu^{\lambda_{\theta(n),\hat{P}_{\theta(n),1}(n),\hat{P}_{\theta(n),2}(n),\delta}}(\underline{d}, \underline{i}) \longrightarrow \mu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d}, \underline{i}) \quad \text{as} \quad n \to \infty \quad \text{almost surely.} \tag{4.138}$$

Because $(\theta(n), \hat{P}_{\theta(n),1}(n), \hat{P}_{\theta(n),2}(n)) \longrightarrow (h, P_1, P_2)$ as $n \to \infty$ almost surely under the policy $\pi^\star_{\mathsf{ns}}(L, \delta)$ and under the arms configuration $C = (h, P_1, P_2)$, it follows that for all n sufficiently large, $(\hat{P}_{\theta(n),1}(n), \hat{P}_{\theta(n),2}(n)) = (\hat{P}_{h,1}(n), \hat{P}_{h,2}(n)) \in O$ almost surely. Invoking [45, Lemma 2.10] with the bounded cost function $c(x_m, x_{m+1}, z_m)$ therein set as $c(x_m, x_{m+1}, z_m) = 1 - \mathbb{I}_{\{x_m = (\underline{d},\underline{i})\}}$, we get

$$\begin{aligned} \liminf_{n\to\infty} \frac{N(n, \underline{d}, \underline{i})}{n} &\geq \inf_{(P,Q)\in O} \mu^{\lambda_{h,P,Q,\delta}}(\underline{d}, \underline{i}) \\ &\geq \frac{\mu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d}, \underline{i})}{1+\delta} \quad \text{almost surely.} \end{aligned} \tag{4.139}$$

Here, $N(n, \underline{d}, \underline{i}) = \sum_{a=1}^{K} N(n, \underline{d}, \underline{i}, a)$. Invoking the null set B from the proof of Proposition 13 in Appendix 4.6.5, restricting the almost sure inequality in (4.139) to outside the null set B, and using the convergence in (4.134), we get that for all $\omega \notin B$, $(\underline{d}, \underline{i}) \in \mathbb{S}$ and $a \in \mathcal{A}$,

$$\alpha(\underline{d}, \underline{i}, a, \omega) \geq \left(\inf_{(P,Q)\in O} \mu^{\lambda_{h,P,Q,\delta}}(\underline{d}, \underline{i}) \right) \left(\frac{\eta}{K} + (1-\eta)\, \lambda_{h,P_1,P_2,\delta}(a \mid \underline{d}, \underline{i}) \right)$$

$$
\begin{aligned}
&\geq \frac{\mu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d},\underline{i})}{1+\delta}\left(\frac{\eta}{K}+(1-\eta)\,\lambda_{h,P_1,P_2,\delta}(a\mid \underline{d},\underline{i})\right)\\
&= \frac{\nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d},\underline{i},a)}{1+\delta}. \qquad (4.140)
\end{aligned}
$$

Plugging (4.140) into (4.131), we see that under the arms configuration $C=(h,P_1,P_2)$, for all $\omega\notin B$,

$$
\begin{aligned}
&\liminf_{n\to\infty}\frac{M_{hh'}(n)(\omega)}{n}\\
&\geq \frac{1}{1+\delta}\Bigg[\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\Bigg\{\nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d},\underline{i},h)\cdot D(P_1^{d_h}(\cdot|i_h)\|P^{d_h}(\cdot|i_h))\\
&\qquad\qquad +\sum_{a\neq h,h'}\nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d},\underline{i},a)\cdot D(P_2^{d_a}(\cdot|i_a)\|P^{d_a}(\cdot|i_a))\Bigg\}\Bigg]\\
&\geq \frac{1}{1+\delta}\inf_{P_2'}\Bigg[\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\Bigg\{\nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d},\underline{i},h)\cdot D(P_1^{d_h}(\cdot|i_h)\|(P_2')^{d_h}(\cdot|i_h))\\
&\qquad\qquad +\sum_{a\neq h,h'}\nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d},\underline{i},a)\cdot D(P_2^{d_a}(\cdot|i_a)\|(P_2')^{d_a}(\cdot|i_a))\Bigg\}\Bigg]\\
&= \frac{1}{1+\delta}\inf_{\substack{P_1',P_2':\\ P_1'\neq P_2'}}\Bigg[\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\Bigg\{\nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d},\underline{i},h)\cdot D(P_1^{d_h}(\cdot|i_h)\|(P_2')^{d_h}(\cdot|i_h))\\
&\qquad\qquad +\nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d},\underline{i},h')\cdot D(P_2^{d_{h'}}(\cdot|i_h)\|(P_1')^{d_{h'}}(\cdot|i_{h'}))\\
&\qquad\qquad +\sum_{a\neq h,h'}\nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d},\underline{i},a)\cdot D(P_2^{d_a}(\cdot|i_a)\|(P_2')^{d_a}(\cdot|i_a))\Bigg\}\Bigg]\\
&\geq \frac{1}{1+\delta}\inf_{\substack{C'=(h',P_1',P_2'):\\ h'\neq h,\\ P_1'\neq P_2'}}\Bigg[\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\Bigg\{\nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d},\underline{i},h)\cdot D(P_1^{d_h}(\cdot|i_h)\|(P_2')^{d_h}(\cdot|i_h))\\
&\qquad\qquad +\nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d},\underline{i},h')\cdot D(P_2^{d_{h'}}(\cdot|i_h)\|(P_1')^{d_{h'}}(\cdot|i_{h'}))\\
&\qquad\qquad +\sum_{a\neq h,h'}\nu^{\lambda_{h,P_1,P_2,\delta}}(\underline{d},\underline{i},a)\cdot D(P_2^{d_a}(\cdot|i_a)\|(P_2')^{d_a}(\cdot|i_a))\Bigg\}\Bigg]\\
&\geq \frac{R^*(h,P_1,P_2)}{(1+\delta)^2}, \qquad (4.141)
\end{aligned}
$$

where the last line above follows from the definition of $\lambda_{h,P_1,P_2,\delta}$ in (4.17). The desired result is thus established.

4.6.8 Proof of Proposition 16

It suffices to show that under the arms configuration $C = (h, P_1, P_2)$ and under the policy $\pi = \pi_2^\star(L, \delta)$,

$$\limsup_{L\to\infty} P^\pi(\tau(\pi) \le m \mid C) = 0 \quad \text{for all} \quad m \ge 1. \tag{4.142}$$

Because the policy $\pi_2^\star(L, \delta)$ selects arm 1 at time $t = 0$, arm 2 at time $t = 1$ and so on until arm K at time $t = K$, it suffices to show that for each $m \ge K$,

$$\limsup_{L\to\infty} P^\pi(\tau(\pi) \le m \mid C) = 0. \tag{4.143}$$

Note that

$$\begin{aligned}
&P^\pi(\tau(\pi) \le m \mid C) \\
&= P^\pi(\exists\ K \le n \le m \text{ and } \tilde{h} \text{ such that } M_{\tilde{h}}(n) \ge \log((K-1)L)) \\
&\overset{(a)}{\le} \sum_{\tilde{h}\in\mathcal{A}} \sum_{n=K}^{m} P^\pi(M_{\tilde{h}}(n) \ge \log((K-1)L) \mid C) \\
&= \sum_{\tilde{h}\in\mathcal{A}} \sum_{n=K}^{m} E^\pi[\mathbb{I}_{\{M_{\tilde{h}}(n)\ge\log((K-1)L)\}} \mid C] \\
&\overset{(b)}{\le} \sum_{\tilde{h}\in\mathcal{A}} \sum_{n=K}^{m} E^\pi\left[\frac{M_{\tilde{h}}(n)}{\log((K-1)L)} \cdot \mathbb{I}_{\{M_{\tilde{h}}(n)\ge\log((K-1)L)\}} \middle| C\right] \\
&= \frac{1}{\log((K-1)L)} \sum_{\tilde{h}\in\mathcal{A}} \sum_{n=K}^{m} \Big(E^\pi[M_{\tilde{h}}(n) \mid C] - E^\pi[M_{\tilde{h}}(n) \cdot \mathbb{I}_{\{M_{\tilde{h}}(n)<\log((K-1)L)\}} \mid C] \Big),
\end{aligned} \tag{4.144}$$

where (a) above is due to the union bound, (b) follows from the fact that on the set $\{M_{\tilde{h}}(n) \ge \log((K-1)L)\}$, we have $1 \le \frac{M_{\tilde{h}}(n)}{\log((K-1)L)}$, and in writing (4.144), we use the fact that

$$E^\pi[M_{\tilde{h}}(n) \mid C] = E^\pi[M_{\tilde{h}}(n) \cdot \mathbb{I}_{\{M_{\tilde{h}}(n)\ge\log((K-1)L)\}} \mid C] + E^\pi[M_{\tilde{h}}(n) \cdot \mathbb{I}_{\{M_{\tilde{h}}(n)<\log((K-1)L)\}} \mid C].$$

In order to handle the second term inside the brackets in (4.144), we note that for any $\tilde{h}$,

$$M_{\tilde{h}}(n) = \min_{h'\neq\tilde{h}} M_{\tilde{h}h'}(n). \tag{4.145}$$

Fixing an arbitrary $h' \neq \tilde{h}$ and using (4.37) noting that $T_3(n)$ and $T_4(n)$ are non-negative, we have

$$M_{\tilde{h}h'}(n)$$
$$\geq \log \mathbb{E}\Bigg[\exp\Bigg(\log\Bigg(\sum_{i\in\mathcal{S}} \phi(i)\, P^{\tilde{h}-1}(X_{\tilde{h}-1}^{\tilde{h}}|i)\Bigg) + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} N(n,\underline{d},\underline{i},\tilde{h},j)\ \log P^{d_{\tilde{h}}}(j|i_{\tilde{h}})\Bigg)\Bigg] \tag{4.146}$$
$$+ \log \mathbb{E}\Bigg[\exp\Bigg\{ \sum_{a\neq\tilde{h}} \log\Bigg(\sum_{i\in\mathcal{S}} \phi(i)\, Q^{a-1}(X_{a-1}^{a}|i)\Bigg)$$
$$+ \sum_{a\neq\tilde{h}} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} N(n,\underline{d},\underline{i},a,j)\ \log Q^{d_a}(j|i_a)\Bigg\}\Bigg]. \tag{4.147}$$

From an earlier exposition, we note that the expectation in (4.146) can be lower bounded by

$$\mathbb{E}\Bigg[\exp\Bigg(\log\Bigg(\sum_{i\in\mathcal{S}} \phi(i)\, P^{\tilde{h}-1}(X_{\tilde{h}-1}^{\tilde{h}}|i)\Bigg) + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} N(n,\underline{d},\underline{i},h,j)\ \log P^{d_{\tilde{h}}}(j|i_{\tilde{h}})\Bigg)$$
$$\cdot\, \mathbb{I}\Big(P\in\mathscr{P}(\bar{\varepsilon}^*)\Big)\Bigg]$$
$$\geq \exp\Bigg(\log(\phi_{\mathsf{min}}\cdot\bar{\varepsilon}^*) + (\log\bar{\varepsilon}^*)\cdot \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} N(n,\underline{d},\underline{i},h,j)\Bigg)\cdot \mathbb{P}(P\in\mathscr{P}(\bar{\varepsilon}^*))$$
$$\geq \exp\Big(\log\phi_{\mathsf{min}} + \log\bar{\varepsilon}^* + n\,\log\bar{\varepsilon}^*\Big)\cdot\mathbb{P}(P\in\mathscr{P}(\bar{\varepsilon}^*)). \tag{4.148}$$

In writing the second inequality above, we use the fact that $\sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} N(n,\underline{d},\underline{i},h,j) \leq n$. Taking logarithm on both sides of (4.148), it follows that (4.146) may be lower bounded by

$$\log\phi_{\mathsf{min}} + (n+1)\,\log\bar{\varepsilon}^* + \log\mathbb{P}(P\in\mathscr{P}(\bar{\varepsilon}^*)). \tag{4.149}$$

On similar lines, (4.147) may be lower bounded by

$$(K-1)\,\log\phi_{\mathsf{min}} + (n+K-1)\log\bar{\varepsilon}^* + \log\mathbb{P}(Q\in\mathscr{P}(\bar{\varepsilon}^*)), \tag{4.150}$$

Combining (4.149) and (4.150), and noting that $\mathbb{P}(Q\in\mathscr{P}(\bar{\varepsilon}^*)) = \mathbb{P}(P\in\mathscr{P}(\bar{\varepsilon}^*))$, we get

$$M_{\tilde{h}h'}(n) \geq K\,\log\phi_{\mathsf{min}} + (n+K)\,\log\bar{\varepsilon}^* + 2\,\log\mathbb{P}(P\in\mathscr{P}(\bar{\varepsilon}^*)). \tag{4.151}$$

Noting that the lower bound in (4.151) holds for all $h' \neq \tilde{h}$ and is therefore a lower bound for $M_{\tilde{h}}(n)$, (4.144) may be upper bounded by

$$\frac{1}{\log((K-1)L)} \sum_{\tilde{h} \in \mathcal{A}} \sum_{n=K}^{m} E^{\pi}[M_{\tilde{h}}(n) \mid C]$$
$$+ \frac{1}{\log((K-1)L)} \sum_{\tilde{h} \in \mathcal{A}} \sum_{n=K}^{m} \left((n+K) \log \frac{1}{\bar{\varepsilon}^*} + 2 \log \frac{1}{\mathbb{P}(P \in \mathscr{P}(\bar{\varepsilon}^*))} + K \log \phi_{\mathsf{min}} \right). \quad (4.152)$$

Recall from an earlier exposition that $\mathbb{P}(P \in \mathscr{P}(\bar{\varepsilon}^*)) > 0$. Thus, it follows that

$$\begin{aligned}
&\limsup_{L \to \infty} P^{\pi}(\tau(\pi) \leq m \mid C) \\
&\leq \limsup_{L \to \infty} \left\{ \frac{1}{\log((K-1)L)} \sum_{\tilde{h} \in \mathcal{A}} \sum_{n=K}^{m} E^{\pi}[M_{\tilde{h}}(n) \mid C] \right. \\
&\qquad \left. + \frac{1}{\log((K-1)L)} \sum_{\tilde{h} \in \mathcal{A}} \sum_{n=K}^{m} \left((n+K) \log \frac{1}{\bar{\varepsilon}^*} + 2 \log \frac{1}{\mathbb{P}(P \in \mathscr{P}(\bar{\varepsilon}^*))} + K \log \phi_{\mathsf{min}} \right) \right\} \\
&= \limsup_{L \to \infty} \frac{1}{\log((K-1)L)} \sum_{\tilde{h} \in \mathcal{A}} \sum_{n=K}^{m} E^{\pi}[M_{\tilde{h}}(n) \mid C], \quad (4.153)
\end{aligned}$$

where the last line above follows by noting that the limit supremum of the second term in (4.152) is equal to zero.

We now show that for each $n \in \{K, \ldots, m\}$, the expectation term $E^{\pi}[M_{\tilde{h}}(n) \mid C]$ is finite, from which the desired result follows. We carry out the analysis by considering the cases $\tilde{h} = h$ and $\tilde{h} \neq h$ separately. Here, h is the odd arm in the arms configuration $C = (h, P_1, P_2)$.

4.6.8.1 Case $\tilde{h} = h$

In this case, we have

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

$$= \log \int_{\Omega} \frac{\bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_h)}{\hat{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_{h'})} \cdot f(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid C)\, dP^{\pi}(\omega) \quad \forall\, h' \neq h$$

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

$$\leq \log \int_{\Omega} \Bigg\{ \frac{\bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_h)}{\hat{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_{h'})} \cdot \exp(-Z_{hh'}(n)(\omega)) \cdot \hat{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_{h'})\, dP^{\pi}(\omega) \Bigg\} \quad \forall\, h' \neq h$$

$$= \log \int_{\Omega} \bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_h) \cdot \exp(-Z_{hh'}(n)(\omega)) \cdot dP^{\pi}(\omega) \quad \forall\, h' \neq h, \tag{4.154}$$

where the last line follows from an application of the change of measure technique presented in [32, Lemma 18]. Also, $Z_{hh'}(n)$ in the above set of equations is given by

$$\begin{aligned} Z_{hh'}(n) &= \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \sum_{a=1}^{K} N(n, \underline{d}, \underline{i}, a, j) \log \frac{(P_h^a)^{d_a}(j|i_a)}{(P_{h'}^a)^{d_a}(j|i_a)} \\ &= \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h, j) \log \frac{P_1^{d_h}(j|i_h)}{P_2^{d_h}(j|i_h)} + N(n, \underline{d}, \underline{i}, h', j) \log \frac{P_2^{d_{h'}}(j|i_{h'})}{P_1^{d_{h'}}(j|i_{h'})}. \end{aligned} \tag{4.155}$$

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

$$\log \frac{1}{\bar{\varepsilon}^*} \geq \log \frac{P_1^d(j|i)}{P_2^d(j|i)} \geq \log \bar{\varepsilon}^* \quad \forall\, d \geq 1,\ i, j \in \mathcal{S}. \tag{4.156}$$

Therefore, it follows that almost surely,

$$\begin{aligned} Z_{hh'}(n) &\geq (\log \bar{\varepsilon}^*) \cdot \Bigg(\sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h, j) + N(n, \underline{d}, \underline{i}, h', j) \Bigg) \\ &\geq (\log \bar{\varepsilon}^*) \cdot \Bigg(\sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} \sum_{a=1}^{K} N(n, \underline{d}, \underline{i}, a, j) \Bigg) \qquad (4.157) \\ &\geq n \cdot (\log \bar{\varepsilon}^*). \qquad (4.158) \end{aligned}$$

Using (4.158) in (4.154), we get

$$E^{\pi}[M_{\tilde{h}}(n) \mid C] \leq \log \int_{\Omega} \bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_h) \ \cdot \ \exp(-n \cdot (\log \bar{\varepsilon}^*)) \ \cdot \ dP^{\pi}(\omega)$$
$$= n \ \log\left(\frac{1}{\bar{\varepsilon}^*}\right), \tag{4.159}$$

where the last line follows by noting that $\int_{\Omega} \bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_h) \ dP^{\pi}(\omega) = 1$ because $\bar{f}$ is an average likelihood.

4.6.8.2 Case $\tilde{h} \neq h$

In this case, we note that

$$\begin{aligned}
E^{\pi}[M_{\tilde{h}}(n) \mid C] &= E^{\pi}\left[\min_{h' \neq \tilde{h}} \log \frac{\bar{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_{\tilde{h}})}{\hat{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_{h'})} \middle| C\right] \\
&\leq E^{\pi}\left[\log \frac{\bar{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_{\tilde{h}})}{\hat{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_{h})} \middle| C\right] \\
&\overset{\text{Jensen's}}{\leq} \log E^{\pi}\left[\frac{\bar{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_{\tilde{h}})}{\hat{f}(B^n, A^n, \bar{X}^n \mid \mathcal{H}_{h})} \middle| C\right] \\
&= \log \int_{\Omega} \frac{\bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_{\tilde{h}})}{\hat{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_{h})} \ \cdot \ f(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid C) \ dP^{\pi}(\omega) \\
&\leq \log \int_{\Omega} \bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_{\tilde{h}}) \ dP^{\pi}(\omega) \\
&= 0,
\end{aligned} \tag{4.160}$$

where the last line follows by noting, as before, that $\int_{\Omega} \bar{f}(B^n(\omega), A^n(\omega), \bar{X}^n(\omega) \mid \mathcal{H}_{\tilde{h}}) \ dP^{\pi}(\omega) = 1$ because $\bar{f}$ is an average likelihood.

Using (4.159) and (4.160) in (4.153), it follows that for each $m \geq K$,

$$\limsup_{L \to \infty} P(\tau(\pi_2^{\star}(L, \delta)) \leq m \mid C) \leq \limsup_{L \to \infty} \frac{1}{\log((K-1)L)} \sum_{\tilde{h} \in \mathcal{A}} \sum_{n=K}^{m} n \ \log\left(\frac{1}{\bar{\varepsilon}^*}\right)$$
$$= 0. \tag{4.161}$$

This establishes the desired result.

4.6.9 Proof of Proposition 18

Here, we demonstrate that for each $\delta > 0$, the family $\{\tau(\pi_2^\star(L,\delta))/\log L : L > 1\}$ is uniformly integrable. Clearly, it suffices to show that under the policy $\pi = \pi_2^\star(L,\delta)$ and under the arms configuration $C = (h, P_1, P_2)$,

$$\limsup_{L\to\infty} E^\pi\left[\left(\frac{\tau(\pi)}{\log L}\right)^2 \middle| C\right] < \infty. \tag{4.162}$$

4.6.9.1 Showing that $P^\pi(M_h(n) < \log((K-1)L)|C)$ is $O(1/n^3)$

Before showing that (4.162) holds, we record the following important result.

Lemma 28. *Fix $C = (h, P_1, P_2)$, $L > 1$ and $\delta > 0$. There exists $0 < B < \infty$ independent of L such that for all sufficiently large values of n, under the policy $\pi = \pi_2^\star(L,\delta)$ and under the arms configuration C,*

$$P^\pi(M_h(n) < \log((K-1)L)|C) \le \frac{B}{n^3}. \tag{4.163}$$

Proof of Lemma 28. Note that

$$\begin{aligned}
&P^\pi(M_h(n) < \log((K-1)L)|C)\\
&= P^\pi\left(\min_{h'\neq h} M_{hh'}(n) < \log((K-1)L) \;\middle|\; C\right)\\
&\overset{(a)}{\le} \sum_{h'\neq h} P^\pi\left(M_{hh'}(n) < \log((K-1)L) \;\middle|\; C\right)\\
&= \sum_{h'\neq h} P^\pi\left(\frac{M_{hh'}(n)}{n} < \frac{\log((K-1)L)}{n} \;\middle|\; C\right)\\
&= \sum_{h'\neq h} P^\pi\left(\frac{T_1(n)}{n} + \frac{T_2(n)}{n} + \frac{T_3(n)}{n} + \frac{T_4(n)}{n} < \frac{\log((K-1)L}{n} \;\middle|\; C\right),
\end{aligned} \tag{4.164}$$

where (a) above follows from the union bound, and the last line follows from (4.37). In order to prove (4.163), it suffices to prove that each term inside the summation in (4.164) is $O(1/n^3)$.

By virtue of Assumption 2, we know that $P_1, P_2 \in \mathscr{P}(\bar{\varepsilon}^*)$ for some $\bar{\varepsilon}^* \in (0,1)$. From (4.38), it follows that

$$\begin{aligned}
&\frac{T_1(n)}{n} - \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},h,j)}{n} \log P_1^{d_h}(j|i_h)\\
&= \frac{1}{n}\log\mathbb{E}\left[\exp\left\{\log\left(\sum_{i\in\mathcal{S}} \phi(i)\, P^{h-1}(X_{h-1}^h|i)\right) + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} N(n,\underline{d},\underline{i},h,j)\, \log P^{d_h}(j|i_h)\right\}\right]
\end{aligned}$$

$$-\frac{1}{n}\log\exp\left\{\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}N(n,\underline{d},\underline{i},h,j)\ \log P_1^{d_h}(j|i_h)\right\}$$

$$=\frac{1}{n}\log\mathbb{E}\Bigg[\exp\left\{\log\left(\sum_{i\in\mathcal{S}}\phi(i)\,P^{h-1}(X_{h-1}^h|i)\right)+\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}N(n,\underline{d},\underline{i},h,j)\ \log\frac{P^{d_h}(j|i_h)}{P_1^{d_h}(j|i_h)}\right\}\Bigg]$$

$$\geq\frac{1}{n}\log\mathbb{E}\Bigg[\exp\left\{\log\left(\sum_{i\in\mathcal{S}}\phi(i)\,P^{h-1}(X_{h-1}^h|i)\right)+\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}N(n,\underline{d},\underline{i},h,j)\ \log\frac{P^{d_h}(j|i_h)}{P_1^{d_h}(j|i_h)}\right\}$$
$$\cdot\,\mathbb{I}\Big(P\in\mathscr{P}(\bar{\varepsilon}^*)\Big)\Bigg]$$

$$\overset{(a)}{\geq}\frac{1}{n}\log\left[\exp\left\{\log(\phi_{\mathsf{min}}\cdot\bar{\varepsilon}^*)+(\log\bar{\varepsilon}^*)\cdot\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}N(n,\underline{d},\underline{i},h,j)\right\}\cdot\mathbb{P}(P\in\mathscr{P}(\bar{\varepsilon}^*))\right]$$

$$\overset{(b)}{=}\frac{1}{n}\log\phi_{\mathsf{min}}+\frac{1}{n}\log\bar{\varepsilon}^*+\frac{N(n,h)}{n}(\log\bar{\varepsilon}^*)+\frac{1}{n}\log\mathbb{P}(P\in\mathscr{P}(\bar{\varepsilon}^*)),\tag{4.165}$$

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

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

and in (b) above,

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

Let $N(n,a)$ be defined similarly for all $a\neq h$. It follows from (4.165) that almost surely

$$\frac{T_1(n)}{n}\geq\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}\frac{N(n,\underline{d},\underline{i},h,j)}{n}\ \log P_1^{d_h}(j|i_h)+\frac{N(n,h)}{n}(\log\bar{\varepsilon}^*)+\frac{1}{n}\log\phi_{\mathsf{min}}$$
$$+\frac{1}{n}\log\bar{\varepsilon}^*+\frac{1}{n}\log\mathbb{P}(P\in\mathscr{P}(\bar{\varepsilon}^*)).\tag{4.166}$$

Similarly, it can be shown that almost surely,

$$\frac{T_2(n)}{n}\geq\sum_{a\neq h}\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}\frac{N(n,\underline{d},\underline{i},a,j)}{n}\ \log P_2^{d_a}(j|i_a)+(\log\bar{\varepsilon}^*)\left(\sum_{a\neq h}\frac{N(n,a)}{n}\right)$$
$$+\frac{K-1}{n}\log\phi_{\mathsf{min}}+\frac{K-1}{n}\log\bar{\varepsilon}^*+\frac{1}{n}\log\mathbb{P}(Q\in\mathscr{P}(\bar{\varepsilon}^*)).\tag{4.167}$$

Next, we note from (4.40) that for each $h'\neq h$, almost surely,

$$\frac{T_3(n)}{n}$$

$$
\begin{aligned}
&\geq \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},h',j)}{n} \log \frac{1}{(\hat{P}_{h',1}(n))^{d_{h'}}(j|i_{h'})} \\
&= \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},h',j)}{n} \log \frac{1}{P_2^{d_{h'}}(j|i_{h'})} + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},h',j)}{n} \log \frac{P_2^{d_{h'}}(j|i_{h'})}{(\hat{P}_{h',1}(n))^{d_{h'}}(j|i_{h'})} \\
&\geq \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},h',j)}{n} \log \frac{1}{P_2^{d_{h'}}(j|i_{h'})} + (\log\bar{\varepsilon}^*) \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},h',j)}{n} \\
&= \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},h',j)}{n} \log \frac{1}{P_2^{d_{h'}}(j|i_{h'})} + (\log\bar{\varepsilon}^*) \frac{N(n,h')}{n}, \qquad (4.168)
\end{aligned}
$$

where the inequality above follows by noting that because $P_2 \in \mathscr{P}(\bar{\varepsilon}^*)$, we have

$$
\forall\ d \geq 1 \text{ and } i,j \in \mathcal{S}, \quad P_2^d(j|i) \geq \bar{\varepsilon}^* \text{ whenever } P_2^d(j|i) > 0,
$$

as a consequence of which we have

$$
\frac{P_2^d(j|i)}{(\hat{P}_{h',1}(n))^d(j|i)} \geq \bar{\varepsilon}^* \quad \text{whenever } (\hat{P}_{h',1}(n))^d(j|i) > 0.
$$

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

$$
\begin{aligned}
&\frac{T_4(n)}{n} \\
&\geq \sum_{a\neq h'} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},a,j)}{n} \log \frac{1}{(\hat{P}_{h',2}(n))^{d_a}(j|i_a)} \\
&= \sum_{a\neq h'} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},a,j)}{n} \log \frac{1}{P_*^{d_a}(j|i_a)} \\
&\qquad\qquad + \sum_{a\neq h'} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},a,j)}{n} \log \frac{P_*^{d_a}(j|i_a)}{(\hat{P}_{h',2}(n))^{d_a}(j|i_a)} \\
&\geq \sum_{a\neq h'} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},a,j)}{n} \log \frac{1}{P_*^{d_a}(j|i_a)} + (\log\bar{\varepsilon}^*) \sum_{a\neq h'} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},a,j)}{n} \\
&= \sum_{a\neq h'} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},a,j)}{n} \log \frac{1}{P_*^{d_a}(j|i_a)} + (\log\bar{\varepsilon}^*) \Bigg(\sum_{a\neq h'} \frac{N(n,a)}{n} \Bigg), \qquad (4.169)
\end{aligned}
$$

Combining the lower bounds in (4.166)-(4.169), it follows that for each $h' \neq h$, almost surely,

$$\begin{aligned}
&\frac{M_{hh'}(n)}{n} \\
&\geq 2(\log \bar{\varepsilon}^*)\left(\sum_{a=1}^{K} \frac{N(n,a)}{n}\right) + \frac{2}{n}\ \log \mathbb{P}(P \in \mathscr{P}(\bar{\varepsilon}^*)) + \frac{K}{n}\ \log \phi_{\mathsf{min}} + \frac{K}{n}\ \log \bar{\varepsilon}^* \\
&+ \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},h,j)}{n}\ \log \frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)} + \sum_{a\neq h,h'} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},a,j)}{n}\ \log \frac{P_2^{d_a}(j|i_a)}{P_*^{d_a}(j|i_a)} \\
&= 2(\log \bar{\varepsilon}^*)\left(\sum_{a=1}^{K} \frac{N(n,a)}{n}\right) + \frac{2}{n}\ \log \mathbb{P}(P \in \mathscr{P}(\bar{\varepsilon}^*)) + \frac{K}{n}\ \log \phi_{\mathsf{min}} + \frac{K}{n}\ \log \bar{\varepsilon}^* \qquad (4.170) \\
&+ \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \left(\frac{N(n,\underline{d},\underline{i},h,j)}{n} - \frac{N(n,\underline{d},\underline{i},h)}{n}\ P_1^{d_h}(j|i_h)\right) \log \frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)} \qquad (4.171) \\
&+ \sum_{a\neq h,h'} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \left(\frac{N(n,\underline{d},\underline{i},a,j)}{n} - \frac{N(n,\underline{d},\underline{i},a)}{n}\ P_2^{d_a}(j|i_a)\right) \log \frac{P_2^{d_a}(j|i_a)}{P_*^{d_a}(j|i_a)} \qquad (4.172) \\
&+ \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},h)}{n}\ D(P_1^{d_h}(\cdot|i_h)\|P_*^{d_h}(\cdot|i_h)) \\
&\qquad + \sum_{a\neq h,h'} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \frac{N(n,\underline{d},\underline{i},a)}{n}\ D(P_2^{d_a}(\cdot|i_a)\|P_*^{d_a}(\cdot|i_a)). \qquad (4.173)
\end{aligned}$$

Using (4.170)-(4.173), the probability term $P^\pi\left(\frac{M_{hh'}(n)}{n} < \frac{\log((K-1)L)}{n} \,\middle|\, C\right)$ may be written as

$$P^\pi\left(\frac{M_{hh'}(n)}{n} < \frac{\log((K-1)L)}{n} \,\middle|\, C\right) \leq U_1(n) + U_2(n) + U_3(n) + U_4(n), \qquad (4.174)$$

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

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

$$U_1(n) = P^\pi\left(\frac{K}{n}\ \log \phi_{\mathsf{min}} + \frac{K}{n}\ \log \bar{\varepsilon}^* + \frac{2}{n}\ \log \mathbb{P}(P \in \mathscr{P}(\bar{\varepsilon}^*)) < -\varepsilon \,\middle|\, C\right). \qquad (4.175)$$

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

$$U_2(n) = P^\pi\left(\sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \left\{\left(\frac{N(n,\underline{d},\underline{i},h,j)}{n} - \frac{N(n,\underline{d},\underline{i},h)}{n}\ P_1^{d_h}(j|i_h)\right) \log \frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)}\right\} < -\varepsilon \,\middle|\, C\right). \qquad (4.176)$$

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

$$U_3(n) = P^{\pi}\Bigg(\sum_{a\neq h,h'} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{j\in\mathcal{S}} \Bigg\{ \Bigg(\frac{N(n,\underline{d},\underline{i},a,j)}{n} - \frac{N(n,\underline{d},\underline{i},a)}{n}\ P_2^{d_a}(j|i_a) \Bigg) \log \frac{P_2^{d_a}(j|i_a)}{P_*^{d_a}(j|i_a)} \Bigg\} < -\varepsilon \ \Bigg| \ C \Bigg). \quad (4.177)$$

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

$$\begin{aligned} U_4(n) = P^{\pi}\Bigg(& \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \frac{N(n,\underline{d},\underline{i},h)}{n}\ D(P_1^{d_h}(\cdot|i_h)\|P_*^{d_h}(\cdot|i_h)) \\ & + \sum_{a\neq h,h'} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \frac{N(n,\underline{d},\underline{i},a)}{n}\ D(P_2^{d_a}(\cdot|i_a)\|P_*^{d_a}(\cdot|i_a)) \\ & + 2(\log\bar{\varepsilon}^*)\Bigg(\sum_{a=1}^{K} \frac{N(n,a)}{n} \Bigg) + 3\varepsilon < \frac{\log((K-1)L)}{n} \ \Bigg| \ C \Bigg). \end{aligned} \quad (4.178)$$

The relations in (4.175)-(4.178) hold for all $\varepsilon > 0$. We shall shortly demonstrate how to choose $\varepsilon > 0$.

We now show that each of the terms $U_1(n)$-$U_4(n)$ is either 0 or $O(1/n^3)$. This will then establish (4.163).

4.6.9.2 Handling $U_1(n)$

Because $\mathbb{P}(P \in \mathscr{P}(\bar{\varepsilon}^*)) > 0$, it follows that inside the probability term in (4.175), the left hand side goes to zero as $n \to \infty$, whereas the right hand side is strictly negative. Thus, there exists $N_1 = N_1(\varepsilon)$ such that $U_1(n) = 0$ for all $n \geq N_1$.

4.6.9.3 Handling $U_2(n)$

Next, we note that

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

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

$$
\begin{aligned}
&E^\pi\Bigg[\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}\Big(N(n,\underline{d},\underline{i},h,j)-N(n,\underline{d},\underline{i},h)\ P_1^{d_h}(j|i_h)\Big)\ \log\frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)}\ \Big|\ B^{n-1},A^{n-1},\bar{X}^{n-1},\ C\Bigg]\\
&=E^\pi\Bigg[\sum_{t=K}^{n}\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}\Bigg\{\mathbb{I}_{\{\underline{d}(t)=\underline{d},\ \underline{i}(t)=\underline{i},\ A_t=h\}}\Big(\mathbb{I}_{\{\bar{X}_t=j\}}-P_1^{d_h}(j|i_h)\Big)\\
&\qquad\qquad\cdot\log\frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)}\Bigg\}\ \Big|\ B^{n-1},A^{n-1},\bar{X}^{n-1},\ C\Bigg]\\
&=\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}\Big(N(n-1,\underline{d},\underline{i},h,j)-N(n-1,\underline{d},\underline{i},h)\ P_1^{d_h}(j|i_h)\Big)\ \log\frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)}\\
&+\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}\mathbb{I}_{\{\underline{d}(n)=\underline{d},\ \underline{i}(n)=\underline{i}\}}\Big(P^\pi(A_n=h,\ \bar{X}_n=j|B^{n-1},A^{n-1},\bar{X}^{n-1},\ C)\\
&\qquad\qquad-P^\pi(A_n=h|B^{n-1},A^{n-1},\bar{X}^{n-1},\ C)\cdot P_1^{d_h}(j|i_h)\Big)\ \log\frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)}\\
&=\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}\Big(N(n-1,\underline{d},\underline{i},h,j)-N(n-1,\underline{d},\underline{i},h)\ P_1^{d_h}(j|i_h)\Big)\ \log\frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)},
\end{aligned}
\tag{4.179}
$$

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

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

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

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

$$\leq n\left(4\left(\log\frac{1}{\bar{\varepsilon}^*}\right)^2|\mathcal{S}|\right) \tag{4.180}$$

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

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

We then have

$$\begin{aligned}
&U_2(n)\\
&= P^\pi\left(\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}\left(N(n,\underline{d},\underline{i},h,j) - N(n,\underline{d},\underline{i},h)\ P_1^{d_h}(j|i_h)\right)\ \log\frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)} < -n\varepsilon \;\middle|\; C\right)\\
&= P^\pi\left(\left|\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}\left(N(n,\underline{d},\underline{i},h,j) - N(n,\underline{d},\underline{i},h)\ P_1^{d_h}(j|i_h)\right)\ \log\frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)}\right| > n\varepsilon \;\middle|\; C\right)\\
&\leq P^\pi\left(\sup_{K\leq t\leq n}\left|\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}\left(N(t,\underline{d},\underline{i},h,j) - N(t,\underline{d},\underline{i},h)\ P_1^{d_h}(j|i_h)\right)\ \log\frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)}\right| > n\varepsilon \;\middle|\; C\right)\\
&\overset{(a)}{\leq} \frac{1}{n^6\ \varepsilon^6}\ E^\pi\left[\left(\sup_{K\leq t\leq n}\left|\sum_{(\underline{d},\underline{i})\in\mathbb{S}}\sum_{j\in\mathcal{S}}\left\{\left(N(t,\underline{d},\underline{i},h,j) - N(t,\underline{d},\underline{i},h)\ P_1^{d_h}(j|i_h)\right)\right.\right.\right.\\
&\qquad\qquad\left.\left.\left.\cdot\log\frac{P_1^{d_h}(j|i_h)}{P_*^{d_h}(j|i_h)}\right|\right)^6\right\}\;\middle|\; C\right]\\
&\overset{(b)}{\leq} \frac{A}{n^6\ \varepsilon^6}\ E^\pi[|\langle M_n\rangle|^3|C]\\
&\overset{(c)}{\leq} \frac{A}{n^6\ \varepsilon^6}\ n^3\left(4\left(\log\frac{1}{\bar{\varepsilon}^*}\right)^2|\mathcal{S}|\right)^3\\
&= \frac{A'}{n^3},
\end{aligned} \tag{4.181}$$

where (a) above is due to Markov's inequality, (b) is due to Burkholder inequality [53, p. 414], and (c) follows from (4.180). The constant A in (b) above comes from Burkholder inequality. We have thus shown that $U_2(n)$ is $O(1/n^3)$.

4.6.9.4 Handling $U_3(n)$

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

4.6.9.5 Handling $U_4(n)$

We first note that

$$U_4(n) \leq P^\pi\Bigg(\sum_{(\underline{d},\underline{i})\in\mathbb{S}} \frac{N(n,\underline{d},\underline{i},h)}{n}\, D(P_1^{d_h}(\cdot|i_h)\|P_*^{d_h}(\cdot|i_h)) + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a\neq h,h'} \frac{N(n,\underline{d},\underline{i},a)}{n}\, D(P_2^{d_a}(\cdot|i_a)\|P_*^{d_a}(\cdot|i_a)) + 2(\log\bar{\varepsilon}^*)\Bigg(\sum_{a=1}^K \frac{N(n,a)}{n}\Bigg) + 3\varepsilon < \frac{\log((K-1)L)}{n} \;\Bigg|\; C\Bigg)$$
$$\leq P^\pi\Bigg(\frac{N(n,\underline{d},\underline{i},h)}{n}\, D(P_1^{d_h}(\cdot|i_h)\|P_*^{d_h}(\cdot|i_h)) + \sum_{a\neq h,h'} \frac{N(n,\underline{d},\underline{i},a)}{n}\, D(P_2^{d_a}(\cdot|i_a)\|P_*^{d_a}(\cdot|i_a)) + 2(\log\bar{\varepsilon}^*) + 3\varepsilon < \frac{\log((K-1)L)}{n} \;\Bigg|\; C\Bigg) \quad \text{for all } (\underline{d},\underline{i})\in\mathbb{S}. \tag{4.182}$$

In writing (4.182), we use the fact that $\sum_{a=1}^K \frac{N(n,a)}{n} \leq 1$. Going further, let us fix $\underline{d} = (K, K-1, \ldots, 1)$ and $\underline{i} = (i^*, i^*, \ldots, i^*)$ for some $i^* \in \mathbb{S}$. Let E_n be the event inside the probability term in (4.182). From (4.140), we know that almost surely,

$$\liminf_{n\to\infty} \frac{N(n,\underline{d},\underline{i},a)}{n} \geq \frac{\nu^{\lambda_h,P_1,P_2,\delta}(\underline{d},\underline{i},a)}{1+\delta}. \tag{4.183}$$

Exploiting this, and denoting the right hand side of (4.183) as $C(\underline{d},\underline{i},a)$, the probability term in (4.182) may be upper bounded as

$$P^\pi(E_n|C) \leq P^\pi\Bigg(E_n \cap \Bigg\{\frac{N(n,\underline{d},\underline{i},a)}{n} \geq C(\underline{d},\underline{i},a)(1-\epsilon')\ \forall a\in\mathcal{A}\Bigg\} \;\Bigg|\; C\Bigg) \tag{4.184}$$
$$+ P^\pi\Bigg(E_n \cap \Bigg\{\exists\, a\in\mathcal{A} : \frac{N(n,\underline{d},\underline{i},a)}{n} < C(\underline{d},\underline{i},a)(1-\epsilon')\Bigg\} \;\Bigg|\; C\Bigg) \tag{4.185}$$

for all $\epsilon' \in (0,1)$. We shall soon specify how to choose ϵ'. The probability term in (4.184) may then be upper bounded by

$$P^\pi\Bigg((1-\epsilon')\Bigg(C(\underline{d},\underline{i},h)\; D(P_1^{d_h}(\cdot|i_h)\|P_*^{d_h}(\cdot|i_h)) + \sum_{a\neq h,h'} C(\underline{d},\underline{i},a)\; D(P_2^{d_a}(\cdot|i_a)\|P_*^{d_a}(\cdot|i_a))\Bigg) + 2(\log\bar{\varepsilon}^*) + 3\varepsilon < \frac{\log((K-1)L)}{n} \;\Bigg|\; C\Bigg). \tag{4.186}$$

Recall that $P_* \in \mathscr{P}(\bar{\varepsilon}^*)$, $P_* \neq P_1, P_2$ as indicated prior to (4.169). Let $\varepsilon > 0$ be chosen such that

$$3\varepsilon > 2\,\log\frac{1}{\bar{\varepsilon}^*} - (1-\epsilon')\left(C(\underline{d},\underline{i},h)\;D(P_1^{d_h}(\cdot|i_h)\|P_*^{d_h}(\cdot|i_h)) + \sum_{a\neq h,h'} C(\underline{d},\underline{i},a)\;D(P_2^{d_a}(\cdot|i_a)\|P_*^{d_a}(\cdot|i_a))\right). \tag{4.187}$$

Such a choice of ε is possible because the right hand side of (4.187) is strictly positive by virtue of the fact that

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

for all $a \neq h, h'$, and therefore

$$\begin{aligned}
&(1-\epsilon')\left(C(\underline{d},\underline{i},h)\;D(P_1^{d_h}(\cdot|i_h)\|P_*^{d_h}(\cdot|i_h)) + \sum_{a\neq h,h'} C(\underline{d},\underline{i},a)\;D(P_2^{d_a}(\cdot|i_a)\|P_*^{d_a}(\cdot|i_a))\right)\\
&\leq \left(\log\frac{1}{\bar{\varepsilon}^*}\right)(1-\epsilon')\sum_{a\neq h'} C(\underline{d},\underline{i},a)\\
&\leq \log\frac{1}{\bar{\varepsilon}^*}.
\end{aligned} \tag{4.188}$$

For instance, it suffices to choose

$$\begin{aligned}
\varepsilon = \frac{2}{3}\Bigg(&2\,\log\frac{1}{\bar{\varepsilon}^*}\\
&-(1-\epsilon')\left(C(\underline{d},\underline{i},h)\;D(P_1^{d_h}(\cdot|i_h)\|P_*^{d_h}(\cdot|i_h)) + \sum_{a\neq h,h'} C(\underline{d},\underline{i},a)\;D(P_2^{d_a}(\cdot|i_a)\|P_*^{d_a}(\cdot|i_a))\right)\Bigg).
\end{aligned} \tag{4.189}$$

With ε as chosen above, it follows that inside the probability term in (4.186), the left hand side is strictly positive whereas the right hand side goes to 0 as $n \to \infty$. Therefore, for all sufficiently large values of n, the probability term in (4.186) is equal to zero.

We now turn attention to the probability term in (4.185). Using the union bound, this term may be upper bounded by

$$\sum_{a=1}^{K} P^{\pi}\left(\frac{N(n,\underline{d},\underline{i},a)}{n} < C(\underline{d},\underline{i},a)(1-\epsilon') \;\middle|\; C\right) = \sum_{a=1}^{K} P^{\pi}\left(N(n,\underline{d},\underline{i},a) < n\,C(\underline{d},\underline{i},a)(1-\epsilon') \;\middle|\; C\right). \tag{4.190}$$

In order to complete the proof, it suffices to show that each term inside the summation in

(4.190) is $O(1/n^3)$ for a suitable choice of $\epsilon' > 0$.

Fix an arbitrary $a \in \mathcal{A}$. From the exposition in Appendix 4.6.4, we know that the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ has the property that for all $T_0 \geq K$, the relation

$$P^\pi(\underline{d}(T_0+M+K+1) = \underline{d}, \underline{i}(T_0+M+K+1) = \underline{i} \mid \underline{d}(T_0) = \underline{d}'', \underline{i}(T_0) = \underline{i}'', C) \geq \rho \quad \forall\, (\underline{d}'', \underline{i}'') \in \mathbb{S}, \tag{4.191}$$

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

$$P_1^M(j|i) > 0, \quad P_2^M(j|i) > 0 \quad \forall\, i, j \in \mathcal{S}. \tag{4.192}$$

The existence of such an integer M is guaranteed by [41, Proposition 1.7]. From (4.191), it follows that

$$P^\pi(\underline{d}(t) = \underline{d},\ \underline{i}(t) = \underline{i}|C) \geq \left(\frac{\eta}{K}\right)^{2K} (\bar{\varepsilon}^*)^K \quad \forall\, t \geq M + 3K + 1. \tag{4.193}$$

Additionally, we know that under the policy $\pi = \pi_2^\star(L, \delta)$,

$$P^\pi(A_t = a|\underline{d}(t) = \underline{d},\ \underline{i}(t) = \underline{i},\ B^{t-1},\ A^{t-1},\ \bar{X}^{t-1},\ C) \geq \frac{\eta}{K} \quad \forall\, t \geq K. \tag{4.194}$$

As a consequence of (4.194), it follows that $P^\pi(A_t = a|\underline{d}(t) = \underline{d},\ \underline{i}(t) = \underline{i}.\ C) \geq \frac{\eta}{K}$ for all $t \geq K$. Combining this with (4.193), we see that for all $t \geq M + 3K + 1$,

$$P^\pi(\underline{d}(t) = \underline{d},\ \underline{i}(t) = \underline{i},\ A_t = a|C) \geq \left(\frac{\eta}{K}\right)^{2K+1} (\bar{\varepsilon}^*)^K. \tag{4.195}$$

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

$$\begin{aligned} N(n, \underline{d}, \underline{i}, a) &= \sum_{t=K}^{M+2K} \mathbb{I}_{\{\underline{d}(t)=\underline{d},\ \underline{i}(t)=\underline{i},\ A_t=a\}} + \sum_{t=M+2K+1}^{n} \mathbb{I}_{\{\underline{d}(t)=\underline{d},\ \underline{i}(t)=\underline{i},\ A_t=a\}} \\ &\geq \sum_{t=M+2K+1}^{n} \mathbb{I}_{\{\underline{d}(t)=\underline{d},\ \underline{i}(t)=\underline{i},\ A_t=a\}} \end{aligned} \tag{4.196}$$

almost surely. Denoting the right hand side of (4.196) as $N'(n, \underline{d}, \underline{i}, a)$, it follows from (4.195) that

$$E^\pi[N'(n, \underline{d}, \underline{i}, a)|C] \geq (n - M - 2K) \left(\frac{\eta}{K}\right)^{2K+1} (\bar{\varepsilon}^*)^K \quad \forall\, n \geq M + 3K + 1. \tag{4.197}$$

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

$$
\begin{aligned}
&P^{\pi}\Bigg(N(n,\underline{d},\underline{i},a) < n\,C(\underline{d},\underline{i},a)(1-\epsilon') \;\Bigg|\; C\Bigg) \\
&\leq P^{\pi}\Bigg(N'(n,\underline{d},\underline{i},a) < n\,C(\underline{d},\underline{i},a)(1-\epsilon') \;\Bigg|\; C\Bigg) \\
&= P^{\pi}\Bigg(N'(n,\underline{d},\underline{i},a) - E^{\pi}[N'(n,\underline{d},\underline{i},a)|C] < n\,C(\underline{d},\underline{i},a)(1-\epsilon') - E^{\pi}[N'(n,\underline{d},\underline{i},a)|C] \;\Bigg|\; C\Bigg) \\
&\leq P^{\pi}\Bigg(N'(n,\underline{d},\underline{i},a) - E^{\pi}[N'(n,\underline{d},\underline{i},a)|C] < n\,C(\underline{d},\underline{i},a)(1-\epsilon') \\
&\qquad\qquad - (n-M-2K)\left(\frac{\eta}{K}\right)^{2K+1}(\bar{\varepsilon}^*)^K \;\Bigg|\; C\Bigg) \\
&\leq P^{\pi}\Bigg(N'(n,\underline{d},\underline{i},a) - E^{\pi}[N'(n,\underline{d},\underline{i},a)|C] < n\Bigg(C(\underline{d},\underline{i},a)(1-\epsilon') \\
&\qquad\qquad - \left(\frac{n-M-2K}{n}\right)\left(\frac{\eta}{K}\right)^{2K+1}(\bar{\varepsilon}^*)^K\Bigg) \;\Bigg|\; C\Bigg) \\
&\leq P^{\pi}\Bigg(N'(n,\underline{d},\underline{i},a) - E^{\pi}[N'(n,\underline{d},\underline{i},a)|C] < n\Bigg(\frac{1-\epsilon'}{1+\delta} \\
&\qquad\qquad - \left(\frac{n-M-2K}{n}\right)\left(\frac{\eta}{K}\right)^{2K+1}(\bar{\varepsilon}^*)^K\Bigg) \;\Bigg|\; C\Bigg),
\end{aligned}
\tag{4.198}
$$

where the last line follows by noting that

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

We now show that for a suitable choice of ϵ', the right hand side of the probability term in (4.198) can be made negative. Because $(n-M-2K)/n \longrightarrow 1$ as $n \to \infty$, it follows that there exists $N_2 = N_2(\delta)$ such that for all $n \geq N_2$,

$$
\frac{n-M-2K}{n} > \frac{1}{1+\delta},
$$

as a consequence of which

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

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

$$1-\epsilon' < \left(\frac{\eta}{K}\right)^{2K+1}(\bar{\varepsilon}^*)^K. \tag{4.199}$$

Such a choice of ϵ' is possible since

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

For instance, it suffices to choose

$$\epsilon' = 1 - \frac{1}{3}\cdot\left(\frac{\eta}{K}\right)^{2K+1}(\bar{\varepsilon}^*)^K. \tag{4.200}$$

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

$$\begin{aligned} &P^\pi\Big(N(n,\underline{d},\underline{i},a) < n\,C(\underline{d},\underline{i},a)(1-\epsilon') \;\Big|\; C\Big) \\ &\leq P^\pi\Big(N'(n,\underline{d},\underline{i},a) - E^\pi[N'(n,\underline{d},\underline{i},a)|C] < -n\Delta \;\Big|\; C\Big), \end{aligned} \tag{4.201}$$

where Δ is given by

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

We now demonstrate that $N'(n,\underline{d},\underline{i},a)$ is sub-gaussian. Subsequently, we use sub-gaussian concentration bounds to show that the right hand side of (4.201) is bounded above exponentially. Recall that a random variable Z is said to be sub-gaussian with variance factor v [54, Section 2.3] if

$$\log E\Big[\exp\Big(\lambda(Z-E[Z])\Big)\Big] \leq \frac{\lambda^2 v}{2} \quad \forall\, \lambda \in \mathbb{R}. \tag{4.202}$$

It is well known that if Z is Bernoulli distributed, then Z is sub-gaussian with variance factor $v = 1/4$. The below lemma demonstrates that given sub-gaussian random variables X and Y

(not necessarily independent), their sum is also sub-gaussian.

Lemma 29. *Suppose X is sub-gaussian with variance factor v_1 and Y (not necessarily independent of X) is sub-gaussian with variance factor v_2. Then, $X+Y$ is sub-gaussian with variance factor v_1+v_2.*

Proof of Lemma 29. Using Holder's inequality, for all $p > 1$, we have

$$\begin{aligned}
&E\left[\exp\left(\lambda(X+Y-(E[X]+E[Y]))\right)\right] \\
&\leq \left(E\left[\exp\left(p\,\lambda(X-E[X])\right)\right]\right)^{\frac{1}{p}} \cdot \left(E\left[\exp\left((1-p)\,\lambda(Y-E[Y])\right)\right]\right)^{\frac{1}{1-p}} \\
&\leq \left(\exp\left(\frac{\lambda^2\,p^2\,v_1}{2}\right)\right)^{\frac{1}{p}} \cdot \left(\exp\left(\frac{\lambda^2\,(1-p)^2\,v_2}{2}\right)\right)^{\frac{1}{1-p}} \\
&= \exp\left(\frac{\lambda^2(v_2+p(v_1-v_2)}{2}\right).
\end{aligned} \tag{4.203}$$

In particular, for $p = 1 + v_2/v_1$, we have

$$\begin{aligned}
E\left[\exp\left(\lambda(X+Y-(E[X]+E[Y]))\right)\right] &\leq \exp\left(\frac{\lambda^2}{2}\left(v_2+\frac{v_1^2-v_2^2}{v_1}\right)\right) \\
&\leq \exp\left(\frac{\lambda^2(v_1+v_2)}{2}\right).
\end{aligned} \tag{4.204}$$

This establishes the desired result. □

Using Lemma 29 and the sub-gaussian property of a Bernoulli random variable mentioned earlier, it follows that $N'(n,\underline{d},\underline{i},a)$ is sub-gaussian with variance factor

$$v^* = \frac{n-M-2K}{4}.$$

Using concentration bounds for sub-gaussian random variables [54, p. 25], we get

$$P^\pi\left(N'(n,\underline{d},\underline{i},a) - E^\pi[N'(n,\underline{d},\underline{i},a)|C] < -n\,\Delta \;\middle|\; C\right) \leq \exp\left(-\frac{n^2\,\Delta^2}{2\,v^*}\right) \leq \exp(-2\,n\,\Delta^2). \tag{4.205}$$

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

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

Therefore, it follows that

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

This completes the handling of $U_4(n)$.

Combining (4.181) and the above result, and choosing A' in (4.181) large if needed, we arrive at (4.163). This completes the proof of Lemma 28. □

4.6.9.6 Completing the Proof of Proposition 18

We first use the result established in Lemma 28 to show (4.162), and later use (4.162) to complete the proof of Proposition 18.

Proof of (4.162). Let

$$u(L) := \frac{\log((K-1)L)}{(\log L)\cdot\frac{3}{2}\cdot\varepsilon}, \tag{4.206}$$

where ε is as given in (4.189). Let $\psi(L) = \max\{u(L), N_1, N_2, N_3, M+3K+1\}$, where the constants N_1, N_2, N_3 are as determined in the proof of Lemma 28. Then, we have

$$\begin{aligned}
&E^{\pi}\left[\left(\frac{\tau(\pi)}{\log L}\right)^2 \;\middle|\; C\right] \\
&= \int_0^{\infty} P^{\pi}\left(\left(\frac{\tau(\pi)}{\log L}\right)^2 > x \;\middle|\; C\right) dx \\
&= \int_0^{\infty} P^{\pi}\left(\frac{\tau(\pi)}{\log L} > \sqrt{x} \;\middle|\; C\right) dx \\
&= \int_0^{\infty} P^{\pi}\left(\tau(\pi) > \lfloor\sqrt{x}\,\log L\rfloor \;\middle|\; C\right) dx \\
&= \int_0^{\psi(L)} P^{\pi}\left(\tau(\pi) > \lfloor\sqrt{x}\,\log L\rfloor \;\middle|\; C\right) dx + \int_{\psi(L)}^{\infty} P^{\pi}\left(\tau(\pi) > \lfloor\sqrt{x}\,\log L\rfloor \;\middle|\; C\right) dx \\
&= \frac{1}{(\log L)^2} \int_0^{\log L\sqrt{\psi(L)}} P^{\pi}\left(\tau(\pi) > \lfloor u\rfloor \;\middle|\; C\right) 2\,u\,du \;+ \int_{\psi(L)}^{\infty} P^{\pi}\left(\tau(\pi) > \lfloor\sqrt{x}\,\log L\rfloor \;\middle|\; C\right) dx
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{(\log L)^2} \int_0^{\log L \sqrt{\psi(L)}} 2\,u\,du \;+ \sum_{n \geq \lfloor \log L \sqrt{\psi(L)} \rfloor} \left[\left(\frac{n+1}{\log L} \right)^2 - \left(\frac{n}{\log L} \right)^2 \right] P^\pi \left(\tau(\pi) > n \;\middle|\; C \right) \\
&\overset{(a)}{\leq} \psi(L) \;+ \sum_{n \geq \lfloor \log L \sqrt{\psi(L)} \rfloor} \frac{2n+1}{(\log L)^2} \, P^\pi \left(M_h(n) < \log((K-1)L) \;\middle|\; C \right) \\
&\overset{(b)}{\leq} \psi(L) + \frac{1}{(\log L)^2} \sum_{n \geq \lfloor \log L \sqrt{\psi(L)} \rfloor} (2n+1) \, \frac{B}{n^3} \\
&\leq \psi(L) + \frac{1}{(\log L)^2} \sum_{n \geq 1} (2n+1) \, \frac{B}{n^3}, \qquad (4.207)
\end{aligned}
$$

where (a) above follows by upper bounding noting that $\{\tau(\pi) > n\} \subset \{M_h(n) < \log((K-1)L)\}$, and (b) is due to Lemma 28. Noting that the summation (4.207) is finite, we get

$$
\begin{aligned}
\limsup_{L \to \infty} E^\pi \left[\left(\frac{\tau(\pi)}{\log L} \right)^2 \;\middle|\; C \right] &\leq \limsup_{L \to \infty} \psi(L) \\
&< \infty. \qquad (4.208)
\end{aligned}
$$

This establishes (4.162). □

Combining the almost sure upper bound in (4.54) and the uniform integrability result of (4.162), we arrive at the upper bound for the expected stopping time of the policy $\pi_2^\star(L, \delta)$ in (4.56). This completes the proof of Proposition 18.

4.7 Summary

1. The main results of this chapter are the following. (a) We gave an asymptotic lower bound on the growth rate of the expected time required to find the odd arm, subject to an upper bound on the error probability (PAC setting). The asymptotics is as the error probability vanishes. The growth rate of the expected time to find the odd arm, see (4.11), is $1/R^*(h, P_1, P_2)$, where $R^*(h, P_1, P_2)$ is defined in (4.12). (b) We gave a policy (called $\pi^*(L, \delta)$) based on the principle of certainty equivalence that uses maximum likelihood (ML) estimates and achieves the lower bound asymptotically.

2. The achievability analysis relied crucially on showing that (a) the ML estimates of the TPMs converge to their true values asymptotically, and (b) given $\delta > 0$ and an arms configuration C, the policy $\pi_2^\star(L, \delta)$ eventually samples the arms according to the δ-optimal solution for C as given by the lower bound. These conditions were established under

Assumption 1, which is that there exists a continuous selection of δ-optimal solutions, and Assumption 2, which is that there exists $\bar{\varepsilon}^* \in (0,1)$ such that for every arms configuration $C = (h, P, Q)$, the TPMs $P, Q \in \mathscr{P}(\bar{\varepsilon}^*)$. Both the assumptions were crucial in establishing that the ML estimates of the TPMs converge to their true values. In [6, 36], an analogue of the continuous selection property was established for the maximisers instead of δ-optimal solutions. However, for more general settings such as the setting of this chapter or the setting considered in [35], a continuous selection assumption seems inevitable and difficult to do away with.

3. The expression for $R^*(h, P_1, P_2)$ contains relative entropies of the corresponding rows of the TPMs P_1 and P_2. The closer the rows of P_1 and P_2 are to each other, the smaller the value of $R^*(h, P_1, P_2)$ and therefore the larger the growth rate. The computability of $R^*(h, P_1, P_2)$ is an issue because it contains an outer supremum over all SRS policies which is difficult to evaluate. An algorithm such as Q-learning may need to be employed to compute $R^*(h, P_1, P_2)$.

4. The policy $\pi_2^\star(L, \delta)$ is based a modified GLR test statistic that consists of an average average likelihood evaluated with respect to an artificial prior (the uniform distribution on the probability simplex $\mathscr{P}(\mathcal{S})$) in the numerator, and the maximum likelihood in the denominator. The computability of the maximum likelihood is an issue because closed-form expressions for the ML estimates of the TPMs are not available. At best, the ML estimates can be evaluated numerically.

5. Repeatedly sampling each arm and using the consecutive observations from the arms to estimate the TPMs may lead to simple closed-form expressions for the TPM estimates. Because the Markov process of each arm is ergodic, by virtue of the ergodic theorem, these estimates converge to their true values. However, in this case, it is not clear if the arms will be sampled eventually according to the δ-optimal solution for the underlying arms configuration (which, to recall, is crucial for achieving the asymptotic lower bound in (4.11)). Also, it is not clear if the desired error probability can be met. Policies that sample the arms repeatedly are known to perform well for the problem of minimising regret [28]. However, it is not clear if such policies perform well for optimal stopping problems such as that studied in this chapter.

6. In writing (4.37), the expression for the modified GLR test statistic $M_{hh'}(n)$, we assume that for each $a \in \mathcal{A}$, the initial state X_0^a of arm a is sampled according to the distribution ϕ that puts a strictly positive mass on each element of $\mathcal{S}$. However, this may not actually

be the case. For instance, even before the decision entity begins the sampling of the arms at time $t = 0$, if the Markov process of each arm has evolved for a sufficiently long duration of time and reached stationarity, then under the configuration $C = (h, P_1, P_2)$, we have $X_0^a \sim \mu_C^a$ where μ_C^a is the stationary distribution of arm a under the configuration C. Because the TPMs P_1 and P_2 are unknown, the distributions μ_C^a too are unknown. In this case, there may be a mismatch between the average and the maximum likelihoods computed using ϕ (as in our policy $\pi_2^\star(L, \delta)$) and the actual values of these likelihoods (based on μ_C^a). Let $\bar{M}_{hh'}(n)$ denote the denote the value of (4.37) with ϕ replaced by μ_C^a for each arm a. Because the Markov process of each arm is irreducible and positive recurrent, we have $\mu_{\mathsf{min}}(C) \coloneqq \min\{\mu_C^a(i) : i \in \mathcal{S},\ a \in \mathcal{A}\} > 0$. Letting $\phi_{\mathsf{min}} = \min\{\phi(i) : i \in \mathcal{S}\}$, we see that

$$\phi_{\mathsf{min}} \leq \frac{\phi(i)}{\mu_h^a(i)} \leq \frac{1}{\mu_{\mathsf{min}}(C)} \quad \text{for all } i \in \mathcal{S}.$$

In this case, it can be shown that $\phi_{\mathsf{min}} \leq |M_{hh'}(n) - \bar{M}_{hh'}(n)| \leq 1/\mu_{\mathsf{min}}(C)$, as a result of which we have

$$\lim_{n\to\infty} \left| \frac{M_{hh'}(n)}{n} - \frac{\bar{M}_{hh'}(n)}{n} \right| = 0.$$

That is, the asymptotic drift of $M_{hh'}(n)$ is identical to that of $\bar{M}_{hh'}(n)$. Therefore, our assumption about $X_0^a \sim \phi$ does not affect the asymptotic analysis in any way.

Chapter 5

Conclusions and Future Directions

The central problem we studied in this thesis was that of identifying an anomalous arm (odd arm) in a multi-armed bandit as quickly as possible subject to an upper bound on the probability of error, when each yields Markov observations. For this problem, we characterised the asymptotic growth rate of the expected time required to find the odd arm, where the asymptotics is as the error probability vanishes. Our analysis was broadly along the following lines. We first derived a problem-instance specific asymptotic lower bound on the growth rate of the expected time required to find the odd arm. The constant in the lower bound captured the hardness of the problem, and suggested a natural arm selection policy as per whose frequencies the arms must be selected in the long run in order to meet the lower bound. Accordingly, we devised policies that select the arms at the correct frequencies in the long run, consequently leading to upper bounds that matched with the lower bounds. A key, common feature in all of our policies was the idea of forced exploration – selecting each arm with a strictly positive probability at any given time. We carried out the analysis separately for the settings of rested arms and restless arms when the TPMs of the arms are known beforehand or unknown.

A key ingredient in our analysis of the setting of rested arms was the identification of the fact that for the Markov process of each arm, the long term fraction of entries into a state is equal to the long term fraction of exits from the state (global balance). Both these quantities are in turn equal to the probability of observing the state under the arm's stationary distribution. As for the setting of restless arms, the focal point of our analysis was the recognition of the fact that for each pair of arm delays and last observed states, say $(\underline{d}, \underline{i})$, the long term fraction of entries into $(\underline{d}, \underline{i})$ equals the long term fraction of exits from $(\underline{d}, \underline{i})$. The recognition of such 'invariant' quantities lies at the heart of our analysis. We showed that the arm delays and the last observed states form a controlled Markov process. The ergodicity of this process played a crucial role in the analysis. Our 'lift' approach of considering the delays and the last observed

states of all the arms jointly offered a global perspective in contrast to the local perspective of dealing with the delay and the last observed state of each arm separately as suggested by the prior works.

When the TPMs of the arms are unknown, we used the principle of maximum likelihood estimation to estimate the TPMs. The key challenge here was to show that these estimates converge to their true values in the long-term, i.e., the system is identifiable. In the setting of rested arms, we proved system identifiability by a simple application of the ergodic theorem. This was possible because closed-form expressions for the ML estimates were available. In the setting of restless arms, closed-form expressions for the ML estimates were not available, and therefore the convergence of ML estimates to their true values could not be asserted directly. We proved system identifiability under two sufficient conditions – a regularity condition on the unknown TPMs and the existence of a continuous selection of near-optimal solutions to the lower bound.

Computation of the Lower Bound

In the setting of rested arms, we showed that the lower bound can be computed in closed form. We also showed that for any arms configuration, the optimal (unconditional) probability distribution on the arms could be identified in closed-form.

In the setting of restless arms, we argued that computing the lower bound in closed form is challenging because of the presence of the countably infinite-valued arm *delays*. In this case, one must deal with conditional probability distributions on the arms (conditioned on the arm *delays* and the *last observed states*) which are more complicated to deal with than their unconditional counterparts. An algorithm such as Q-learning for restless arms [37] may be needed to compute the optimal conditional distribution.

Open questions and future directions We conclude the thesis with some open questions and some possible future directions to explore.

- **Switching costs**: Our problem setup does not include any penalty for switching between the arms [55, 56]. This means that the learning agent is free to switch between the arms any number of times and at any rate, until stoppage. However, in practice, switching between the arms may incur a cost. For example, when a secondary user in a cognitive radio network attempts to find a free channel for transmission, switching between the various channels in the network may be expensive. In such cases, it is important to

incorporate switching costs into the problem formulation. A possible formulation is to minimise the total cost, equal to the expected time to find the odd arm plus the overall switching cost, subject to an upper bound on the error probability. A detailed analysis of this problem for the case of independent observations from the arms is available in [8]. Extensions to the case of Markov observations may be interesting to study.

- **The case of no trembling hand**: The focal point in our analysis of the setting of restless arms (Chapters 3 and 4) was the ergodicity of the controlled Markov process of arm delays and last observed states (Lemma 9 of Chapter 3). Recall that this ergodicity property was established under the trembling hand model (i.e., the trembling hand parameter $\eta > 0$). The trembling hand model may be viewed as a regularisation that gives ergodicity for free. It would be interesting to study the case when $\eta = 0$. Our analysis in Section 3.7 shows that merely setting $\eta = 0$ in the expressions for the lower and the upper bounds derived for the case $\eta > 0$ may not yield the corresponding bounds for the case $\eta = 0$.

  For fixed $\eta, \delta > 0$ and arms configuration C, if $\lambda_{\eta,\delta,C}$ is a δ-optimal solution to the lower bound when the trembling hand parameter is η and the true arms configuration is C, and $\mu^{\lambda_{\eta,\delta,C}}$ is the stationary distribution associated with the SRS policy $\pi^{\lambda_{\eta,\delta,C}}$, then establishing the *tightness* of the family $\{\mu^{\lambda_{\eta,\delta,C}} : \eta > 0\}$ of stationary distributions may be crucial in order to show that the lower and the upper bounds for the case $\eta = 0$ match [57, Theorem 3.1, pp. 61]. It may be interesting to explore this further.

- **Second order asymptotics**: In Chapters 2, 3 and 4, given an arms configuration C, we characterised the value of

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

  Denoting the value of the above quantity as $\alpha = \alpha(C)$, our exposition shows that for all $\pi \in \Pi(\epsilon)$, we have $E^{\pi}[\tau(\pi)|C] \approx \alpha \cdot \log(1/\epsilon)$ for sufficiently small values of ϵ. Alternatively, the first order term in the asymptotic expansion for $E^{\pi}[\tau(\pi)|C]$ is given by α. It would be interesting to characterise the second order term of this expansion. By this, we mean the following: for what choice of function g does the expansion

$$E^{\pi}[\tau(\pi)|C] \approx \alpha \cdot \log(1/\epsilon) + \beta \cdot g(\epsilon) + o(g(\epsilon))$$

  hold for all $\pi \in \Pi(\epsilon)$? In the above expansion, $\beta = \beta(C)$ is a strictly positive, arms configuration-dependent constant that captures the second order growth rate of the expected time to find the odd arm. It is instructive to note that in our derivation of the

lower bounds, we showed along the lines of [1] that for each $\epsilon > 0$,

$$\inf_{\pi \in \Pi(\epsilon)} E^{\pi}[\tau(\pi)|C] \geq \alpha \cdot d(\epsilon, 1-\epsilon),$$

where $d(\epsilon, 1-\epsilon) = \epsilon \log \frac{\epsilon}{1-\epsilon} + (1-\epsilon) \log \frac{1-\epsilon}{\epsilon}$. Deriving the second and higher order terms in the above lower bound could serve as a starting point towards inferring the function g mentioned above.

- **General sequential hypothesis testing with Markov observations**: While the focus of this thesis has been on the analysis of the problem of identifying an anomalous arm in a multi-armed bandit with Markov observations from each arm, it may interesting to study more general sequential hypothesis testing problems such as best[1] arm identification, multi-bandit best arm identification [58] where the arms are categorised into one or more groups with possible overlaps between the various groups and the goal is to identify the best arm in each group, L-anomalous arms identification where the goal is to identify a set of $L > 1$ anomalous arms in a multi-armed bandit, etc, all in the context of Markov observations from the arms. An analysis of best arm identification for the setting of rested arms appears in the recent work of Moulos [20]. Extensions to the setting of restless arms could be interesting to explore. An analysis of general sequential hypothesis testing problems for the case of independent observations from the arms appears in the recent works [34, 35]. Extensions to the case of Markov observations from the arms could be interesting to study.

- **Sophisticated visual search models**: As noted at the beginning of Chapter 4 and in [43], it is likely that in visual search experiments, the human subject scans multiple images at once before narrowing down the search to the oddball image. In particular, the human eye has the flexibility to scan multiple images at once at the cost of not capturing the fine details of each image, or at the other extreme, to focus on one image in order to capture its fine details. It would be interesting to incorporate such nuances in the selection of arms. The trembling hand model, although rendering the problem amenable to analysis, does not capture the aforementioned nuances.

  In addition, the human subject participating in the experiment has limited memory and may not remember the entire history of images observed in the past. However, recall that the policies of Chapters 2, 3 and 4 used the entire history of arm selections and

[1]Best arm in the context of Markov observations from the arms may be defined as the arm with the largest stationary mean.

observations in order to estimate the odd arm at each time instant. It may be interesting to fix attention to policies whose memory is a certain fixed time unit, say T, into the past, and derive lower and upper bounds for the class of such memory-constrained policies.

- **Nonasymptotic regime**

  Throughout the thesis, we study the problem of fixing the error probability and determining the asymptotic growth rate of the expected time to find the odd arm subject to the error probability constraint, where the asymptotics is as the error probability vanishes. It will be interesting to study a practically more relevant version of the problem – the nonasymptotic version wherein the error probability is fixed at some $\epsilon > 0$. Our analyses of the lower bounds for the asymptotic regime reveal that for any given $\epsilon > 0$ and arm configurations C, the expected value of the stopping time of any policy whose error probability is $\leq \epsilon$ is lower bounded by

  $$d(\epsilon, 1-\epsilon) \cdot \alpha(C),$$

  where $\alpha(C)$ is an arms configuration dependent constant. Here, $d(x, y)$ is the KL divergence between the Bernoulli distributions with parameters x and y. This lower bound on the expected time to find the odd arm clearly carries over to the nonasymptotic regime.

  The upper bound in our analyses is asymptotic. For the nonasymptotic regime, it is suggestive to look at the next order term in the expression for the expected time to find the odd arm as explained on pp. 207. It is worth noting that the first order term is $O(\log 1/\epsilon)$.

Bibliography

[1] E. Kaufmann, O. Cappé, and A. Garivier, "On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1–42, 2016. 1, 3, 10, 17, 28, 33, 208

[2] S. Bubeck, R. Munos, and G. Stoltz, "Pure Exploration in Finitely-armed and Continuous-armed Bandits," *Theor. Comput. Sci.*, vol. 412, pp. 1832–1852, Apr. 2011. 2, 9

[3] A. P. Sripati and C. R. Olson, "Global Image Dissimilarity in Macaque Inferotemporal Cortex Predicts Human Visual Search Efficiency," *Journal of Neuroscience*, vol. 30, no. 4, pp. 1258–1269, 2010. 2

[4] N. K. Vaidhiyan, S. P. Arun, and R. Sundaresan, "Neural Dissimilarity Indices that Predict Oddball Detection in Behaviour," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 4778–4796, 2017. 2, 3, 64, 65, 67, 69, 77, 84, 120, 124, 127

[5] N. K. Vaidhiyan, S. Arun, and R. Sundaresan, "Active Sequential Hypothesis Testing with Application to a Visual Search Problem," in *2012 IEEE International Symposium on Information Theory Proceedings*, pp. 2201–2205, IEEE, 2012. 2, 3, 64, 65, 67, 69, 77, 124, 127

[6] N. K. Vaidhiyan and R. Sundaresan, "Learning to Detect an Oddball Target," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 831–852, 2017. 2, 3, 10, 11, 17, 24, 64, 65, 67, 69, 77, 124, 127, 133, 142, 202

[7] P. M. Krueger, M. K. van Vugt, P. Simen, L. Nystrom, P. Holmes, and J. D. Cohen, "Evidence Accumulation Detected in BOLD Signal Using Slow Perceptual Decision Making," *Journal of Neuroscience Methods*, vol. 281, pp. 21–32, 2017. 2

[8] G. R. Prabhu, S. Bhashyam, A. Gopalan, and R. Sundaresan, "Learning to Detect an Oddball Target with Observations from an Exponential Family," *arXiv preprint arXiv:1712.03682*, 2017. 3, 10, 17, 18, 64, 207

[9] H. Chernoff, "Sequential Design of Experiments," *The Annals of Mathematical Statistics*, vol. 30, no. 3, pp. 755–770, 1959. 3, 16, 120

[10] A. E. Albert, "The Sequential Design of Experiments for Infinitely Many States of Nature," *The Annals of Mathematical Statistics*, pp. 774–799, 1961. 3, 11, 22, 24, 62

[11] A. Garivier and E. Kaufmann, "Optimal Best Arm Identification with Fixed Confidence," in *Conference on Learning Theory*, pp. 998–1027, 2016. 3, 24

[12] B. Hemo, K. Cohen, and Q. Zhao, "Asymptotically Optimal Search of Unknown Anomalies," in *Signal Processing and Information Technology (ISSPIT), 2016 IEEE International Symposium on*, pp. 75–80, IEEE, 2016. 3

[13] S. Nitinawarat and V. V. Veeravalli, "Universal Scheme for Optimal Search and Stop," *Bernoulli*, vol. 23, pp. 1759–1783, 08 2017. 3

[14] M. Naghshvar and T. Javidi, "Active M-ary Sequential Hypothesis Testing," in *2010 IEEE International Symposium on Information Theory*, pp. 1623–1627, IEEE, 2010. 3

[15] M. Naghshvar and T. Javidi, "Information Utility in Active Sequential Hypothesis Testing," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 123–129, IEEE, 2010. 3

[16] M. Naghshvar and T. Javidi, "Performance Bounds for Active Sequential Hypothesis Testing," in *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pp. 2666–2670, IEEE, 2011. 3

[17] M. Naghshvar, T. Javidi, *et al.*, "Active Sequential Hypothesis Testing," *The Annals of Statistics*, vol. 41, no. 6, pp. 2703–2738, 2013. 3

[18] J. C. Gittins, "Bandit Processes and Dynamic Allocation Indices," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 148–177, 1979. 8, 65

[19] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotically Efficient Adaptive Allocation Schemes for Controlled Markov Chains: Finite Parameter Space," *IEEE Transactions on Automatic Control*, vol. 34, no. 12, pp. 1249–1259, 1989. 8, 9, 10

[20] V. Moulos, "Optimal best Markovian Arm Identification with Fixed Confidence," in *Advances in Neural Information Processing Systems*, pp. 5606–5615, 2019. 9, 66, 208

[21] A. Wald, "On Cumulative Sums of Random Variables," *The Annals of Mathematical Statistics*, vol. 15, no. 3, pp. 283–296, 1944. 10

[22] I. Csiszar and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011. 16

[23] H. Victor, "A General Class of Exponential Inequalities for Martingales and Ratios," *The Annals of Probability*, vol. 27, no. 1, pp. 537–564, 1999. 39, 60, 104

[24] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, vol. 38. Springer Science & Business Media,, 2009. 41, 49

[25] L. M. Ausubel and R. J. Deneckere, "A Generalized Theorem of the Maximum," *Economic Theory*, vol. 3, no. 1, pp. 99–107, 1993. 48

[26] Q. Zhao, B. Krishnamachari, and K. Liu, "On Myopic Sensing for Multi-Channel Opportunistic Access: Structure, Optimality, and Performance," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5431–5440, 2008. 65

[27] P. Whittle, "Restless Bandits: Activity Allocation in a Changing World," *Journal of applied probability*, vol. 25, no. A, pp. 287–298, 1988. 65

[28] H. Liu, K. Liu, and Q. Zhao, "Learning in a Changing World: Restless Multiarmed Bandit with Unknown Dynamics," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1902–1916, 2012. 66, 135, 202

[29] R. Ortner, D. Ryabko, P. Auer, and R. Munos, "Regret Bounds for Restless Markov Bandits," in *International Conference on Algorithmic Learning Theory*, pp. 214–228, Springer, 2012. 66

[30] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-Time Analysis of the Multiarmed Bandit Problem," *Journal of Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002. 66

[31] S. Grünewälder and A. Khaleghi, "Approximations of the Restless Bandit Problem," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 514–550, 2019. 66

[32] E. Kaufmann, O. Cappé, and A. Garivier, "On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1–42, 2016. 66, 76, 94, 97, 98, 140, 156, 161, 185

[33] A. Deshmukh, S. Bhashyam, and V. V. Veeravalli, "Controlled Sensing for Composite Multihypothesis Testing with Application to Anomaly Detection," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pp. 2109–2113, IEEE, 2018. 66

[34] A. Deshmukh, V. V. Veeravalli, and S. Bhashyam, "Sequential Controlled Sensing for Composite Multihypothesis Testing," *Sequential Analysis*, pp. 1–38, 2021. 66, 208

[35] G. R. Prabhu, S. Bhashyam, A. Gopalan, and R. Sundaresan, "Sequential Multi-Hypothesis Testing in Multi-Armed Bandit Problems: An Approach for Asymptotic Optimality," *arXiv preprint arXiv:2007.12961*, 2020. 66, 142, 202, 208

[36] G. R. Prabhu, S. Bhashyam, A. Gopalan, and R. Sundaresan, "Optimal Odd Arm Identification with Fixed Confidence," *arXiv preprint arXiv:1712.03682*, 2017. 67, 69, 77, 124, 127, 133, 142, 202

[37] K. Avrachenkov and V. S. Borkar, "Whittle Index Based Q-Learning for Restless Bandits with Average Reward," 2020. 68, 128, 206

[38] P. Milgrom and I. Segal, "Envelope Theorems for Arbitrary Choice Sets," *Econometrica*, vol. 70, no. 2, pp. 583–601, 2002. 69, 84, 87, 90, 168

[39] V. S. Borkar, "Control of Markov Chains with Long-Run Average Cost Criterion," in *Stochastic Differential Systems, Stochastic Control Theory and Applications*, pp. 57–77, Springer, 1988. 73, 74, 91

[40] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014. 76, 91, 126, 127

[41] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times*, vol. 107. American Mathematical Soc., 2017. 91, 103, 105, 143, 196

[42] I. Kontoyiannis, L. A. Lastras-Montaño, and S. P. Meyn, "Relative Entropy and Exponential Deviation Bounds for General Markov Chains," in *Proceedings. International Symposium on Information Theory, 2005. ISIT 2005.*, pp. 1563–1567, IEEE, 2005. 123

[43] M. Naghshvar and T. Javidi, "Two-Dimensional Visual Search," in *2013 IEEE International Symposium on Information Theory*, pp. 1262–1266, IEEE, 2013. 130, 208

[44] P. Mandl, "Estimation and Control in Markov Chains," *Advances in Applied Probability*, pp. 40–60, 1974. 131, 132

[45] V. Borkar and P. Varaiya, "Identification and Adaptive Control of Markov Chains," *SIAM Journal on Control and Optimization*, vol. 20, no. 4, pp. 470–489, 1982. 131, 132, 149, 166, 167, 169, 171, 180

[46] K. J. Åström and B. Wittenmark, "On Self Tuning Regulators," *Automatica*, vol. 9, no. 2, pp. 185–199, 1973. 132

[47] B. Doshi and S. E. Shreve, "Strong Consistency of a Modified Maximum Likelihood Estimator for Controlled Markov Chains," *Journal of Applied Probability*, pp. 726–734, 1980. 132

[48] V. Borkar and P. Varaiya, "Adaptive Control of Markov Chains I: Finite Parameter Set," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 953–957, 1979. 132, 134, 144

[49] P. R. Kumar, "A Survey of Some Results in Stochastic Adaptive Control," *SIAM Journal on Control and Optimization*, vol. 23, no. 3, pp. 329–380, 1985. 132

[50] L. M. Ausubel and R. J. Deneckere, "A Generalized Theorem of the Maximum," *Economic Theory*, vol. 3, no. 1, pp. 99–107, 1993. 142

[51] A. Federgruen, A. Hordijk, and H. C. Tijms, "A Note on Simultaneous Recurrence Conditions on a Set of Denumerable Stochastic Matrices," *Journal of Applied Probability*, pp. 842–847, 1978. 166, 167, 169

[52] J. Munkres, *Topology*. Featured Titles for Topology, Prentice Hall, Incorporated, 2000. 172

[53] Y. S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales*. Springer Science & Business Media, 2012. 193

[54] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. 198, 199

[55] S. Krishnasamy, P. Akhil, A. Arapostathis, R. Sundaresan, and S. Shakkottai, "Augmenting Max-Weight with Explicit Learning for Wireless Scheduling with Switching Costs," *IEEE/ACM Transactions on Networking*, vol. 26, no. 6, pp. 2501–2514, 2018. 206

[56] N. K. Vaidhiyan and R. Sundaresan, "Active Search with a Cost for Switching Actions," in *2015 Information Theory and Applications Workshop (ITA)*, pp. 17–24, IEEE, 2015. 206

[57] W. Fleming and P.-L. Lions, *Stochastic Differential Systems, Stochastic Control Theory and Applications: Proceedings of a Workshop, held at IMA, June 9-19, 1986*, vol. 10. Springer Science & Business Media, 2012. 207

[58] J. Scarlett, I. Bogunovic, and V. Cevher, "Overlapping Multi-Bandit Best Arm Identification," in *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 2544–2548, IEEE, 2019. 208