Sequential Multi-Hypothesis Testing in Multi-Armed Bandit Problems: An Approach for Asymptotic Optimality

Gayathri R. Prabhu, Srikrishna Bhashyam, *Senior Member, IEEE*, Aditya Gopalan, and Rajesh Sundaresan, *Senior Member, IEEE*

***Abstract*— We consider a multi-hypothesis testing problem involving a K-armed bandit. Each arm's signal follows a distribution from a vector exponential family. The actual parameters of the arms are unknown to the decision maker. The decision maker incurs a delay cost for delay until a decision and a switching cost whenever he switches from one arm to another. His goal is to minimise the overall cost until a decision is reached on the true hypothesis. Of interest are policies that satisfy a given constraint on the probability of false detection. This is a sequential decision making problem where the decision maker gets only a limited view of the true state of nature at each stage, but can control his view by choosing the arm to observe at each stage. An information-theoretic lower bound on the total cost (expected time for a reliable decision plus total switching cost) is first identified, and a variation on a sequential policy based on the generalised likelihood ratio statistic is then studied. Due to the vector exponential family assumption, the signal processing at each stage is simple; the associated conjugate prior distribution on the unknown model parameters enables easy updates of the posterior distribution. The proposed policy, with a suitable threshold for stopping, is shown to satisfy the given constraint on the probability of false detection. Under a continuous selection assumption, the policy is also shown to be asymptotically optimal in terms of the total cost among all policies that satisfy the constraint on the probability of false detection.**

***Index Terms*— Action planning, active sensing, conjugate prior, exponential family, hypothesis testing, multi-armed bandit, relative entropy, search problems, sequential analysis, switching cost.**

I. Introduction

We consider a multi-hypothesis testing problem involving a K-armed bandit. The observations from each arm i, $1 \leq i \leq K$, follow a distribution from a *vector exponential family* parameterised by its *natural (vector) parameter* $\boldsymbol{\eta}_i$. The parameters $\overline{\boldsymbol{\eta}} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K)$ of the arms are unknown. The parameter $\overline{\boldsymbol{\eta}}$ belongs to one of the sets $\Theta_1, \ldots, \Theta_M$. The goal is to identify the set Θ_l to which it belongs. At each successive stage or round, the decision maker chooses exactly one among the K arms for observation. The decision maker therefore has only a limited view of the true state of nature at each stage. But the decision maker can control his view by choosing the arm to observe. The decision maker also incurs a cost whenever he switches from one arm to another. Specifically, the decision maker has to minimise the overall cost of expected time for a reliable decision plus total switching cost, subject to a constraint on the probability of false detection.

We can model the above problem as a sequential hypothesis testing problem with control [1] and unknown distributions [2] or parameters [3]. The control here is in the choice of arm for observation at each stage which is determined by the sampling strategy of the policy. Many problems fall within the aforementioned framework, for e.g., anomaly detection, best arm identification, A/B testing, etc.; see Section III for several examples.

Given a constraint α on the probability of false detection, let $\Pi(\alpha)$ be the set of admissible policies that satisfy this constraint. Note that the hypotheses are composite, and therefore, each policy in $\Pi(\alpha)$ should satisfy the constraint on probability of false detection for each $\overline{\boldsymbol{\eta}} \in \Theta_l$, for each l. Our objective is to identify a policy $\pi \in \Pi(\alpha)$ such that it stops within a finite time and to characterize for each l and each $\overline{\boldsymbol{\eta}} \in \Theta_l$, the asymptotic growth rate of the total cost with respect to $\log(1/\alpha)$ as α goes to 0. For such a sequential composite hypothesis testing problem with control, a typical sequential policy has two components: a stopping rule that decides whether to stop and make a decision or continue sampling, and, when we continue, a sampling rule that specifies which arm to sample at each time. The stopping rule is typically based on the comparison of a test statistic computed from the observations with a threshold. For our proposed policy, the design of these components and the methodology leading to the design are described in Section I-B.

A. Remarks on the Model Assumptions

1) Exponential Families: Our interest in exponential families is for three reasons.

- It unifies most of the widely used statistical models such as the Gaussian, the Binomial, the Poisson, the Gamma distributions, among countless others.

Manuscript received July 25, 2020; revised November 14, 2021; accepted January 26, 2022. Date of publication March 14, 2022; date of current version June 15, 2022. This work was supported by the Science and Engineering Research Board, Department of Science and Technology, under Grant EMR/2016/002503. *(Corresponding author: Srikrishna Bhashyam.)*

Gayathri R. Prabhu and Srikrishna Bhashyam are with the Department of Electrical Engineering, IIT Madras, Chennai 600036, India (e-mail: skrishna@ee.iitm.ac.in).

Aditya Gopalan is with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India.

Rajesh Sundaresan is with the Department of Electrical Communication Engineering and the Robert Bosch Centre for Cyber-Physical Systems, Indian Institute of Science, Bangalore 560012, India.

Communicated by N. Santhanam, Associate Editor for Signal Processing and Source Coding.

Digital Object Identifier 10.1109/TIT.2022.3159600

- The generalisation forces us to rely on, and therefore bring out, the key properties of exponential families that make the analysis tractable. These include the usefulness of the convex conjugate (or convex dual) of the log partition function, the existence of easily amenable formulae for relative entropy, and the usefulness of the conjugate prior in the analysis.
- The conjugate prior enables extremely easy signal processing for posterior updates. This is of great value in practice.

2) Switching Cost: Incorporation of switching cost is motivated by a number of applications.

- In visual search tasks, where one is searching for say an odd object among many objects, a switching action implies a change in the location of visual focus via movement of the eyes. This is called a *saccade* and results in a delay cost [4].
- In robotics applications, relocation of robots (or other autonomous decision makers such as unmanned aerial vehicles) incurs considerable cost in terms of energy or delay [5].
- In manufacturing industry applications, switching refers to reconfiguration of a production line and causes extra delays [6].

B. Results and Methodology

Converse: We use the results from [7] to obtain an information-theoretic lower bound on the conditional expected total cost for any policy that satisfies an upper bound constraint on the probability of false detection, say α. The lower bound suggests that the conditional expected total cost is asymptotically proportional to $\log(1/\alpha)$, i.e., it grows as $\log(1/\alpha)/D^*(\overline{\boldsymbol{\eta}})$, where $D^*(\overline{\boldsymbol{\eta}})$ is a relative-entropy based constant which we shall study in some detail in this paper. An examination of this lower bound reveals that it may be viewed as the best performance achievable when a more informed decision maker that knows the parameters to be either $\overline{\boldsymbol{\eta}} \in \Theta_l$ or its 'nearest alternative', a suitable $\overline{\boldsymbol{\eta}}' \in \Theta_{-l} := \bigcup_{m=1}^{M} \Theta_m \backslash \Theta_l$, is attempting to decide which of these is true, as quickly as possible in a sequential fashion and with a control on the false alarm probability.

Achievability: A commonly used test in problems with unknown parameters is the generalised likelihood ratio (GLR) test; see for example the text book [8]. The basic idea of the policy, for stopping problems with control, dates back at least to Chernoff's *Procedure A* [1]. In our case, taking a cue from [3], we use a modified GLR where the numerator of the generalised likelihood ratio is replaced by an averaged likelihood function. The average is computed with respect to an artificial prior on the unknown parameters. Each hypothesis is tested against its *nearest alternative* by taking the minimum, across the alternatives, of the modified GLRs. This yields a suitable statistic that quantifies the decision maker's confidence on each hypothesis.

At each stage, then, we choose the hypothesis with the largest statistic. If the statistic exceeds a pre-defined threshold, we declare this hypothesis as the one likely to be true and stop further sampling. Else, we decide randomly, based on a coin toss, whether to sample the current arm or choose another one according to the policy's sampling strategy. This slowed switching is to handle the switching costs. The bias of the coin determines the speed of switching thereby providing a control on the switching cost. The threshold for a decision in our policy, and therefore stoppage of further sampling, depends only on the tolerable probability of false detection (α) and the number of hypotheses (M); in particular, the threshold is not time-varying. We show that such a policy meets the constraint on the probability of false detection (i.e., the policy is admissible). It is in proving this admissibility where the modification to the GLR comes in handy.

As remarked earlier, our approach involves the computation of generalised likelihoods. These provide best estimates of the unknown parameters obtained by (estimation-theoretic) constrained optimisation. We then adopt the principle of *certainty equivalence*, i.e., we assume that the latest estimated parameters are correct, solve an associated (decision-theoretic) optimisation problem for identifying the optimal sampling strategy, and then take actions according to this optimal prescription. The estimated parameters, at best, can approach the true parameters for, after all, the parameters take values in a continuum. This leads to two requirements. First, to enable the convergence to the true parameters, there should be *sufficient exploration*. Second, the arg-max of the decision-theoretic optimisation problem at each stage, assuming the estimated parameters at that stage, may have several solutions and therefore several sampling prescriptions; we then need a *continuous selection* of the arg-max. Otherwise, information will not be gained at the required rate $D^*(\overline{\boldsymbol{\eta}})$ to meet the lower bound.

When a continuous selection exists, with just barely sufficient exploration, we show that the sampling strategy of our proposed policy has performance that is asymptotically close to the lower bound; the asymptotics is as the target probability of false detection α goes down to zero. We also show that, asymptotically, the total cost scales as $\log(1/\alpha)/D^*(\overline{\boldsymbol{\eta}})$, where $D^*(\overline{\boldsymbol{\eta}})$ is the optimal scaling factor suggested by the lower bound. We then show that the continuous selection assumption holds for some examples.

Under the vector exponential family assumption, the information processing at each stage is extremely simple. The decision maker maintains the parameters of the associated conjugate priors, corresponding to the posterior distributions of the model parameters, via easy-to-implement update rules.

C. Closely Related Prior Works

Two special cases of great interest in literature are the cases of *best arm identification* and *odd arm identification*. Garivier and Kaufmann [9] have characterised the complexity of best arm identification in one-parameter bandit problems in the fixed confidence setting. Kaufmann *et al.* [7] have discussed the case of identifying m best arms in a stochastic multi-armed bandit model for both fixed confidence and fixed budget settings. In [4], the authors have considered the odd arm identification problem with switching costs, but the statistics

of the observations were assumed to be known and Poisson-distributed. In [3], the authors have considered a learning setting where the parameters of the Poisson distribution were not known but the switching costs were not taken into account. In earlier versions of this work, [10] (a workshop paper) and [11] (technical report), under some restricting assumptions, we considered the odd arm identification problem with switching costs when the distributions are from a vector exponential family. This work substantially extends the results in [10], [11] to much more general parameter structure classes.

A related problem is discussed in [12], [13] where the authors have considered a multi-hypothesis problem with controlled sensing of the observations. The parameter set is partitioned into various subsets, one for each hypothesis. Each subset is further assumed to be a finite union of convex sets; then projections on closures of such sets exist (uniqueness holds in each closed convex part). In [14], the authors have considered the problem of identifying the partition to which a set of arms belong, given a finitely partitioned universe of such set of arms. In all these works, [12]–[14], the authors have assumed that the observations come from a single parameter exponential family of distributions. The important and practical issue of switching costs is also not taken into consideration. Further, their choice of the statistic requires the employment of a time-varying threshold.

Our work thus provides a significant generalisation of the results in [12]–[14] to general vector exponential families, analyses the effect of switching cost on search complexity, all in the presence of learning. More importantly, we provide a fuller understanding of the subtleties associated with learning the true parameters: the use of forced exploration, the usefulness of the existence of a continuous selection, and the analysis methods that show convergence despite the adaptation based on the estimated parameters that change with time.

For connections to, and limitations of, the classical works of Chernoff [1] and Albert [2], see a detailed summary in [3, Sec. I-A].

D. Our Contributions

- We provide a significant generalisation of the odd arm identification problem in [11] to a much more general sequential hypothesis testing setting. At least six problems already studied in the literature are highlighted as special cases; see Section III-B.
- We provide generalisations of the problems discussed in closely related papers [12], [13] in three aspects:
  - Observations come from a vector exponential family of distributions.
  - We incorporate switching costs based on the idea of slowed switching; see [4], [15] and [16].
  - The threshold for decision is time invariant.
- We show that the proposed policy, which incorporates learning, is asymptotically optimal even with switching costs. This is in the sense that the growth rate of the total cost with switching, as the probability of false detection and the switching parameter that controls the speed of switching are both driven to zero, is the same as that without switching costs. Of course, optimality is only in terms of the growth rate (slope). As our simulations indicate, the slowed switching leads to an additional delay. However, the delay does not affect the growth rate since it does not show up in the slope.
- We highlight why the continuous selection assumption for the arg-max over sampling strategies may be essential to meet, asymptotically, the lower bound.
- We demonstrate the usefulness of forced exploration for learning the parameters, and suggest a range of exploration rates useful in our context.

E. Organisation of the Paper

The paper is organised as follows. In Section II we describe preliminaries related to exponential families. In Section III, we describe the problem model and discuss several examples that fall within our framework. We then preview the main result. In Section IV, we discuss a lower bound on the expected search time for admissible policies. In Section V, we describe the proposed policy that can be made to come arbitrarily close to meeting the lower bound. In Section VI, we provide insightful simulation results that corroborate the developed theory. The proofs and the verification of the assumption of continuous selection for some examples are all relegated to the appendices.

II. Preliminaries: Exponential Family Basics

In this section we discuss formulae associated with the exponential family that will help in our analysis. Those familiar with exponential families may skip this section.

A probability distribution is a member of a vector exponential family if its probability density function (or probability mass function) can be written as

$$f\left(x|\boldsymbol{\eta}\right) = h\left(x\right)\exp\left(\boldsymbol{\eta}^T\mathbf{T}(x) - \mathcal{A}\left(\boldsymbol{\eta}\right)\right) \quad \forall x \in \mathbb{R}, \tag{1}$$

where $\boldsymbol{\eta}$ is the vector parameter of the family, with $\boldsymbol{\eta}$ in some open convex subset Ψ of $\mathbb{R}^d$, $\mathbf{T}(x) \in \mathbb{R}^d$ is the sufficient statistic for the family, and $\mathcal{A}\left(\boldsymbol{\eta}\right)$ is the log partition function given by

$$\mathcal{A}\left(\boldsymbol{\eta}\right) = \log \int_{\mathbb{R}^d} h\left(x\right)\exp\left(\boldsymbol{\eta}^T\mathbf{T}(x)\right)dx.$$

The expression in (1) gives the *canonical* parameterisation of the exponential family. We restrict ourselves to minimal representations [17, p. 40] which enables us to represent the distributions in the family using the *expectation* parameter defined as

$$\boldsymbol{\kappa}(\boldsymbol{\eta}) := E_{\boldsymbol{\eta}}[\mathbf{T}(x)] = \nabla_{\boldsymbol{\eta}}\mathcal{A}\left(\boldsymbol{\eta}\right) \tag{2}$$

whenever $\mathcal{A}(\cdot)$ is continuously differentiable.

We now discuss a few members of the family to get familiar with the notation.

1) **Poisson distribution:** For the Poisson distribution with alphabet $\mathbb{Z}_+$, we have the probability mass function

$$p\left(x|\lambda\right) \;=\; \frac{e^{-\lambda}}{x!}\lambda^x = \frac{1}{x!}\exp\{x\log\lambda - \lambda\}.$$

This belongs to the exponential family with $\boldsymbol{\eta} = \log \lambda$, $\mathbf{T}(x) = x$, $\mathcal{A}(\boldsymbol{\eta}) = \lambda = e^{\boldsymbol{\eta}}$, and $h(x) = \frac{1}{x!}$. The expectation parameter $\boldsymbol{\kappa}(\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}} \mathcal{A}(\boldsymbol{\eta}) = e^{\boldsymbol{\eta}} = \lambda$.

2) **Bernoulli distribution:** For the Bernoulli distribution with parameter p, we have the probability mass function

$$P(x;p) = p^x (1-p)^{(1-x)} \tag{3}$$

$$= \exp\left\{ \left(x \log \frac{p}{1-p} \right) + \log(1-p) \right\}. \tag{4}$$

Here $\boldsymbol{\eta} = \frac{p}{1-p}$, $\mathbf{T}(x) = x$, $\mathcal{A}(\boldsymbol{\eta}) = -\log(1-p)$, $h(x) = 1$, and $\boldsymbol{\kappa} = p$.

3) **Gaussian distribution:** For the Gaussian distribution defined by the mean parameter μ and the variance parameter σ^2, the probability density function is

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} = \frac{1}{\sqrt{2\pi}} \exp\left\{ \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 - \frac{\mu^2}{2\sigma^2} - \log \sigma \right\}.$$

We consider three different cases: a) unknown mean and known variance, b) known mean and unknown variance, and c) both mean and variance unknown. In the second case, we can subtract the mean value and consider them to be distributions with zero mean.

a) **Unknown mean and known variance:** In this case, we have $\boldsymbol{\eta} = \frac{\mu}{\sigma}$, $\mathcal{A}(\boldsymbol{\eta}) = \frac{\mu^2}{2\sigma^2} = \frac{\boldsymbol{\eta}^2}{2}$, $\mathbf{T}(x) = \frac{x}{\sigma}$, $h(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-x^2}{2\sigma^2} \right\}$, and $\boldsymbol{\kappa} = \frac{\mu}{\sigma}$.

b) **Zero mean and unknown variance:** In this case, we have $\boldsymbol{\eta} = \frac{-1}{2\sigma^2}$, $\mathbf{T}(x) = x^2$, $\mathcal{A}(\boldsymbol{\eta}) = \log \sigma$, $h(x) = \frac{1}{\sqrt{2\pi}}$, and $\boldsymbol{\kappa} = \sigma^2$.

c) **Both mean and variance unknown, vector parameter case:** In this case, we have $\boldsymbol{\eta} = \left[\frac{\mu}{\sigma^2} \;\; \frac{-1}{2\sigma^2} \right]^T$, $\mathbf{T}(x) = \left[x \;\; x^2 \right]^T$, and $\mathcal{A}(\boldsymbol{\eta}) = \frac{\mu^2}{2\sigma^2} + \log \sigma = -\frac{\boldsymbol{\eta}_1^2}{4\boldsymbol{\eta}_2} - \frac{1}{2}\log(-2\boldsymbol{\eta}_2)$, $h(x) = \frac{1}{\sqrt{2\pi}}$. The expectation parameter

$$\boldsymbol{\kappa} = \begin{bmatrix} \mu \\ \mu^2 + \sigma^2 \end{bmatrix}.$$

We now continue with some additional observations on exponential families. The mapping $\boldsymbol{\eta} \mapsto \mathcal{A}(\boldsymbol{\eta})$ is strictly convex [17, Prop. 3.1], a fact that can be easily verified via the Hölder inequality. The strictness comes from the minimality of the representation. If $\mathcal{A}(\cdot)$ is twice differentiable, then the Hessian $\mathsf{Hess}(\mathcal{A})(\boldsymbol{\eta})$ is just the covariance of $\mathbf{T}(x)$ when the canonical parameter is $\boldsymbol{\eta}$. If $\mathcal{A}(\cdot)$ twice continuously differentiable, then the covariance matrix is a continuous function of its parameter.

The convex conjugate of $\mathcal{A}(\boldsymbol{\eta})$ evaluated at an arbitrary $\boldsymbol{\kappa}$ and denoted $\mathcal{F}(\boldsymbol{\kappa})$ is given by

$$\mathcal{F}(\boldsymbol{\kappa}) := \sup_{\boldsymbol{\eta} \in \mathbb{R}^d} \{ \boldsymbol{\eta}^T \boldsymbol{\kappa} - \mathcal{A}(\boldsymbol{\eta}) \}; \tag{5}$$

this is also a convex function. Since $\mathcal{A}(\cdot)$ is convex, we can recover $\mathcal{A}(\cdot)$ as the convex conjugate of $\mathcal{F}(\cdot)$, i.e.,

$$\mathcal{A}(\boldsymbol{\eta}) := \sup_{\boldsymbol{\kappa} \in \mathbb{R}^d} \{ \boldsymbol{\eta}^T \boldsymbol{\kappa} - \mathcal{F}(\boldsymbol{\kappa}) \}. \tag{6}$$

We will assume henceforth that $\mathcal{F}(\cdot)$ and $\mathcal{A}(\cdot)$ are strictly convex and C^2 functions (twice continuously differentiable) at all points in their domains of definition. Optimising (5) over $\boldsymbol{\eta}$, recalling the strict convexity of $\mathcal{A}(\cdot)$, we get that the optimising $\boldsymbol{\eta}$ is unique and satisfies $\boldsymbol{\kappa} = \nabla_{\boldsymbol{\eta}} \mathcal{A}(\boldsymbol{\eta})$ which is the expectation parameter (2) evaluated at the optimising $\boldsymbol{\eta}$. Similarly, optimising (6) over $\boldsymbol{\kappa}$, we get an equation analogous to (2), i.e., $\boldsymbol{\eta} = \nabla_{\boldsymbol{\kappa}} \mathcal{F}(\boldsymbol{\kappa})$. Thus the optimising $\boldsymbol{\kappa}$ and $\boldsymbol{\eta}$ are dual to each other and are in one-to-one correspondence. Indeed, we can move from $\boldsymbol{\eta}$ to its corresponding $\boldsymbol{\kappa}$ and from $\boldsymbol{\kappa}$ to its corresponding $\boldsymbol{\eta}$ via

$$\boldsymbol{\kappa}(\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}} \mathcal{A}(\boldsymbol{\eta}) \text{ and } \boldsymbol{\eta}(\boldsymbol{\kappa}) = \nabla_{\boldsymbol{\kappa}} \mathcal{F}(\boldsymbol{\kappa}). \tag{7}$$

From this one-to-one relation between $\boldsymbol{\eta}$ and $\boldsymbol{\kappa}$ in (7), we also have

$$\begin{aligned} \mathcal{F}(\boldsymbol{\kappa}) &= \boldsymbol{\eta}(\boldsymbol{\kappa})^T \boldsymbol{\kappa} - \mathcal{A}(\boldsymbol{\eta}(\boldsymbol{\kappa})), \\ \mathcal{A}(\boldsymbol{\eta}) &= \boldsymbol{\eta}^T \boldsymbol{\kappa}(\boldsymbol{\eta}) - \mathcal{F}(\boldsymbol{\kappa}(\boldsymbol{\eta})). \end{aligned} \tag{8}$$

When we know that a particular $\boldsymbol{\eta}$ and a particular $\boldsymbol{\kappa}$ are dual to each other, we simplify the notation in (8) to

$$\mathcal{F}(\boldsymbol{\kappa}) + \mathcal{A}(\boldsymbol{\eta}) = \boldsymbol{\eta}^T \boldsymbol{\kappa}. \tag{9}$$

That the dual parameter $\boldsymbol{\kappa}(\boldsymbol{\eta})$ (respectively, $\boldsymbol{\eta}(\boldsymbol{\kappa})$) is involved should be clear from the context since, in (9), the supremum that appears in (6) (respectively, (5)) is absent. See [18, Section 3.3.2] for these basic properties on convex duals.

The expressions for Kullback-Leibler (KL) divergence or relative entropy in terms of the natural parameter and in terms of the expectation parameter (by (9)) are

$$D(\boldsymbol{\eta}_1 \parallel \boldsymbol{\eta}_2) := D(f(\cdot|\boldsymbol{\eta}_1) \parallel f(\cdot|\boldsymbol{\eta}_2)) = (\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2)^T \boldsymbol{\kappa}_1 - \mathcal{A}(\boldsymbol{\eta}_1) + \mathcal{A}(\boldsymbol{\eta}_2) \tag{10}$$

$$= (\boldsymbol{\kappa}_2 - \boldsymbol{\kappa}_1)^T \boldsymbol{\eta}_2 + \mathcal{F}(\boldsymbol{\kappa}_1) - \mathcal{F}(\boldsymbol{\kappa}_2). \tag{11}$$

Note that we have used the duality relation between $\boldsymbol{\kappa}_i$ and $\boldsymbol{\eta}_i$, i.e., $\boldsymbol{\kappa}_i = \boldsymbol{\kappa}(\boldsymbol{\eta}_i)$, $i = 1, 2$. Let $\mathsf{Hess}(\mathcal{A})(\cdot)$ denote the Hessian associated with the function $\mathcal{A}(\cdot)$. Another useful formula is obtained by expanding (10) using the Taylor series centred at $\boldsymbol{\eta}_1$:

$$\begin{aligned} D(\boldsymbol{\eta}_1 \parallel \boldsymbol{\eta}_2) &= \mathcal{A}(\boldsymbol{\eta}_2) - \mathcal{A}(\boldsymbol{\eta}_1) - (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1)^T \nabla \mathcal{A}(\boldsymbol{\eta}_1) \\ &= \frac{1}{2} (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1)^T \cdot \int_0^1 (1-t) \mathsf{Hess}(\mathcal{A})(\boldsymbol{\eta}_1 + t(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1)) \, dt \\ &\qquad \cdot (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1). \end{aligned} \tag{12}$$

These useful formulae will be exploited in later sections.

III. Problem Model, Specific Examples, and Preview of the Main Result

In this section, we first discuss the model and explain the costs under consideration. We then provide several examples considered in the literature that are encompassed by our generalised framework. We end the section with a formal problem statement and a preview of the main result.

A. Problem Model

Let the set of arms be denoted as $\mathcal{K} := \{1, 2, \ldots, K\}$. The distribution of the observations from arm i is a member of the vector exponential family with the natural (canonical) parameter $\boldsymbol{\eta}_i \in \Psi_i$ (open, convex subset of a Euclidean space). Let

$$\Omega = \Psi_1 \times \Psi_2 \times \ldots \times \Psi_K. \tag{13}$$

Let $\overline{\boldsymbol{\eta}} \in \Omega$ denote the tuple of vector parameters associated with the set of arms:

$$\overline{\boldsymbol{\eta}} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_K). \tag{14}$$

Let $\mathcal{M} = \{1, 2, \ldots, M\}$ denote the set of hypotheses. Under the hypothesis $m \in \mathcal{M}$, $\overline{\boldsymbol{\eta}} \in \Theta_m$, where $\Theta_m \subset \Omega$. We assume that the sets Θ_m, $m \in \mathcal{M}$ are disjoint. In addition, we also make an assumption that Θ_m is open in aff(Θ_m), i.e., $\Theta_m =$ relint(Θ_m), where aff($\cdot$) is the affine hull and relint($\cdot$) is the relative interior. Recall that we also assume that $\mathcal{A}(\cdot)$ is strictly convex and C^2; so the second central moment exists for each $\overline{\boldsymbol{\eta}} \in \Theta_m$ and for each m, namely, $E_{\boldsymbol{\eta}}[(\mathbf{T}(x) - \boldsymbol{\kappa}(\boldsymbol{\eta}))(\mathbf{T}(x) - \boldsymbol{\kappa}(\boldsymbol{\eta}))^T]$ exists and is finite.

Let $\mathcal{P}(\mathcal{K})$ be the set of probability distributions on $\mathcal{K}$. Let $a_n \in \mathcal{K}$ denote the index of the arm chosen for observation at the instant n, and let x_n denote the value of the observation during instant n. We write x^n for $(x_1, x_2, \ldots, x_n)$ and a^n for $(a_1, a_2, \ldots, a_n)$. At any stage, say n, given the past observations and actions up to time $n-1$, a policy π must choose an action $\overline{A}_n$ of the form:

- $\overline{A}_n = (stop, \delta)$ which is a decision to stop and decide the hypothesis as $\delta \in \mathcal{M}$, or
- $\overline{A}_n = (continue, \boldsymbol{\lambda}, \delta = \mathsf{null})$ which is a decision to continue and sample the next arm to pull according to a probability measure $\boldsymbol{\lambda}$ on the finite set of arms.

We define the stopping time of the policy π as

$$\tau(\pi) := \inf\{n \geq 1 : \overline{A}_n = (stop, \cdot)\}. \tag{15}$$

Given the false detection probability constraint α, with $0 < \alpha < 1$, let $\Pi(\alpha)$ be the set of *admissible policies* that meet the following constraint on the probability of false detection:

$$\Pi(\alpha) \;=\; \{\pi : P(\delta \neq l \mid \overline{\boldsymbol{\eta}}) \leq \alpha, \quad \forall \overline{\boldsymbol{\eta}} \in \Theta_l, \quad \forall l\}, \tag{16}$$

with δ being the decision made when the algorithm stops. It is important to note that a policy is in $\Pi(\alpha)$ only if it does well for each $\overline{\boldsymbol{\eta}} \in \Theta_l$, for each l. Policies tuned to specific $\overline{\boldsymbol{\eta}}$'s or specific Θ_l will likely fail when other hypotheses are in vogue.

B. Examples

We discuss several examples already studied in the literature and show how they fit within our general framework. In each case, we first pose the problem and show how to embed it within our framework. The embedding sheds light on the structural aspects, associated with the parameter sets, of the specific problem under consideration.

- *Generalised best arm identification in a multi-armed bandit setting:* We consider a set of K arms, each following a distribution from the vector exponential family. Our objective is to identify the *best* arm $i \in \mathcal{K}$ such that

$$\mathbf{c}^T \boldsymbol{\eta}_i > \mathbf{c}^T \boldsymbol{\eta}_j, \forall j \in \mathcal{K} \backslash i,$$

where $\mathbf{c} \in \mathbb{R}^d$. We assume it is a priori known that there is exactly one such arm.

We cast this problem into our framework as follows. Let the number of hypotheses $M = K$. The parameter set under hypothesis m can be taken to be

$$\Theta_m = \{\overline{\boldsymbol{\eta}} \in \Omega : \mathbf{c}^T \boldsymbol{\eta}_m > \mathbf{c}^T \boldsymbol{\eta}_{m'}, \forall m' \neq m\}.$$

For the scalar parameter case where the parameter is the mean and $\mathbf{c} = 1$, we have the best arm identification problem; see Fig. 1.

This problem was posed at least as early as Chernoff [1] and was subsequently studied by Albert [2]. For a more recent study with better performance for the best arm problem see Garivier and Kaufmann [9]. For results in the nonasymptotic regime see [19], [20].

- *Multi-bandit best arm identification:* We consider a set of K arms, each following a distribution from the vector exponential family. We assume that there are b possibly overlapping group of arms denoted by the subsets $B_1, B_2, \ldots, B_b$ of $\{1, 2, \ldots, K\}$. Each arm is present in at least one of these groups and each group has at least two arms and a unique best arm. Our objective is to find the best arm in each of these groups, i.e., to find $m = \{m_1, m_2, \ldots, m_b\} \subset \mathcal{K}$ such that for $m_k \in B_k$, and for each $k \in \{1, 2, \ldots, b\}$

$$\mathbf{c}^T \boldsymbol{\eta}_{m_k} > \mathbf{c}^T \boldsymbol{\eta}_j, \quad \forall j \in B_k, \quad j \neq m_k.$$

We can embed this problem into our setting as follows. Let the number of hypothesis $M = |B_1| \times |B_2| \times \cdots \times |B_b|$. Take $m = (m_1, m_2, \ldots, m_b) \in B_1 \times B_2 \times \cdots \times B_p$. The parameter set under the hypothesis m can be defined as

$$\Theta_m = \{\overline{\boldsymbol{\eta}} \in \Omega : \mathbf{c}^T \boldsymbol{\eta}_{m_k} > \mathbf{c}^T \boldsymbol{\eta}_j, \ \forall j \in B_k \setminus \{m_k\}, \quad \forall k\}.$$

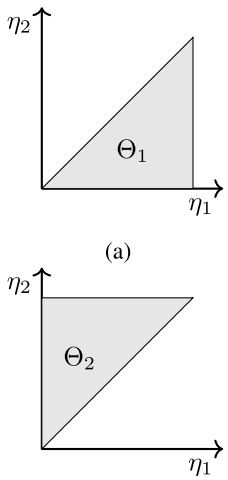

Fig. 1. Scalar best arm identification with $K = 2$. Under hypothesis 1, we have $\eta_1 > \eta_2$ and Θ_1 is the shaded region indicated in (a). Under hypothesis 2, Θ_2 is the shaded region in (b). It is assumed that $\eta_1 \neq \eta_2$.

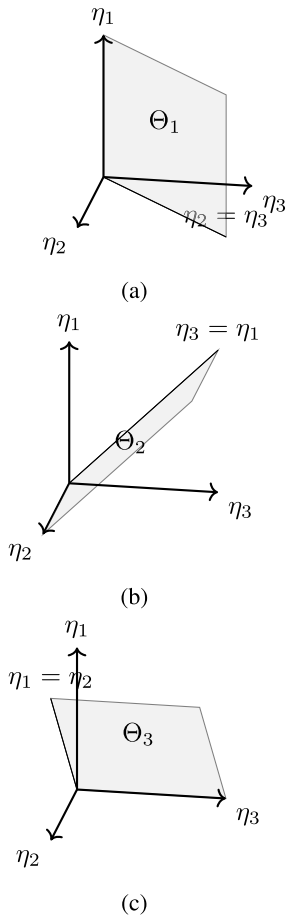

Fig. 2. Odd arm identification with $K = 3$. Under hypothesis 1, $\eta_2 = \eta_3 \neq \eta_1$ and Θ_1 is the shaded portion indicated in (a) excluding the line $\eta_1 = \eta_2 = \eta_3$. In (b) and (c), the shaded portions excluding the line $\eta_1 = \eta_2 = \eta_3$ indicate the parameter sets Θ_2 and Θ_3, respectively.

A special case of this problem was studied in [21], where $\boldsymbol{\eta}$ is the mean parameter and the distributions are sub-Gaussian, i.e., $E[e^{sX}] \leq e^{\frac{\sigma^2 s^2}{2}}, \forall s \in \mathbb{R}$ with $\sigma \leq 1/2$.

- *Odd arm identification in multi-armed bandit setting:* We consider a set of K arms, each following a distribution from the vector exponential family, in which all but one have the same distribution. The objective is to identify the odd arm, i.e., to find the $i \in \mathcal{K}$ such that

$$\boldsymbol{\eta}_i = \boldsymbol{\theta}, \text{ and } \boldsymbol{\eta}_j = \boldsymbol{\theta}', \forall j \in \mathcal{K} \backslash \{i\} \text{ for some } \boldsymbol{\theta}' \neq \boldsymbol{\theta}.$$

The decision maker knows, a priori, that there is such an odd arm.
This problem too can be embedded in our setting as follows. Let the number of hypotheses $M = K$, and let the parameter set under hypothesis m be defined as

$$\Theta_m = \{\overline{\boldsymbol{\eta}} \in \Omega : \boldsymbol{\eta}_m = \boldsymbol{\theta}, \text{ and } \boldsymbol{\eta}_{m'} = \boldsymbol{\theta}', \forall m' \neq m, \ \boldsymbol{\theta}' \neq \boldsymbol{\theta}\},$$

where the hypothesis m indicates that the arm m is the odd one.
This problem was considered in the conference version [10] and its accompanying technical report [11] as the exponential family generalisation of [3], [4] where the observations were restricted to be Poisson random variables. Note that the odd arm detection problem is a particular case of structured best arm identification [22].

- *L-anomalous arms identification in the multi-armed bandit setting:* Here, we consider the case when we have multiple anomalous arms, i.e., out of the K arms, L arms have a distribution different from the rest.
We can embed this problem into our setting as follows. Let the number of hypothesis $M = \binom{K}{L}$ and enumerate the subsets as $S_1, S_2, \ldots, S_M$, $|S_m| = L$, $1 \leq m \leq M$. The parameter set under hypothesis m associated with S_m is defined as

$$\Theta_m = \{\overline{\boldsymbol{\eta}} \in \Omega : \boldsymbol{\eta}_i = \boldsymbol{\theta}, \forall i \in S_m \text{ and } \boldsymbol{\eta}_j = \boldsymbol{\theta}', \forall j \notin S_m, \boldsymbol{\theta}' \neq \boldsymbol{\theta}\}.$$

The hypothesis m indicates that the arms in S_m are anomalous.
This has been considered in [23]. For a summary of various other kinds of anomaly detection problems, see [24].

- *High reward outlier detection in the multi-armed bandit setting:* Consider the problem of identifying outlier arms with extremely high expected reward compared to the other arms. An arm is defined as an outlier if the expectation parameter is greater than the mean plus k times standard deviation of the expectation parameters of all the arms, i.e., arm i is an outlier when

$$\kappa_i > \mu + k\sigma = \theta,$$

where κ_i is the expectation parameter of arm i, μ and σ are calculated as

$$\mu = \frac{1}{K}\sum_{i=1}^{K} \kappa_i \text{ and } \sigma = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(\kappa_i - \mu)^2},$$

respectively. It is a priori known that this set is nonempty.
We can embed this problem too into our setting as follows. Consider a set of K arms, each following a distribution from the exponential family. Let the number of hypothesis $M = 2^K - 1$ and enumerate the subsets P_m of $2^{\mathcal{K}} \backslash \emptyset$, $m = 1, 2, \ldots, M$. Let the parameter set under hypothesis m be defined as, with $\overline{\boldsymbol{\eta}} = (\eta_1, \ldots, \eta_K)$,

$$\Theta_m = \{\overline{\boldsymbol{\eta}} \in \Omega : \kappa_i > \theta, \forall \eta_i \in P_m\}.$$

The hypothesis m indicates that the arms in the set P_m are outliers.
This problem was studied in [25].

- *Partition identification problem in the multi-armed bandit setting:* In this setting, the parameter space is *partitioned* into M sets. The goal is to identify the subset of the partition in which the parameter belongs. The general problem addressed in this paper does not require the $\Theta_m, \ m \in \mathcal{M}$, to be a partition of Ω. Two such problems (with $M = 2$) and their embedding in our framework are described below. We take $\overline{\boldsymbol{\eta}} = (\eta_1, \ldots, \eta_K)$, where the η_i are (in this set of examples) scalars.

– *Threshold crossing problem [14]:* In this setting, our objective is to check if there is at least one arm whose parameter is above a given threshold value u. Define the parameter set Θ_1 to be

$$\Theta_1 = \left\{ \overline{\boldsymbol{\eta}} \in \Omega : \max_{i \leq K} \eta_i > u \right\},$$

and $\Theta_2 = \text{relint}(\Theta_1^c)$.

– *Half-space identification problem [14]:* In this setting, our objective is to identify the half-space that contains the parameters. Fix constants $(a_1, a_2, \ldots, a_k, b)$. Define the parameter set Θ_1 to be

$$\Theta_1 = \left\{ \overline{\boldsymbol{\eta}} \in \Omega : \sum_{i=1}^{K} a_i \eta_i > b \right\},$$

and $\Theta_2 = \text{relint}(\Theta_1^c)$.

– *Norm-threshold problem:* For a given set of parameters, the objective here is to check if the norm of the parameter tuple lies within a threshold. This is embedded in the above framework as follows. Define the parameter set Θ_1 to be

$$\Theta_1 = \{\overline{\boldsymbol{\eta}} \in \Omega : ||\overline{\boldsymbol{\eta}}||_2 < r_1\},$$

and Θ_2 to be

$$\Theta_2 = \{\overline{\boldsymbol{\eta}} \in \Omega : r_1 < ||\overline{\boldsymbol{\eta}}||_2 < r_2\}.$$

This problem is an example of a situation where the parameter set is not a finite union of convex sets.

As one can see from the rich set of examples above, our framework is sufficiently general to cover all these examples considered in the literature.

C. Costs

The total cost is the sum of the switching cost and the delay cost in arriving at a decision, as in [15]. We now make this precise.

1) Switching Cost: Let $g(a, a')$ denote the cost of switching from arm a to arm a'. This is incurred every time a switch of arms is executed. We assume

$$g(a, a') \geq 0 \ \forall a, a' \in \mathcal{K} \text{ and } g(a, a) = 0 \quad \forall a \in \mathcal{K}.$$

The assumption $g(a, a) = 0$ says there is no switching cost if the controller does not switch arms. Define $g_{\max}$ as follows and assume that it is finite:

$$g_{\max} := \max_{a, a' \in \mathcal{K}} g(a, a') < \infty.$$

2) Total Cost: For a policy $\pi \in \Pi(\alpha)$, the total cost $C(\pi)$ is the sum of the stopping time (delay) and the net switching cost:

$$C(\pi) := \tau(\pi) + \sum_{l=1}^{\tau(\pi)-1} g(a_l, a_{l+1}).$$

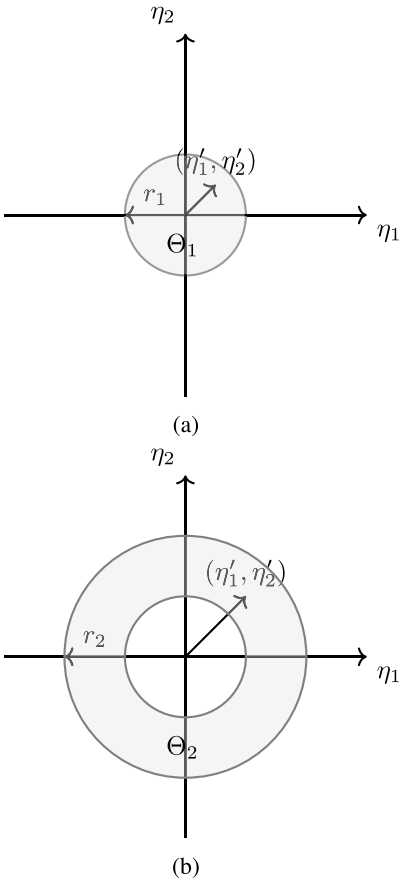

Fig. 3. Norm-threshold problem with $K = 2$. Under hypothesis 1, we have $0 \leq \sqrt{\eta_1^2 + \eta_2^2} < r_1$, and Θ_1 is the shaded portion indicated in (a). Θ_2 for hypothesis 2 is the shaded portion indicated in (b).

D. Problem Statement and a Preview of Main Result

Problem Statement: Our goal is to identify, for each l and for each $\overline{\boldsymbol{\eta}} \in \Theta_l$, the asymptotic growth rate of the cost $\inf_{\pi \in \Pi(\alpha)} E[C(\pi) \mid \overline{\boldsymbol{\eta}}]$ with respect to $\log(1/\alpha)$ as the constraint on the probability of false detection α vanishes. More precisely, we wish to identify

$$\lim_{\alpha \downarrow 0} \inf_{\pi \in \Pi(\alpha)} \frac{E[C(\pi) \mid \overline{\boldsymbol{\eta}}]}{\log(1/\alpha)}.$$

A preview of the main result: We will argue that

$$\lim_{\alpha \downarrow 0} \inf_{\pi \in \Pi(\alpha)} \frac{E[C(\pi) \mid \overline{\boldsymbol{\eta}}]}{\log(1/\alpha)} = \frac{1}{D^*(\overline{\boldsymbol{\eta}})},$$

where $D^*(\overline{\boldsymbol{\eta}})$ is the solution to a max-min problem to be defined later in (18). The converse, as usual, follows from a smart application of the data-processing inequality, and involves the stumbling block of a sequential hypothesis test between $\overline{\boldsymbol{\eta}} \in \Theta_l$ and its nearest alternative in Θ_{-l}.

The achievability result, however, requires us to address several nontrivial and nuanced issues which we now highlight.

- We need the cumulant generating function $\mathcal{A}(\cdot)$ to be strictly convex, a consequence of the minimality of the representation, and further in C^1 (continuously differentiable). The former ensures 1-to-1 correspondence between the $\boldsymbol{\eta}$ and the $\boldsymbol{\kappa}$ parameters. The latter ensures

that the relative entropy is continuous in the parameters of the problem.

- To ensure that the estimated parameters approach the true parameters, one needs sufficient exploration. We use an $O(n^{\beta})$ exploration scheme, where $1/2 < \beta < 1$. See equation (43) in Section V-G.
- For certain concentration results to hold, we need finite second central moments and a regularity condition on relative entropy – that it diverges as the separation between the associated canonical parameters grows without bound. If $\mathcal{A}(\cdot)$ is twice-differentiable, the former corresponds to the positive definiteness of the Hessian matrix and the latter corresponds to its condition number not vanishing too quickly.
- The optimisation problem leads to a set of maximising actions. However these may not be singletons. As a consequence, as we show later, we only get upper semi-continuity of the maximising set-valued correspondence. On the other hand, the use of the certainty equivalence principle involves actions based on maximisers associated with the estimated parameters. One therefore needs a continuous selection for the action mapping.
- As the estimated parameters approach the true parameters, the policies used also vary with time. One needs to show convergence in this complex regime which has time-varying estimates and certainty equivalence based actions.
- On account of zero switching cost for no switching and on account of $g_{\max} < \infty$, the asymptotic growth rate of $E[C(\pi) \mid \overline{\boldsymbol{\eta}}]$ can be made as close to the asymptotic growth rate without switching cost, i.e., the growth rate of $E[\tau(\pi) \mid \overline{\boldsymbol{\eta}}]$, as one wishes. This involves the use of a *sluggish* policy that switches at a very slow rate, yet mimics the stationary distribution associated with the asymptotically optimal policy for no switching costs.

IV. The Converse (Lower Bound on Delay)

The following proposition gives an information theoretic lower bound on the expected conditional stopping time for any policy that belongs to $\Pi(\alpha)$, given the true configuration is $\overline{\boldsymbol{\eta}} \in \Theta_l$. The decision maker a priori does not know either $\overline{\boldsymbol{\eta}}$ or l.

Proposition 1: Fix $0 < \alpha < 1$. Let $\overline{\boldsymbol{\eta}} \in \Theta_l$ be the true configuration. For any $\pi \in \Pi(\alpha)$, we have

$$E[\tau|\overline{\boldsymbol{\eta}}] \geq \frac{d_b(\alpha \parallel 1-\alpha)}{D^*(\overline{\boldsymbol{\eta}})} \tag{17}$$

where $d_b(\cdot \parallel \cdot)$ is the binary relative entropy function, and $D^*(\overline{\boldsymbol{\eta}})$ is defined as

$$D^*(\overline{\boldsymbol{\eta}}) = \sup_{\boldsymbol{\lambda} \in \mathcal{P}(\mathcal{K})} \inf_{\overline{\boldsymbol{\eta}}' \in \Theta_{-l}} \sum_{i=1}^{K} \lambda_i D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i'), \tag{18}$$

where $\overline{\boldsymbol{\eta}}' = (\boldsymbol{\eta}_1', \boldsymbol{\eta}_2', \ldots, \boldsymbol{\eta}_K')$.

Proof: The proof follows the steps in the proof of [3, Prop. 1], which relies on [7, Lem. 1] (see also [26]) and involves an application of the data processing inequality and Wald's lemma. We omit the details. ■

The binary relative entropy function is given by the familiar expression:

$$d_b(\alpha \parallel 1-\alpha) = \alpha \log \frac{\alpha}{1-\alpha} + (1-\alpha) \log \frac{1-\alpha}{\alpha}.$$

As the constraint on the probability of false detection $\alpha \to 0$, we have

$$d_b(\alpha \parallel 1-\alpha) / (\log(1/\alpha)) \to 1.$$

Hence, we get that the conditional expected stopping time of the optimal policy scales at least as $(\log(1/\alpha))/D^*(\overline{\boldsymbol{\eta}})$.

Corollary 2: Fix $0 < \alpha < 1$. Let $\overline{\boldsymbol{\eta}} \in \Theta_l$ be the true configuration. For any $\pi \in \Pi(\alpha)$, we have

$$E[C(\pi) | \overline{\boldsymbol{\eta}}] \geq \frac{d_b(\alpha \parallel 1-\alpha)}{D^*(\overline{\boldsymbol{\eta}})}. \tag{19}$$

Proof: With the switching costs added, we have $C(\pi) \geq \tau(\pi)$. Hence the corollary follows from Proposition 1. ■

Interpretation of the sup-inf optimisation problem in (18)*:* We can interpret $D^*(\overline{\boldsymbol{\eta}})$ as follows. Consider the simpler hypothesis testing problem where the decision maker has to decide between the given $\overline{\boldsymbol{\eta}} \in \Theta_l$ and any alternative chosen from Θ_{-l} by an adversary. The decision maker may choose a sampling strategy $\boldsymbol{\lambda} \in \mathcal{P}(\mathcal{K})$. Knowing this, the adversary may pick, from Θ_{-l}, the nearest alternative to $\overline{\boldsymbol{\eta}}$ that minimises the separation as measured by the $\boldsymbol{\lambda}$-weighted average of the relative entropies of the arms. Realising this, the decision maker will ensure that his chosen $\boldsymbol{\lambda}$ maximises the minimum separation. This is the decision maker's best guarding policy against the (more informed) adversary's strategy of picking the nearest alternative to $\overline{\boldsymbol{\eta}}$ from outside Θ_l, i.e., Θ_{-l}.

In the next section, we discuss how to convert the above intuitive interpretation into a policy that achieves this lower bound.

V. A Sluggish and Modified GLR Test With Forced Exploration

In this section, we describe a suitable policy that can achieve the growth rate in the lower bound in Proposition 1, as the constraint on the probability of false detection is driven to zero. The proposed policy is a modification of the policy π_{SM} discussed in [11] where we replace the random sampling strategy by a forced exploration technique[1] as in [9].

Let us denote by $\mathcal{A}_i(\cdot)$ the cumulant generating function associated with arm i. Recall the assumption that $\mathcal{A}_i(\cdot)$ is C^2. This ensures $\boldsymbol{\kappa}_i(\cdot)$ is continuous and furthermore that the relative entropy $D(\cdot \parallel \boldsymbol{\eta}_i')$ is continuous for each fixed $\overline{\boldsymbol{\eta}}'$. As a consequence, for any i and any $\boldsymbol{\eta}_i, \boldsymbol{\eta}_i' \in \Psi_i$ such that $\boldsymbol{\eta}_i \neq \boldsymbol{\eta}_i'$, we have $D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i') < \infty$. Furthermore, on account of the C^2 condition, the observations have finite second central moments. Indeed, as already highlighted, the Hessian matrix of $\mathcal{A}_i(\boldsymbol{\eta}_i)$ is just the covariance of $\mathbf{T}(x)$ when the parameter is $\boldsymbol{\eta}_i$, the strict convexity of $\mathcal{A}_i(\boldsymbol{\eta}_i)$ at all $\boldsymbol{\eta}_i$ is the same as positive definiteness of the associated covariance matrix, and the C^2 condition on $\mathcal{A}_i(\boldsymbol{\eta}_i)$ is the same as saying that the covariance matrix of $\mathbf{T}(x)$ has entries that are continuous in the parameter $\boldsymbol{\eta}_i$.

[1]The subscript SM in π_{SM} stands for "sluggish" and "modified". We shall use π_{SMF} where the added letter F stands for "forced exploration".

A. Continuous Selection of the Optimal Sampling Strategy

In this section, we will highlight the usefulness of a continuous selection of the sampling strategy. To set the stage, we first prove the following property about the attainment of the supremum in the definition of $D^*(\overline{\boldsymbol{\eta}})$ in (18) and the upper semi-continuity of the supremum map. The possibility that this upper semi-continuity may not extend to continuity in general leads us to introduce Assumption A on the existence of a continuous selection of optimal sampling strategies.

Proposition 3: The supremum in (18) is a maximum, i.e., for each l and for each $\overline{\boldsymbol{\eta}} \in \Theta_l$, we can write

$$\boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}}) = \operatorname*{argmax}_{\boldsymbol{\lambda}\in\mathcal{P}(\mathcal{K})} \inf_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} \sum_{i=1}^{K} \lambda_i D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i'). \tag{20}$$

Furthermore the mapping $\overline{\boldsymbol{\eta}} \mapsto \boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}})$ is an upper semi-continuous convex-valued correspondence.

Proof: See Appendix B. ∎

In any problem instance of the type considered in this paper, the learner does not know the true parameters, and must learn these parameters along the way. Our strategy to attain the lower bound is to estimate the parameters $\overline{\boldsymbol{\eta}}$, say via $\widehat{\overline{\boldsymbol{\eta}}}(n)$ at time instant n, to apply the certainty equivalence principle by taking the estimated $\widehat{\overline{\boldsymbol{\eta}}}(n)$ to be true parameter, and then to apply a sampling strategy from the set $\boldsymbol{\lambda}^*(\widehat{\overline{\boldsymbol{\eta}}}(n))$. Since the estimated parameter can at best approach the true parameter as time $n \to \infty$, for our scheme to work, continuity in the sampling strategy will prove beneficial. If the desired continuity does not hold, the rate at which information and therefore confidence is gathered, based on $\boldsymbol{\lambda}^*(\widehat{\overline{\boldsymbol{\eta}}}(n))$ may not match the rate at which information should be gathered, as per $\boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}})$, to meet the lower bound. Observe however that the mapping $\boldsymbol{\eta} \mapsto \boldsymbol{\lambda}^*(\boldsymbol{\eta})$ is only an upper semicontinuous correspondence, and may not possess, in general, a continuous selection [27, Sec. 9.2]. We shall therefore make the following assumption on the existence of a continuous selection. Let

$$F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}}) := \inf_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} \sum_{i=1}^{K} \lambda_i D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i'). \tag{21}$$

Then $\boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}})$ in (20) optimises $F(\cdot, \overline{\boldsymbol{\eta}})$.

Assumption A: The correspondence $\overline{\boldsymbol{\eta}} \mapsto \boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}})$ admits a continuous selection.

This assumption holds for example when

(i) $\boldsymbol{\lambda}^*(\cdot)$ is single-valued, a condition that holds when $F(\cdot, \overline{\boldsymbol{\eta}})$ is strictly concave for each $\boldsymbol{\eta} \in \Theta_l$, see [28, Th. 9.17]; or

(ii) $\boldsymbol{\lambda}^*(\cdot)$ is lower semicontinuous, see [27, Sec. 9.1].

We shall verify in Appendix F that Assumption A holds in the examples of odd arm and best arm identifications. There are interesting situations where we have not yet been able to establish that Assumption A holds, and they involve nonconvex sets that are not necessarily finite unions of convex sets, e.g., two-dimensional independent Gaussians with variances 1 but unknown means η_1 and η_2, with $\Theta_1 = \{\overline{\boldsymbol{\eta}} = (\eta_1, \eta_2) \ : \ \eta_1^2 + \eta_2^2 < 1\}$, and $\Theta_2 = \{\overline{\boldsymbol{\eta}} = (\eta_1, \eta_2) \ : \ 1 < \eta_1^2 + \eta_2^2 < 2\}$. Whether a continuous selection exists for this setting is still open.

B. Additional Notations

Let N_i^n denote the number of times the arm i was chosen for observation up to time n, i.e.,

$$N_i^n = \sum_{t=1}^{n} 1_{\{a_t = i\}}, \tag{22}$$

where a_t is the arm chosen at time t. Clearly $n = \sum_{i=1}^{K} N_i^n$. Let $\mathbf{Y}_i^n$ denote the sum of sufficient statistic of arm i up to time n, i.e.,

$$\mathbf{Y}_i^n = \sum_{t=1}^{n} \mathbf{T}(x_t) 1_{\{a_t = i\}}. \tag{23}$$

We will use the letter $f(\cdot)$ to denote all probability density functions. Conditional densities will be denoted by $f(\cdot|\cdot)$. The argument(s) will help identify the appropriate random variable(s) whose density (conditional density) is being represented. We also use it to denote *likelihoods* and *conditional likelihoods* without the normalisation needed to make them probability densities or conditional probability densities.

C. Likelihood Function

Let $f(x^n|a^n, \overline{\boldsymbol{\eta}})$ be the likelihood function of the observations upto time n, conditioned on the actions and the parameters $\overline{\boldsymbol{\eta}}$, i.e.,

$$\begin{aligned}
&f(x^n|a^n, \overline{\boldsymbol{\eta}}) \\
&\quad = \prod_{t=1}^{n} f(x_t|a_t, \boldsymbol{\eta}_{a_t}) && (24)\\
&\quad = \prod_{t=1}^{n} h(x_t) \exp\left\{\boldsymbol{\eta}_{a_t}^T \mathbf{T}(x_t) - \mathcal{A}_{a_t}(\boldsymbol{\eta}_{a_t})\right\} && (25)\\
&\quad = \left(\prod_{t=1}^{n} h(x_t)\right) \cdot \\
&\qquad \prod_{i=1}^{K} \exp\left\{\boldsymbol{\eta}_i^T \sum_{t=1}^{n} \mathbf{T}(x_t) 1_{\{a_t=i\}} - N_i^n \mathcal{A}_i(\boldsymbol{\eta}_i)\right\} && (26)\\
&\quad = \left(\prod_{t=1}^{n} h(x_t)\right) \prod_{i=1}^{K} \exp\left\{\boldsymbol{\eta}_i^T \mathbf{Y}_i^n - N_i^n \mathcal{A}_i(\boldsymbol{\eta}_i)\right\}. && (27)
\end{aligned}$$

Then the log likelihood function is

$$\begin{aligned}
&\log f(x^n|a^n, \overline{\boldsymbol{\eta}}) \\
&\quad = \sum_{t=1}^{n} \log h(x_t) + n \sum_{i=1}^{K} \frac{N_i^n}{n} \left\{\boldsymbol{\eta}_i^T \frac{\mathbf{Y}_i^n}{N_i^n} - \mathcal{A}_i(\boldsymbol{\eta}_i)\right\} && (28)\\
&\quad = \sum_{t=1}^{n} \log h(x_t) + n \sum_{i=1}^{K} w_i \left\{\boldsymbol{\eta}_i^T \hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\boldsymbol{\eta}_i)\right\}, && (29)
\end{aligned}$$

where $w_i := N_i^n/n$ and $\hat{\boldsymbol{\kappa}}_i := \mathbf{Y}_i^n/N_i^n$.

D. Maximum Likelihood Function

Consider a sequence $\delta_n \to 0$. Let the maximum likelihood estimates of the natural parameters under the hypothesis m be

defined as

$$\overline{\boldsymbol{\eta}}^*(m) := (\boldsymbol{\eta}_1^*(m), \ldots, \boldsymbol{\eta}_K^*(m)) \in \operatorname*{argmax}_{\overline{\boldsymbol{\eta}} \in \Theta_m} \sum_{i=1}^{K} w_i \left\{ \boldsymbol{\eta}_i^T \hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\boldsymbol{\eta}_i) \right\} \tag{30}$$

if the maximum is attained, and as some $\overline{\boldsymbol{\eta}}^*(m) \in \Theta_m$ so that

$$\left| \sup_{\overline{\boldsymbol{\eta}} \in \Theta_m} \sum_{i=1}^{K} w_i \left[\boldsymbol{\eta}_i^T \hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\boldsymbol{\eta}_i) \right] - \sum_{i=1}^{K} w_i \left[\boldsymbol{\eta}_i^*(m)^T \hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\boldsymbol{\eta}_i^*(m)) \right] \right| \leq \delta_n \tag{31}$$

if the maximum is not attained. When the maximum exists, the expression for relative entropy and some algebraic manipulation shows that

$$\overline{\boldsymbol{\eta}}^*(m) \in \operatorname*{argmin}_{\overline{\boldsymbol{\eta}} \in \Theta_m} \sum_{i=1}^{K} w_i D(\hat{\boldsymbol{\eta}}_i \parallel \boldsymbol{\eta}_i). \tag{32}$$

We differ here from Deshmukh *et al.* [12] in that our estimate is an ML estimate that recognises that $\overline{\boldsymbol{\eta}} \in \Theta_m$ while Deshmukh *et al.* [12] first optimise over all $\overline{\boldsymbol{\eta}}$ and then project this value on to Θ_m. So our approach is more direct, but requires the existence of a continuous selection (Assumption A). In this context, note that Deshmukh et al. [12] also assume continuous selection by asking for the $\boldsymbol{\lambda}^*(\cdot)$ to be single-valued (sufficient condition (i) for Assumption A to hold).

The log ML function is obtained as

$$\begin{aligned} &\log \hat{f}(x^n | a^n, \overline{\boldsymbol{\eta}} \in \Theta_m) \\ &= \sum_{t=1}^{n} \log h(x_t) \\ &\quad + n \sum_{i=1}^{K} w_i \left\{ \boldsymbol{\eta}_i^{*T}(m) \hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\boldsymbol{\eta}_i^*(m)) \right\}. \end{aligned} \tag{33}$$

E. Average Likelihood Function

When the parameters are unknown, a natural conjugate prior on the parameter $\boldsymbol{\eta}_i$ enables easy updates of the posterior distribution based on observations. The conjugate prior, also denoted $f(\overline{\boldsymbol{\eta}} | \overline{\boldsymbol{\eta}} \in \Theta_l)$, is taken to be a product distribution over $i = 1, \ldots, K$, with each marginal coming from an exponential family of the same form characterised by the hyper-parameters $\overline{\boldsymbol{\Upsilon}} = (\boldsymbol{\Upsilon}_1, \boldsymbol{\Upsilon}_2, \ldots, \boldsymbol{\Upsilon}_K)$ and $\mathbf{n_0} = (n_{01}, \ldots, n_{0K})$, i.e.,

$$f(\overline{\boldsymbol{\eta}} | \overline{\boldsymbol{\eta}} \in \Theta_l) = \begin{cases} \mathcal{H}_l(\overline{\boldsymbol{\Upsilon}}, \mathbf{n_0}) \prod_{i=1}^{K} \exp\left\{ \boldsymbol{\eta}_i^T \boldsymbol{\Upsilon}_i - n_{0i} \mathcal{A}_i(\boldsymbol{\eta}_i) \right\}, & \text{if } \overline{\boldsymbol{\eta}} \in \Theta_l \\ 0, & \text{otherwise} \end{cases} \tag{34}$$

with

$$\mathcal{H}_l(\overline{\boldsymbol{\Upsilon}}, \mathbf{n_0}) = \left[\int_{\overline{\boldsymbol{\eta}} \in \Theta_l} \prod_{i=1}^{K} \exp\left\{ \boldsymbol{\eta}_i^T \boldsymbol{\Upsilon}_i - n_{0i} \mathcal{A}_i(\boldsymbol{\eta}_i) \right\} d\overline{\boldsymbol{\eta}} \right]^{-1} \tag{35}$$

as the normalising factor. We remark that the conjugate prior is an *artificial prior*, used mainly as an analytical artifice for easy posterior updates.

The average likelihood function at time n, averaged according to the artificial prior in (34), is

$$\tilde{f}_l(x^n | a^n) := \int_{\overline{\boldsymbol{\eta}} \in \Theta_l} f(x^n | a^n, \overline{\boldsymbol{\eta}}) \cdot f(\overline{\boldsymbol{\eta}} | \overline{\boldsymbol{\eta}} \in \Theta_l) d\overline{\boldsymbol{\eta}} \tag{36}$$

$$= \left(\prod_{t=1}^{n} h(x_t) \right) \mathcal{H}_l(\overline{\boldsymbol{\Upsilon}}, \mathbf{n_0}) \int_{\overline{\boldsymbol{\eta}} \in \Theta_l} \exp\left\{ \sum_{i=1}^{K} \boldsymbol{\eta}_i^T (\mathbf{Y}_i^n + \boldsymbol{\Upsilon}_i) - (N_i^n + n_{0i}) \mathcal{A}_i(\boldsymbol{\eta}_i) \right\} d\overline{\boldsymbol{\eta}} \tag{37}$$

$$= \left(\prod_{t=1}^{n} h(x_t) \right) \frac{\mathcal{H}_l(\overline{\boldsymbol{\Upsilon}}, \mathbf{n_0})}{\mathcal{H}_l(\mathbf{Y} + \overline{\boldsymbol{\Upsilon}}, \mathbf{N} + \mathbf{n_0})}, \tag{38}$$

where $\mathbf{Y} = (\mathbf{Y}_1^n, \ldots, \mathbf{Y}_K^n)$ and $\mathbf{N} = (N_1^n, \ldots, N_K^n)$ (with the dependence of $\mathbf{Y}$ and $\mathbf{N}$ on n understood and suppressed). The equality in (37) is obtained by substituting (27) and (34) in (36). We then use (35) in (37) to get the final expression in (38). Taking the log, we get

$$\begin{aligned} \log \tilde{f}_l(x^n | a^n) &= \sum_{t=1}^{n} \log h(x_t) + \log \mathcal{H}_l(\overline{\boldsymbol{\Upsilon}}, n_0) \\ &\quad - \log \mathcal{H}_l(\mathbf{Y} + \overline{\boldsymbol{\Upsilon}}, \mathbf{N} + \mathbf{n_0}). \end{aligned} \tag{39}$$

F. Modified GLR Statistic

We define the modified GLR of hypothesis l against hypothesis m as

$$Z_{lm}(n) = \log \frac{\tilde{f}_l(x^n | a^n)}{\hat{f}(x^n | a^n, \overline{\boldsymbol{\eta}} \in \Theta_m)} \tag{40}$$

$$\begin{aligned} &= \log \mathcal{H}_l(\overline{\boldsymbol{\Upsilon}}, \mathbf{n_0}) - \log \mathcal{H}_l(\mathbf{Y} + \overline{\boldsymbol{\Upsilon}}, \mathbf{N} + \mathbf{n_0}) \\ &\quad - n \sum_{i=1}^{K} w_i \left\{ \boldsymbol{\eta}_i^{*T}(m) \hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\boldsymbol{\eta}_i^*(m)) \right\}, \end{aligned} \tag{41}$$

which is obtained using (33) and (39). The modification to the standard GLR is that the numerator contains an averaged likelihood function, averaged with respect to the artificial prior, rather than the maximum likelihood function. As we shall soon see, it is this modification that enables us to make the resulting policy admissible (i.e., a policy in $\Pi(\alpha)$).

Now let

$$Z_l(n) = \min_{m \neq l} Z_{lm}(n) \tag{42}$$

denote the modified GLR of hypothesis l against its nearest alternative. The value of $Z_l(n)$ is a measure of the decision maker's confidence in the hypothesis l.

G. Policy

Let us denote our policy as $\pi_{SMF}(L, \gamma, \beta)$ with SMF standing for Sluggish, Modified GLR based test with Forced exploration, where L is a threshold parameter, γ is a switching

Policy $\pi_{SMF}(L,\gamma,\beta)$:

Fix $L \geq 1$, $0 < \gamma \leq 1$, $1/2 < \beta < 1$.
Initialize: Sample the first arm $a_1 = 1$, set $n^a = 1$, $N_1^{n,a} = 1$, $N_i^{n,a} = 0\ \forall i \neq 1$, $N_1^n = 1$, $N_i^n = 0\ \forall i \neq 1$.
For $n = 1, 2, \ldots$, do:

- $l^*(n) = \text{argmax}_l\, Z_l(n)$. Resolve ties uniformly at random.
- If $Z_{l^*(n)} < \log((M-1)L)$ then choose a_{n+1} via the following, and make the associated updates:
  - Generate $U_{n+1} \sim Bern(\gamma)$, independent of all other random variables.
  - If $U_{n+1} = 0$, $a_{n+1} = a_n$.
  - If $U_{n+1} = 1$, then update $n^a = n^a + 1$ and choose a_{n+1} according to

$$a_{n+1} \in \begin{cases} \underset{i}{\text{argmin}}\, N_i^{n,a} & \text{if } \ni i : N_i^{n,a} < (n^a)^\beta - (\beta K)^{\beta/(1-\beta)} \\ \underset{i}{\text{argmax}}\{n^a \lambda_i^*(\overline{\boldsymbol{\eta}}^*(l^*(n)) - N_i^{n,a}\} & \text{otherwise.} \end{cases} \tag{43}$$

    Resolve ties uniformly at random.
    Update $N_i^{n,a}$ as $N_i^{n,a} = N_i^{n,a} + 1$, whenever $a_{n+1} = i$.
  - $N_i^n = N_i^n + 1$, whenever $a_{n+1} = i$.
- If $Z_{l^*(n)} \geq \log((M-1)L)$, then stop and declare $\delta = l^*(n)$ as the decision.

parameter, and β is a forced exploration parameter, to be explained soon. The policy will involve some new variables: n^a is the number of instants when the decision maker *actively* samples using (43) above, and $N_i^{n,a}$ is the number of times the arm i is *actively* sampled. These will be clear from the pseudo-code.

We now explain the sampling rule in words. When $U_{n+1} = 0$, the sampling arm is not changed for the next instant, i.e., there is no switching. Our policy is *sluggish* because of this possibility of no switching. When $U_{n+1} = 1$, we actively sample based on the sampling rule in (43). This is a variation on the *D-tracking* sampling rule of [9] that includes a forced exploration component. In the above policy, as already described, n^a is the number of instants when the decision maker *actively* samples using (43) and $N_i^{n,a}$ is the number of times the arm i is *actively* sampled. N_i^n is the number of times arm i is sampled, actively or otherwise, up to time n. The threshold to stop is $\log((M-1)L)$ which depends on the threshold parameter L. Observe that the threshold is fixed upfront and does not change over time. It is important to note that the switching parameter γ cannot be chosen to be 0 (for ergodicity considerations).

H. Main Result

We can now state and prove the main result.

Theorem 4: Let Assumption A hold. Fix l. Consider K arms with the configuration $\overline{\boldsymbol{\eta}} \in \Theta_l$. Let $(\alpha_n)_{n\geq 1}$ be a sequence of tolerances such that $\lim_{n\to\infty} \alpha_n = 0$. Then, for each n and for each $\gamma > 0$, the policy $\pi_{SMF}(L_n, \gamma, \beta)$ with $L_n = 1/\alpha_n$ belongs to $\Pi(\alpha_n)$. Furthermore,

$$\begin{aligned} &\liminf_{n\to\infty} \inf_{\pi\in\Pi(\alpha_n)} \frac{E[C(\pi)\,|\overline{\boldsymbol{\eta}}]}{\log(L_n)} \\ &= \lim_{\gamma\downarrow 0} \lim_{n\to\infty} \frac{E[C(\pi_{SMF}(L_n,\gamma,\beta))\,|\overline{\boldsymbol{\eta}}]}{\log(L_n)} = \frac{1}{D^*(\overline{\boldsymbol{\eta}})}. \end{aligned} \tag{44}$$

Proof: The main steps of the proof are to verify that the n-indexed sequence of policies $\pi_{SMF}(L_n,\gamma,\beta)$ satisfies the following.

1) For each n, $\pi_{SMF}(L_n,\gamma,\beta)$ stops in finite time.
2) For each n, $\pi_{SMF}(L_n,\gamma,\beta) \in \Pi(\alpha_n)$, i.e., it is admissible with error tolerance α_n.
3) As $n \to \infty$, the sequence of policies $\pi_{SMF}(L_n,\gamma,\beta)$ indexed by n can be made arbitrarily close to being asymptotically optimal by a suitable choice of γ.

We proceed to show these in the Propositions 5, 6, and 7 next.

Let us begin with the assertion that the proposed policy almost surely (a.s.) stops in finite time.

Proposition 5 (Probability of Stopping in Finite Time): Fix the threshold parameter $L > 1$ and switching parameter $0 < \gamma \leq 1$. Fix $l \in \{1, \cdots, M\}$. Let $\overline{\boldsymbol{\eta}} \in \Theta_l$ be the true configuration. Then, the policy $\pi_{SMF}(L,\gamma,\beta)$ stops in finite time with probability 1, i.e.,

$$P(\tau(\pi_{SMF}(L,\gamma,\beta)) < \infty | \overline{\boldsymbol{\eta}}) = 1.$$

Proof: To prove this, we show that under the true hypothesis, almost surely, the test statistic $Z_l(n)$ grows as $\Omega(n^\beta)$ and, therefore, crosses the threshold $\log((M-1)L)$ in finite time. See Appendix C for details. ■

We next assert the admissibility of the proposed policy.

Proposition 6 (Admissibility): Fix $\alpha > 0$, $\gamma > 0$, and let $L = 1/\alpha$. We then have $\pi_{SMF}(L,\gamma,\beta) \in \Pi(\alpha)$.

Proof: The proof exploits the properties of the modified GLR and involves a change of measure argument. See Appendix D. ■

We next assert that our policy is not only admissible, but is also asymptotically arbitrarily close to the lower bound.

Proposition 7 (Achievability): Fix $\gamma > 0$. Consider the policy $\pi_{SMF}(L,\gamma,\beta)$. Let $\overline{\boldsymbol{\eta}} \in \Theta_l$ be the true configuration. Under Assumption A, we have

$$\limsup_{L\to\infty} \frac{\tau(\pi_{SMF}(L,\gamma,\beta))}{\log(L)} \leq \frac{1}{D^*(\overline{\boldsymbol{\eta}})} \quad \text{a.s.}, \tag{45}$$

$$\limsup_{L\to\infty} \frac{E[\tau(\pi_{SMF}(L,\gamma,\beta)) \mid \overline{\boldsymbol{\eta}}]}{\log(L)} \leq \frac{1}{D^*(\overline{\boldsymbol{\eta}})}, \tag{46}$$

and furthermore,

$$\limsup_{L\to\infty} \frac{E\left[C(\pi_{SMF}(L,\gamma,\beta)) \mid \overline{\boldsymbol{\eta}}\right]}{\log(L)} \leq \frac{1}{D^*(\overline{\boldsymbol{\eta}})} + \frac{g_{\max}\gamma}{D^*(\overline{\boldsymbol{\eta}})} \tag{47}$$

Proof: The main ideas are as follows.

The key to showing (45) is that, as the target probability of false alarm goes to 0, the policy must wait longer and longer to decide. But this, along with the chosen sampling strategy and forced exploration, ensures that the estimated parameters approach the true parameters. By Assumption A, the continuously evolving sampling strategy approaches the desired sampling strategy that guards the correct hypothesis against its nearest alternative. Since relative entropy is continuous in its parameters, the logarithm of the GLR grows at the correct rate, almost surely, and reaches the threshold for a decision within the desired time duration.

The main idea behind (46) is to leverage the second moment condition for uniform integrability. This then helps us turn an almost-sure bound into a bound on the expectation. For the second moment condition itself, concentration inequalities play a crucial role.

The proof of (47) leverages the fact that the total cost is upper bounded by $(1+g_{\max}\gamma)$ times the delay cost. Since the $g_{\max}$ is finite and $\gamma > 0$ is arbitrary, we obtain this inequality as well.

See Appendix E for the proof of each of these. ■

Propositions 5, 6, and 7 combined with Corollary 2 establish Theorem 4. ■

Finally, even when the observations are not from the exponential family, we anticipate the expected delay to be $\Theta(\log(1/\alpha))$ except in degenerate cases (e.g. $D^*(\cdot) = \infty$). The lower bound is applicable even for this case. However, a new approach would be required for the analysis of the proposed policy.

VI. Simulation Results

We end the main body of the paper with some simulation results.

A. Odd Arm Identification

In Figs. 4(a) and 4(b), we plot the average delay and average cost versus $\log(L)$, where $\alpha = 1/L$, for the odd arm identification problem with 8 arms. The observations from the arms are Gaussian with unknown means but known variance. The odd arm has mean 1 while the other arms are zero mean. All arms have unit variance. The forced exploration parameter $\beta = 0.5$. Each plot contains the asymptotic lower bound (dashed curve) and five other curves for different values of the sluggishness parameter $\gamma \in \{0.1, 0.2, 0.4, 0.5, 1.0\}$. Smaller values of γ correspond to the sluggish implementation, while $\gamma = 1$ permits switching at each stage. Each simulation point is obtained by averaging over 5000 realizations.

We observe that the slope of the empirical average delay matches the slope of the lower bound, thereby validating the asymptotic optimality of the policy. The slope of the empirical average values of cost also matches the slopes of the lower bounds, as expected, for small γ. For smaller values of γ, the average delay in arriving at a decision increases (due to limited exploration). As γ decreases, the total cost first decreases due to reduced switching. But, as the value of γ further decreases, we observe that the policy becomes sluggish, thereby resulting in an increased average cost. A value of γ of around 0.2 seems to be the best choice for the chosen settings. In our analysis, we noted that the forced exploration parameter can be chosen between 0.5 and 1. In Figs. 4(c) and 4(d), we plot the results when the forced exploration parameter $\beta = 0.75$. As expected, the results are similar to the case when $\beta = 0.5$.

In Figs. 5(a)-(d), we show the results for the odd arm identification problem with 8 arms where the observations are Gaussian with unknown variance but known mean. Here we choose $\mu_1 = 0$, $\sigma_1^2 = 5$, $\mu_2 = 0$, $\sigma_2^2 = 1$. Once again, we observe that the slope of the empirical average delay matches the slope of the lower bound, and the slope of the empirical average values of cost also matches the slope of the lower bound, as expected, for small γ. A value of γ of around 0.2 seems to be the best choice for this setting as well.

In Figs. 6(a)-(d), we show the results for the odd arm identification problem with both mean and variance unknown. This is an example of a vector parameter exponential family. Again, there is a total of $K = 8$ arms. The observations are similar to those for Figs. 4 and 5.

B. Norm-Threshold Problem

Now, we present results for the norm-threshold problem. In this problem, we have two arms with Gaussian observations and mean and variance parameters $\mu_1 = 0$, $\sigma_1^2 = 1$, $\mu_2 = 0.5$, $\sigma_2^2 = 1$. The means are unknown but the variance is known.The problem is to decide if $\mu_1^2 + \mu_2^2 < 1$ or $1 < \mu_1^2 + \mu_2^2 < 2$. As mentioned immediately after Assumption A, we do not know if a continuous selection exists for this problem. Therefore, it is not known if our policy is asymptotically optimal for this problem. In Figs. 7(a)-(b), we plot the average delay and average cost versus $\log(L)$. The forced exploration parameter $\beta = 0.5$. The plot contains the asymptotic lower bound (dashed curve) and four other curves for γ of 0.01, 0.05, 0.2, or 1.0. Even for this setting, we observe that the slope of the empirical average delay of the proposed policy roughly matches the slope of the lower bound, and the slope of the empirical average values of cost also matches the slope of the lower bound, as expected, for small γ. The values of γ of 0.05 and 0.2 seem to good.

C. Best Arm Identification

In Fig. 8, we compare the performance of our proposed policy for best arm identification with that of the policy discussed in [9] and the lower bound. There are 3 arms and the observations are i.i.d. Gaussian with means 0, 2, and 4 and unit variance. We make the following observations.

- We observe that for both policies the slopes of the empirical average delays match the slope given by the lower bound, thereby validating the asymptotic optimality of the policies.

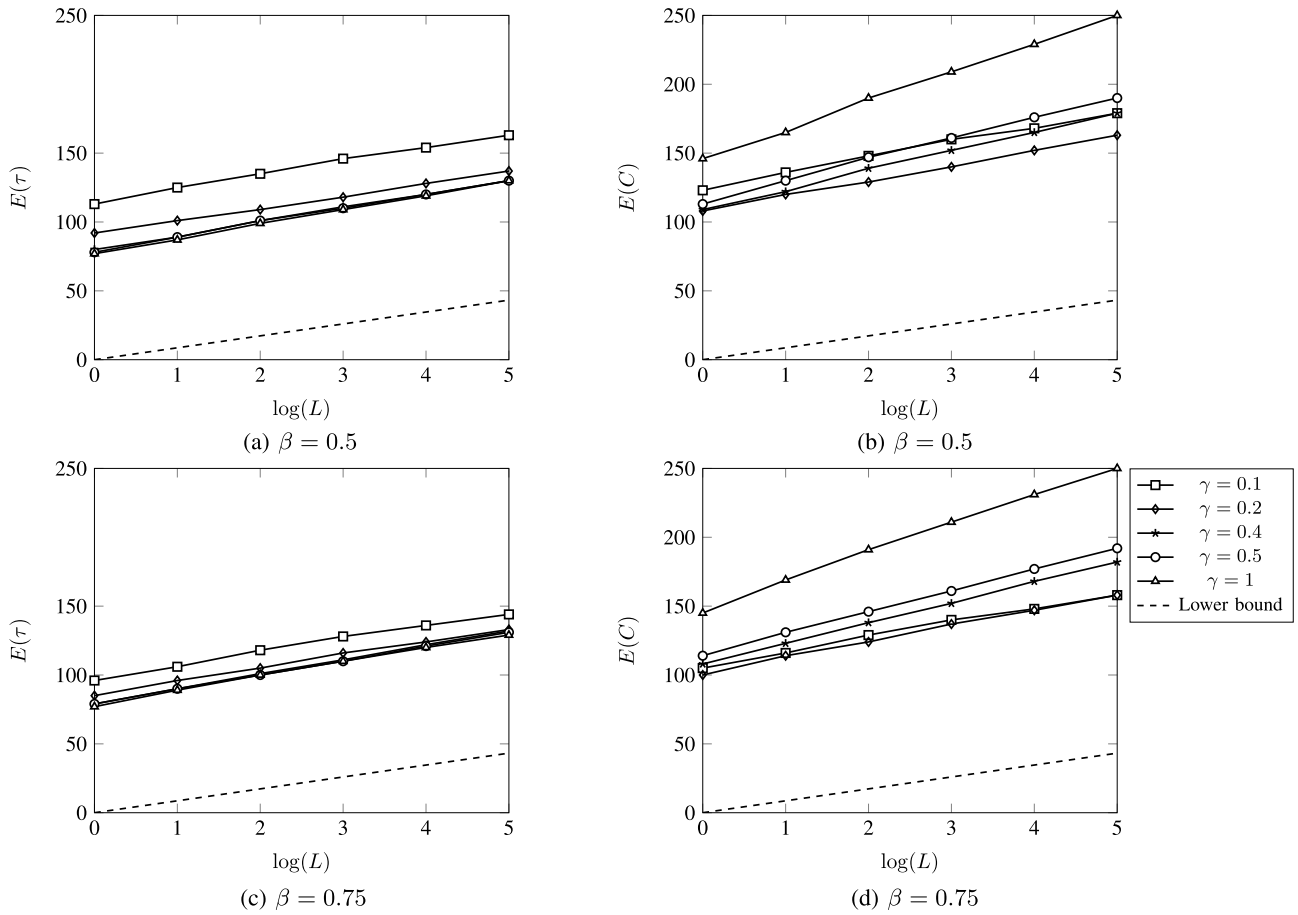

Fig. 4. Gaussian distribution with unknown means and known variances. The true parameters are $\mu_1 = 0$, $\sigma_1^2 = 1$, $\mu_2 = 1$, $\sigma_2^2 = 1$, $K = 8$, $g_{\max} = 1$ and $D^* = 0.1156$.

- While the policy in [9] is specifically designed for the best arm identification setting under single parameter exponential family observations, our policy works for more general settings as discussed in detail earlier. Therefore, for the best-arm setting, the higher average delay of our proposed policy is tolerable given the generality as long as the slopes match.

D. Dependence of D^* on K

We now study the performance for odd arm identification as a function of the number of arms K. Note that for this setting, the number of hypotheses is also K. We consider the same observation model for the odd and non-odd arms as in Section VI-A. Figs. 9(a)-(c) show the results for odd arm identificaton with unknown means and known variances among K arms, with K varying in the range 3 to 50. Fig. 9(a) shows the variation of D^* with K. We observe that with increase in number of arms K, D^* increases and then converges to the average of $D(\boldsymbol{\eta}_1||\tilde{\boldsymbol{\eta}})$ and $D(\boldsymbol{\eta}_2||\tilde{\boldsymbol{\eta}})$. This means that as the number of arms increases, identifying the odd arm becomes easier. Figs. 9(b) and 9(c) show the variation of average delay and average cost versus K for a different values of $\log(L)$. The average delay for $\log(L) = 1$ corresponds mainly to the initial exploration overhead, and increases with K. However, based on the convergence of D^* for large K, and our asymptotic result for large L, we expect this increase in delay with K to be less significant for large L. This can also be observed in the simulations as $\log(L)$ is increased to 5 and 10.

It is worth noting that the dependence of D^* on K is different for different special cases of the general problem we have addressed. For odd arm identification D^* increases with K and converges to a constant. For best arm identification, D^* decreases with K, i.e., identifying the best arm among a larger number of arms is harder. Fig. 10 shows the variation of D^* for best arm problem for Gaussian observations with unknown means and known variances among K arms, with K varying in the range 2 to 50. The means for the K arms are taken to be $1, 2, \ldots, K$, and the variance is 1 for each arm. We observe that with increase in number of arms K, D^* decreases. Therefore, the variation of D^* with K is problem instance dependent whereas our focus is on the general setting.

APPENDIX A
A REGULARITY LEMMA

In this appendix, we prove the following regularity lemma.

Lemma 8: For any i, for any compact set C,

$$\inf_{\boldsymbol{\eta}_i \in C} \|\boldsymbol{\eta}_i' - \boldsymbol{\eta}_i\| \to \infty \;\Rightarrow\; \inf_{\boldsymbol{\eta}_i \in C} D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i') \to \infty.$$

In other words, the relative entropy of (the distribution associated with) a confined parameter $\boldsymbol{\eta}_i$ with respect to

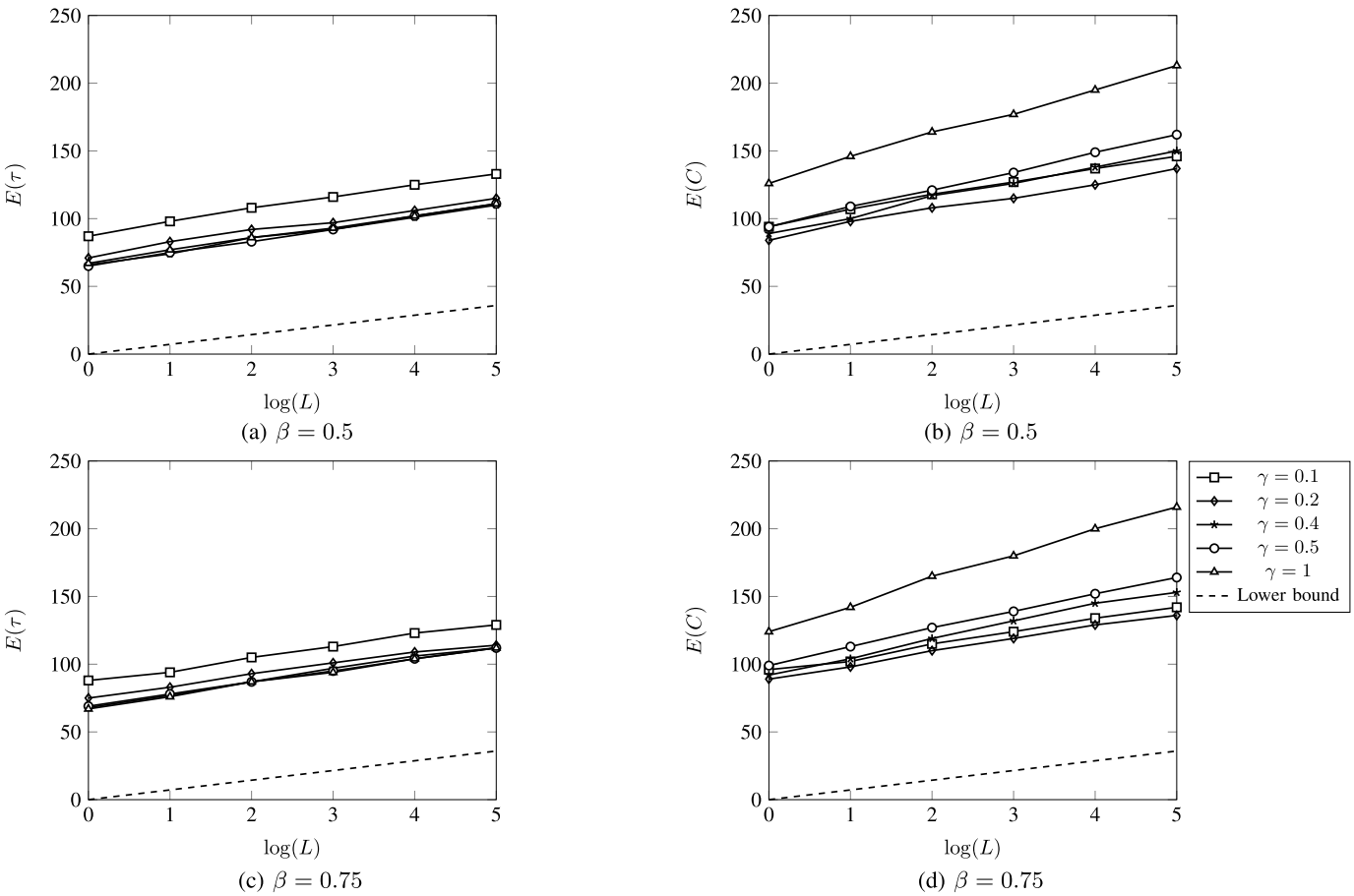

Fig. 5. Gaussian distribution with known means and unknown variances. The true parameters are $\mu_1 = 0$, $\sigma_1^2 = 5$, $\mu_2 = 0$, $\sigma_2^2 = 1$, $K = 8$, $g_{\max} = 1$ and $D^* = 0.1392$.

(the distribution associated with) a parameter $\boldsymbol{\eta}_i'$ approaches infinity as the parameter separation between $\boldsymbol{\eta}_i'$ and $\boldsymbol{\eta}_i$ grows without bound.

Proof: The lemma holds vacuously when the parameter set is bounded. Let $\mathbf{d} := \boldsymbol{\eta}_i' - \boldsymbol{\eta}_i$ and assume that $||\mathbf{d}|| > 1$. Then, from the formula (12) for relative entropy, we get

$$
\begin{aligned}
D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i') &= \frac{1}{2}(\boldsymbol{\eta}_i' - \boldsymbol{\eta}_i)^T \\
&\quad \left[\int_0^1 (1-t)\mathsf{Hess}(\mathcal{A}_i)(\boldsymbol{\eta}_i + t(\boldsymbol{\eta}_i' - \boldsymbol{\eta}_i))dt\right](\boldsymbol{\eta}_i' - \boldsymbol{\eta}_i) \\
&\geq \frac{\mathbf{d}^T}{2}\left[\int_0^{\frac{1}{||\mathbf{d}||}} (1-t)\mathsf{Hess}(\mathcal{A}_i)(\boldsymbol{\eta}_i + t(\boldsymbol{\eta}_i' - \boldsymbol{\eta}_i))dt\right]\mathbf{d} \qquad (48) \\
&\geq \frac{\|\mathbf{d}\|^2}{2}\left(1 - \frac{1}{\|\mathbf{d}\|}\right) \\
&\quad \int_0^{1/||\mathbf{d}||} \lambda_{\min}\left(\mathsf{Hess}(\mathcal{A}_i)(\boldsymbol{\eta}_i + t(\boldsymbol{\eta}_i' - \boldsymbol{\eta}_i))\right)\, dt \qquad (49) \\
&\geq \frac{1}{2}\left(\|\mathbf{d}\| - 1\right) \times \\
&\quad \min\{\lambda_{\min}\left(\mathsf{Hess}(\mathcal{A}_i)(\boldsymbol{\eta}_i + s\frac{\mathbf{d}}{\|\mathbf{d}\|})\right) : s \in [0,1]\} \qquad (50) \\
&\to \infty \text{ as } \|\mathbf{d}\| \to \infty. \qquad (51)
\end{aligned}
$$

In the above sequence of inequalities, (48) follows because of the positive definiteness of the Hessian of $\mathcal{A}_i$. Next, (49) follows from $1-t \geq 1 - 1/\|\mathbf{d}\|$ in the interval under consideration and the fact that $\mathbf{d}^T H \mathbf{d} \geq \lambda_{\min}(H)\|\mathbf{d}\|^2$ where $\lambda_{\min}(H)$ is the smallest eigenvalue of any positive definite matrix H. Finally, (51) follows because $\lambda_{\min}(\mathsf{Hess}(\mathcal{A}_i)(\cdot)$ is strictly positive in the unit neighbourhood around $\boldsymbol{\eta}_i$; indeed, $\lambda_{\min}(\mathsf{Hess}(\mathcal{A}_i)(\cdot))$ is a continuous function of $\tilde{\boldsymbol{\eta}}_i$ and therefore cannot attain the value zero on account of the strict convexity of $\mathcal{A}(\cdot)$ leading to $\lambda_{\min}(\mathcal{A}(\cdot))$ being strictly positive in the unit neighbourhood around $\boldsymbol{\eta}_i$. ■

APPENDIX B
PROOF OF PROPOSITION 3

Define

$$F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}}) := \inf_{\overline{\boldsymbol{\eta}}' \in \Theta_{-l}} \sum_{i=1}^{K} \lambda_i D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i').$$

Lemma 9: Let $(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}}) \to F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}})$ be a continuous function with $\boldsymbol{\lambda} \in \mathcal{P}(\mathcal{K})$ and $\overline{\boldsymbol{\eta}} \in \Theta_l$. Let $D^*(\overline{\boldsymbol{\eta}})$ and $\boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}})$ be defined by

$$
\begin{aligned}
D^*(\overline{\boldsymbol{\eta}}) &= \max_{\boldsymbol{\lambda} \in \mathcal{P}(\mathcal{K})} F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}}), \\
\boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}}) &= \arg\max_{\boldsymbol{\lambda} \in \mathcal{P}(\mathcal{K})} F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}}).
\end{aligned}
$$

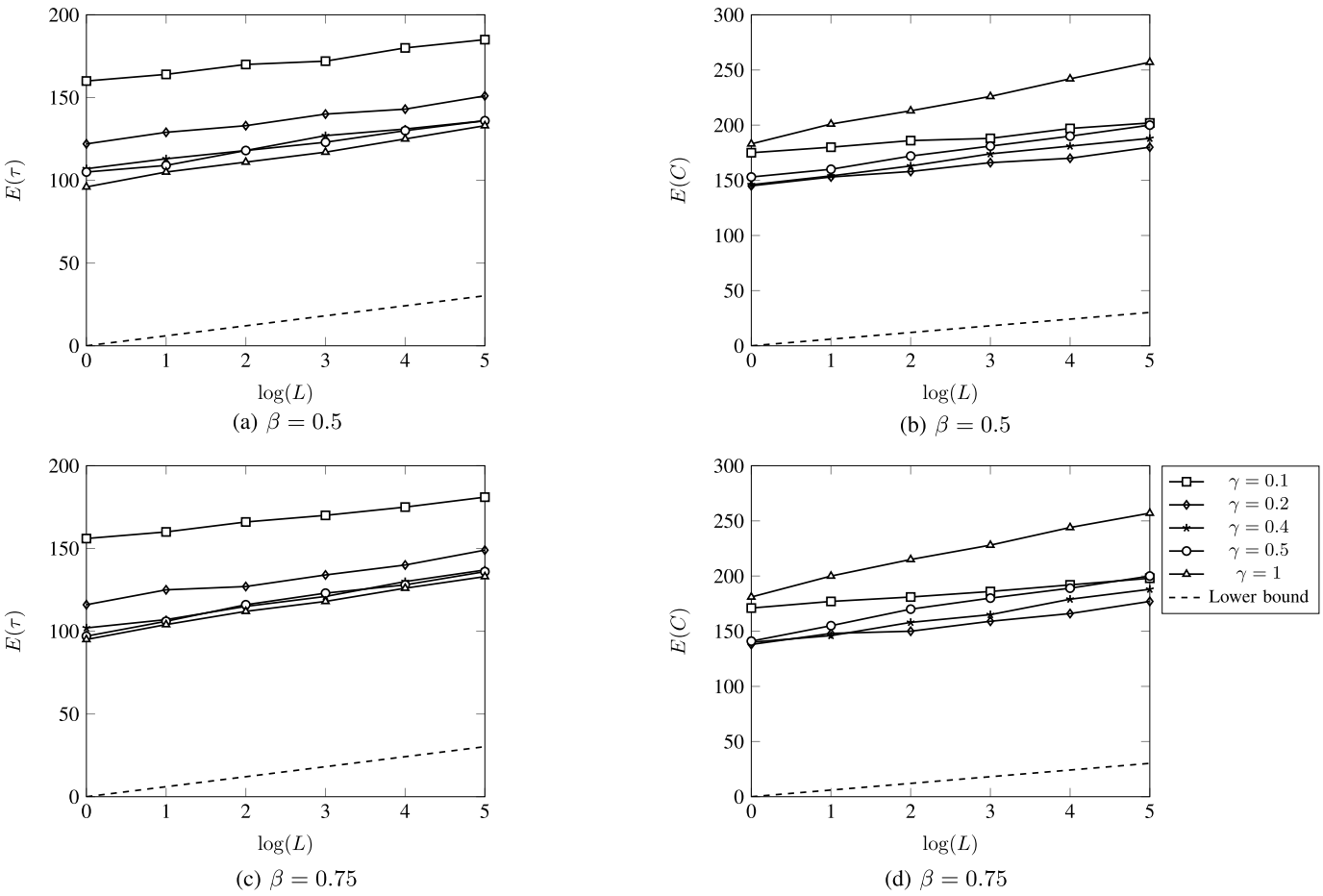

Fig. 6. Gaussian distribution with unknown means and unknown variances. The true parameters are $\mu_1 = 0$, $\sigma_1^2 = 2$, $\mu_2 = 1$, $\sigma_2^2 = 10$, $K = 8$, $g_{\max} = 1$ and $D^* = 0.1653$.

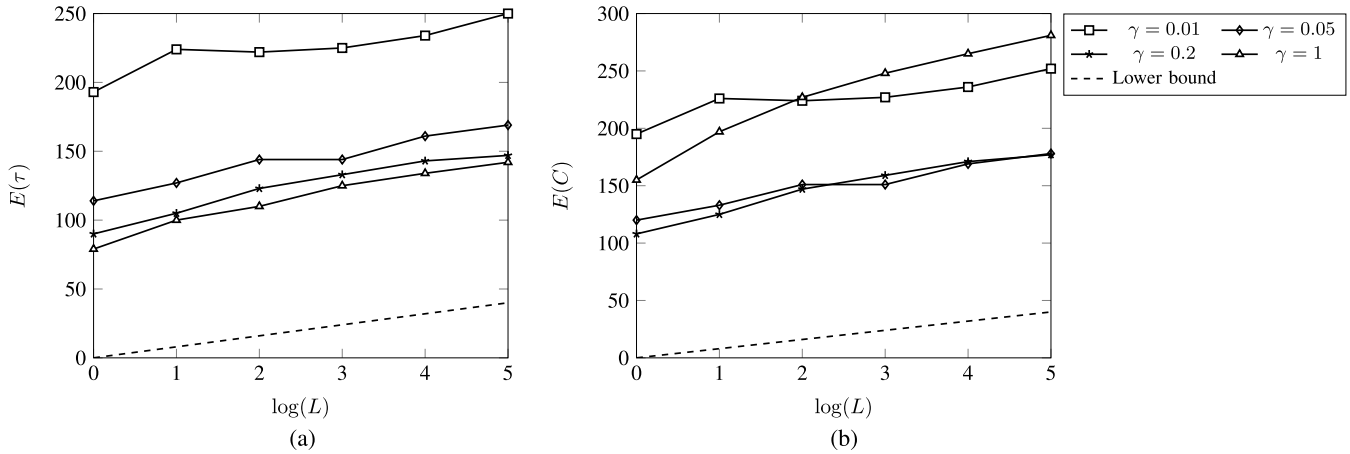

Fig. 7. Norm-threshold problem: Gaussian distribution with unknown means and known variances. The true parameters are $\mu_1 = 0$, $\sigma_1^2 = 1$, $\mu_2 = 0.5$, $\sigma_2^2 = 1$, $K = 2$, $g_{\max} = 1$, $r_1 = 1$, $r_2 = 2$ and $D^* = 0.125$.

Then $D^*(\overline{\boldsymbol{\eta}})$ is a continuous function of $\overline{\boldsymbol{\eta}}$, and $\overline{\boldsymbol{\eta}} \mapsto \boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}})$ is a compact valued, upper semicontiuous correspondence. Also, if $\boldsymbol{\lambda} \mapsto F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}})$ is concave for each $\overline{\boldsymbol{\eta}}$, then $\overline{\boldsymbol{\eta}} \mapsto \boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}})$ is a convex valued correspondence.

Proof: This result is a special case of Berge's maximum theorem [28, Theorems 9.14 and 9.17(2)] for the function $F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}})$. The proof is given in [28, Theorems 9.14 and 9.17(2)]. ■

To prove Proposition 3, it suffices show that $(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}}) \mapsto F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}})$ is continuous everywhere on its domain and that $\boldsymbol{\lambda} \mapsto F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}})$ is concave for each $\overline{\boldsymbol{\eta}}$. Then by Lemma 9 the set $\boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}})$ where the maximum is attained is nonempty, and the set-valued map $\overline{\boldsymbol{\eta}} \mapsto \boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}})$ is upper semi-continuous, compact-valued and convex-valued.

Fix $\overline{\boldsymbol{\eta}} \in \Theta_l$. Since $\boldsymbol{\lambda} \mapsto F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}})$ is an infimum of linear functions parameterised by $\overline{\boldsymbol{\eta}}'$, concavity immediately follows.

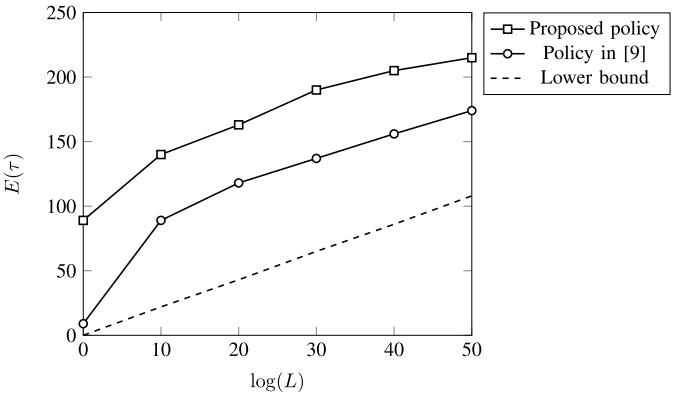

Fig. 8. Best arm identification: Gaussian distribution with unknown means and known variances. The true parameters are $K = 3$, $\mu = \{0, 2, 4\}$, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$.

We now proceed to show that $(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}}) \mapsto F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}})$ is continuous. On account of concavity, the issue of continuity arises only at a boundary. Nevertheless, we give a general proof.

(i) Fix an arbitrary $\varepsilon > 0$. Fix $\overline{\boldsymbol{\eta}}' \in \Theta_{-l}$ so that

$$\sum_i \lambda_i D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i') \leq F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}}) + \varepsilon. \tag{52}$$

Observe that, for each i,

$$\overline{\boldsymbol{\eta}} \mapsto D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i') = (\boldsymbol{\eta}_i - \boldsymbol{\eta}_i')^T \boldsymbol{\kappa}_i(\boldsymbol{\eta}_i) - \mathcal{A}_i(\boldsymbol{\eta}_i) + \mathcal{A}_i(\boldsymbol{\eta}_i')$$

is a continuous function of $\overline{\boldsymbol{\eta}} \in \Theta_l$ by virtue of the continuity of $\boldsymbol{\kappa}_i(\cdot)$ and the continuously differentiable property of $\mathcal{A}_i(\cdot)$. (Indeed, the continuity of $\boldsymbol{\kappa}_i(\cdot)$ itself follows from the continuous differentiability of $\mathcal{A}_i(\cdot)$.) Now consider a sequence $(\boldsymbol{\lambda}(n), \overline{\boldsymbol{\eta}}(n)) \to (\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}}) \in \mathcal{P}(K) \times \Theta_l$ as $n \to \infty$. We then have, for all sufficiently large n,

$$\boldsymbol{\lambda}(n) \leq \boldsymbol{\lambda} + \varepsilon \mathbf{1} \text{ and } D(\boldsymbol{\eta}_i(n) \parallel \boldsymbol{\eta}_i') \leq D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i') + \varepsilon \ \forall i, \tag{53}$$

where the first inequality is to be taken componentwise with $\mathbf{1}$ being the all-1 vector. Hence, for all sufficiently large n, we have

$$\begin{aligned} & F(\boldsymbol{\lambda}(n), \overline{\boldsymbol{\eta}}(n)) \\ & \quad \leq \sum_i \lambda_i(n) D(\boldsymbol{\eta}_i(n) \parallel \boldsymbol{\eta}_i') \\ & \qquad \text{(left-side is an infimum, for each fixed } n) \\ & \quad \leq \sum_i (\lambda_i + \varepsilon)(D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i') + \varepsilon) \\ & \quad = \sum_i \lambda_i D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i') \\ & \qquad + \varepsilon \left(1 + \sum_i D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i')\right) + \varepsilon^2 K \\ & \quad \leq F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}}) + \varepsilon \left(2 + \sum_i D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i') + \varepsilon K\right), \end{aligned}$$

where the last inequality follows from (52). Since $\sum_i D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i')$ is finite for any $\overline{\boldsymbol{\eta}} \in \Theta_l$ and $\overline{\boldsymbol{\eta}}' \in \Theta_{-l}$, and since ε was arbitrary, we get

$$\limsup_{n \to \infty} F(\boldsymbol{\lambda}(n), \overline{\boldsymbol{\eta}}(n)) \leq F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}}). \tag{54}$$

(ii) Once again fix $\varepsilon > 0$, and consider a sequence $(\boldsymbol{\lambda}(n), \overline{\boldsymbol{\eta}}(n)) \to (\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}}) \in \mathcal{P}(K) \times \Theta_l$ as $n \to \infty$, but this time choose a convergent sequence[2] $\overline{\boldsymbol{\eta}}'(n) \in \Theta_{-l}$, such that for every n, we have:

$$F(\boldsymbol{\lambda}(n), \overline{\boldsymbol{\eta}}(n)) \geq \sum_i \lambda_i(n) D(\boldsymbol{\eta}_i(n) \parallel \boldsymbol{\eta}_i'(n)) - \varepsilon, \tag{55}$$

and $\overline{\boldsymbol{\eta}}'(n) \to \overline{\boldsymbol{\eta}}'$. Analogous to (53), using the boundedness of the sequence $\boldsymbol{\eta}'(n)$, for all sufficiently large n, we have

$$\begin{aligned} & \boldsymbol{\lambda}(n) \geq \boldsymbol{\lambda} - \varepsilon \mathbf{1} \text{ and} \\ & D(\boldsymbol{\eta}_i(n) \parallel \boldsymbol{\eta}_i'(n)) \geq D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i'(n)) - \varepsilon \ \forall i. \end{aligned} \tag{56}$$

Using (56) in (55), we get

$$\begin{aligned} & F(\boldsymbol{\lambda}(n), \overline{\boldsymbol{\eta}}(n)) \\ & \quad \geq \sum_i \lambda_i D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i'(n)) \\ & \qquad - \varepsilon \left(2 + \sum_i D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i'(n)) - \varepsilon K\right) \\ & \quad \geq F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}}) - \varepsilon \left(2 + \sum_i D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i'(n)) - \varepsilon K\right), \end{aligned}$$

where the last inequality follows from the observation that $\overline{\boldsymbol{\eta}}'(n) \in \Theta_{-l}$ for all n and by then taking the infimum over all $\overline{\boldsymbol{\eta}}' \in \Theta_{-l}$. Since $\overline{\boldsymbol{\eta}}'(n) \to \overline{\boldsymbol{\eta}}'$, the quantity $\sum_i D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i'(n))$ converges to $\sum_i D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i')$ and is therefore bounded. Since ε was arbitrary, we obtain

$$\liminf_{n \to \infty} F(\boldsymbol{\lambda}(n), \overline{\boldsymbol{\eta}}(n)) \geq F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}}). \tag{57}$$

From (54) and (57), we have the continuity of $F(\boldsymbol{\lambda}, \overline{\boldsymbol{\eta}})$.

APPENDIX C
FINITE STOPPING TIME

We show in a series of steps that the proposed policy stops in finite time. We begin with the proof of ML estimates of the parameters converging to the true parameter values and then show that the statistic grows as $\Omega(n^\beta)$ as time $n \to \infty$. We then show that, with this growth, the statistic crosses any fixed threshold in finite time and, hence, the stopping time is finite almost surely.

Before we begin with the proof, as done in [3], we also consider two variants of $\pi_{SMF}(L, \gamma, \beta)$ which are useful in the analysis.

[2]Observe that if, for some i, $\boldsymbol{\eta}_i'(n) \to \infty$, since $\boldsymbol{\eta}_i(n)$ is confined to a compact neighbourhood of $\boldsymbol{\eta}_i$, by Lemma 8, we must have $D(\boldsymbol{\eta}_i(n) \parallel \boldsymbol{\eta}_i'(n)) \to \infty$. On account of (55), we can replace the offending $\boldsymbol{\eta}_i'(n)$ with another bounded quantity yielding a bounded $D(\boldsymbol{\eta}_i(n) \parallel \boldsymbol{\eta}_i'(n))$, without affecting inequality (55). Hence, without loss of generality, we may take that $\overline{\boldsymbol{\eta}}'(n)$ is bounded. Furthermore, by passing to a subsequence if necessary, we may further take $\overline{\boldsymbol{\eta}}'(n) \to \overline{\boldsymbol{\eta}}'$ for some limit $\overline{\boldsymbol{\eta}}'$.

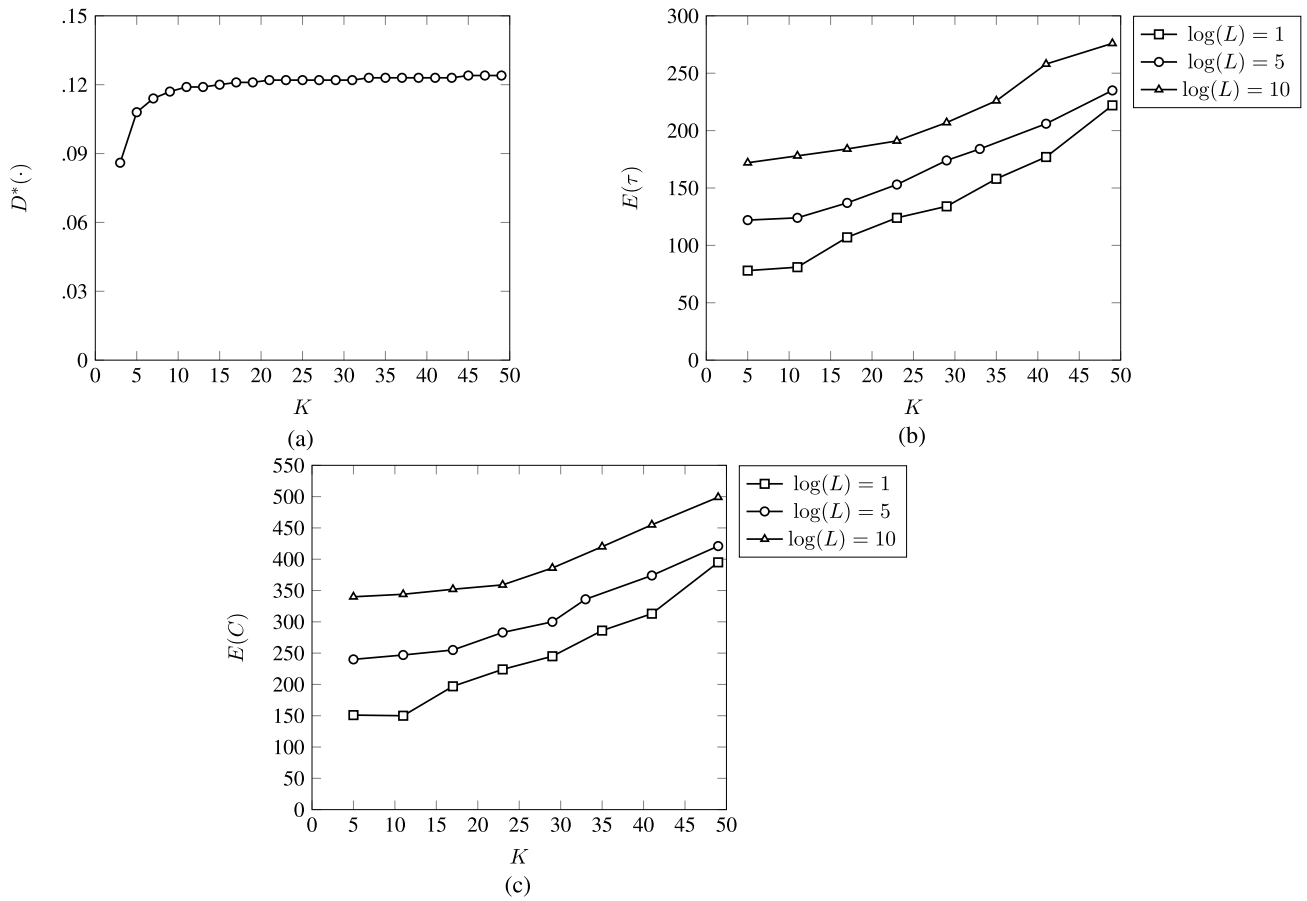

Fig. 9. Odd arm Identification: Gaussian distribution with unknown means and known variances. The true parameters are $\mu_1 = 0$, $\sigma_1^2 = 1$, $\mu_2 = 1$, $\sigma_2^2 = 1$, $\gamma = 1$ and $g_{max} = 1$.

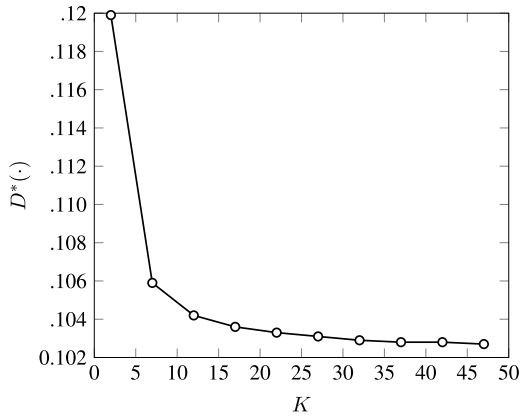

Fig. 10. Best arm Identification: Gaussian distribution with unknown means and known variances. μ_i's are chosen from the set $\{1, 2, \ldots, K\}$ and ${\sigma_i}^2 = 1$ for $i = 1, 2, \ldots, K$.

1) *Policy* $\pi^l_{SMF}(L, \gamma, \beta)$ is like policy $\pi_{SMF}(L, \gamma, \beta)$ but stops only at decision l, when $Z_l(n) \geq \log((M-1)L)$.
2) *Policy* $\tilde{\pi}_{SMF}(\gamma, \beta)$ is also like $\pi_{SMF}(L, \gamma, \beta)$ but never stops. So the policy does not depend on the policy parameter L.

The policy $\tilde{\pi}_{SMF}(\gamma, \beta)$ will be used in Proposition 13 and the policy $\pi^l_{SMF}(L, \gamma, \beta)$ will be used in the proof of Proposition 5.

A. Convergence Results

Let us begin with a Lemma from [9].

Lemma 10: [9, Lemma 17] Let K be a positive integer and let Σ_K be a simplex of dimension $K-1$. Let $g : \mathbb{N} \to \mathbb{R}$ be a non-decreasing function such that $g(0) = 0$, $g(n)/n \to 0$ as $n \to \infty$ and for all $k \geq 1$ and $\forall m \geq 1$,

$$\inf\{k \in \mathbb{N} : g(k) \geq m\} > \inf\{k \in \mathbb{N} : g(k) \geq m-1\} + K.$$

Let $\hat{\boldsymbol{\lambda}}(k)$ be a sequence of elements in Σ_K such that there exists $\boldsymbol{\lambda}^* \in \Sigma_K$, there exists $\epsilon > 0$ and an integer $n_0(\epsilon)$ such that

$$\forall n \geq n_0, \ \sup_{1 \leq i \leq K} |\hat{\boldsymbol{\lambda}}_i(k) - \boldsymbol{\lambda}^*_i| \leq \epsilon.$$

Define $N^{n,a}(0) = 0$, and for every $k \in \{0, \ldots, n-1\}$, $U_k = \{i : N_i^{n,a}(k) < g(k)\}$ and

$$I_{k+1} \in \begin{cases} \underset{i \in U_k}{\operatorname{argmin}} N_i^{n,a} & \text{if } U_k \neq \Phi \\ \underset{i \in \{1,2,\ldots,K\}}{\operatorname{argmax}} \{n^a \lambda^*_i(\overline{\boldsymbol{\eta}}^*(l^*(n)) - N_i^{n,a}\} & \text{otherwise,} \end{cases}$$

and for all i, $N_i^{n,a}(k+1) = N_i^{n,a}(k) + \mathbb{1}_{(I_{k+1}=i)}$. Then for all $i \in \{1, 2, \ldots, K\}$, $N_i^{n,a}(n) > g(n) - 1$ and there exists $n_1 \geq n_0$ (that depends on ϵ) such that for all $n \geq n_1$,

$$\max_{1 \leq i \leq K} \left| \frac{N_i^{n,a}}{n^a} - \boldsymbol{\lambda}^*_i \right| \leq 3(K-1)\epsilon.$$

Proof: This Lemma is the same as [9, Lemma 17] and is included here for easy reference. See [9, Appendix B.2] for proof. ■

Next, we show a preliminary result about the forced exploration rule.

Lemma 11: Let $\overline{\boldsymbol{\eta}} \in \Theta_l$ be the true configuration. The proposed sampling rule ensures that

$$N_i^{n,a} \geq [(n^a)^\beta - (\beta(K+1))^{\beta/(1-\beta)}]_+ - 1.$$

Furthermore, for all $\epsilon > 0$ and for all n_0, setting $n_\epsilon = \max\{n_0, \epsilon^{-1}\}/(3\epsilon)$, we have the implication:

$$\sup_{n^a \geq n_0} \max_i |\boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}}^*(l)) - \boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}})| \leq \epsilon$$
$$\Rightarrow \sup_{n^a \geq n_\epsilon} \max_i \left| \frac{N_i^{n,a}}{n^a} - \lambda_i^*(\overline{\boldsymbol{\eta}}) \right| \leq 3(K-1)\epsilon. \quad (58)$$

Proof: The proof follows from Lemma 10. Specifically, we need to check that, if $g(n) = [n^\beta - (\beta(K+1))^{\beta/(1-\beta)}]_+$, then $g(0) = 0$, $g(n)/n \to 0$ as $n \to \infty$, and for every $m \geq 1$

$$\inf\{k \in \mathbb{N} : g(k) \geq m\} > \inf\{k \in \mathbb{N} : g(k) \geq m-1\} + K.$$

The first two conditions are straightforward. To check the third condition, observe that $\inf\{k \in \mathbb{N} : g(k) \geq m\} = \lceil [m + (\beta(K+1))^{\beta/(1-\beta)}]^{1/\beta} \rceil$. Thus, with $u = m - 1 + (\beta(K+1))^{\beta/(1-\beta)}$, we have

$$\begin{aligned} \inf\{k \in \mathbb{N} : g(k) \geq m\} &- \inf\{k \in \mathbb{N} : g(k) \geq m-1\} \\ &\geq (u+1)^{1/\beta} - \lceil u^{1/\beta} \rceil \\ &> \frac{1}{\beta} u^{(1-\beta)/\beta} - 1 \\ &\geq \frac{1}{\beta} u_{\min}^{(1-\beta)/\beta} - 1 \\ &= K \end{aligned}$$

where the strict inequality follows from strict convexity of the function $u^{1/\beta}$, the following inequality follows from monotonicity of $u^{(1-\beta)/\beta}$, and the last equality follows from the observation that $u_{\min} = (\beta(K+1))^{\beta/(1-\beta)}$ (obtained when $m = 1$). Since the conditions for [9, Lemma 17] hold, the rest of the proof follows [9, Appendix B.2], whose examination indicates that we may take $n_\epsilon = \max\{n_0, \epsilon^{-1}\}/(3\epsilon)$. ■

We next establish a concentration lemma. The notation $\mathbf{a} \succ \mathbf{b}$ stands for component-wise strict inequality.

Lemma 12: Let $\boldsymbol{\epsilon} \in \mathbb{R}^d$ be made of strictly positive entries. Then there exists a finite positive constant C such that

$$P\left(\left| \frac{\mathbf{Y}_i^n}{N_i^n} - \boldsymbol{\kappa}_i \right| \succ \boldsymbol{\epsilon} \right) \leq \frac{C}{n^4}.$$

Proof: Thanks to the union bound, it suffices to focus on one component, say l where $1 \leq l \leq d$. Let the lth components be represented as $Y_i^n(l)$, $\kappa_i(l)$, and ϵ_l. We will show that

$$P\left(\left| \frac{Y_i^n(l)}{N_i^n} - \kappa_i(l) \right| > \epsilon_l \right) \leq \frac{C_l}{n^4}$$

for some finite positive constant C_l.

Fix l. Observe that $M_n := Y_i^n(l) - \kappa_i(l) N_i^n$ is a martingale whose quadratic variation process $\langle M \rangle_n$ has the following property:

$$\sum_{t=1}^n E\left[((Y_{i,t}(l) - \kappa_i(l)) 1\{a_t = i\})^2 \mid \mathbf{Y}_i^{t-1}, A^{t-1} \right] \leq n\sigma^2(l),$$

a consequence of the existence of the second central moment for the observations. Hence, we have the following inequalities for all sufficiently n:

$$\begin{aligned} P&\left(\left| \frac{Y_i^n(l)}{N_i^n} - \kappa_i(l) \right| > \epsilon_l \right) \\ &\leq P\Big(|Y_i^n(l) - \kappa_i(l) N_i^n| > N_i^n \epsilon_l \Big) \\ &\leq P\Big(|Y_i^n(l) - \kappa_i(l) N_i^n| > (n^a)^\beta \epsilon_l/2 \Big) \\ &\quad \text{(since } N_i^n \geq N_i^{n,a} \geq (n^a)^\beta/2 \text{ for all suff. large } n) \\ &\leq P\Big(|Y_i^n(l) - \kappa_i(l) N_i^n| > (\gamma n/2)^\beta \epsilon_l/2, \quad n^a \geq \gamma n/2 \Big) \\ &\quad + P\Big(n^a < \gamma n/2 \Big) \\ &\leq P\Big(\sup_{1 \leq t \leq n} |Y_i^t(l) - \kappa_i(l) N_i^t| > n^\beta (\gamma/2)^\beta \epsilon_l/2 \Big) \\ &\quad + P\Big(n^a - \gamma n < -\gamma n/2 \Big) \\ &\leq \frac{E[\big(\sup_{1 \leq t \leq n} |Y_i^t(l) - \kappa_i(l) N_i^t|\big)^p]}{(\epsilon_l/2)^p (\gamma/2)^{\beta p} n^{\beta p}} \quad \text{(Markov inequality)} \\ &\quad + e^{-n\gamma^2/16} \quad \text{(Bernstein inequality)} \\ &\leq \frac{c_p}{(\epsilon_l/2)^p (\gamma/2)^{\beta p} n^{\beta p}} E[|\langle M \rangle_n|^{p/2}] + e^{-n\gamma^2/16} \\ &\quad \text{(Burkholder inequality)} \\ &\leq \frac{c_p}{(\epsilon_l/2)^p (\gamma/2)^{\beta p} n^{\beta p}} \cdot \sigma^p(l) n^{p/2} + e^{-n\gamma^2/16} \\ &\quad \text{(from quadratic variation bound)} \\ &\leq \frac{C_l}{n^4} \quad \text{(taking } p = 4/(\beta - 1/2) \text{ and choosing } C_l \text{ suitably);} \end{aligned}$$

recall that in the inequality above where the Burkholder inequality [29, p.414] is employed with $p = 4/(\beta - 1/2)$, we made use of finiteness of the variances of the observations. This establishes the lemma. Choosing a suitably larger C_l if needed, we can make the probability inequality be upper bounded by C_l/n^4 for all n. ■

Proposition 13: Let $\overline{\boldsymbol{\eta}} \in \Theta_l$ be the true configuration. Consider the non-stopping policy $\tilde{\pi}_{SMF}(\gamma, \beta)$. Then, the following convergences hold almost surely as $n \to \infty$:

$$\hat{\boldsymbol{\kappa}}_i := \frac{\mathbf{Y}_i^n}{N_i^n} \to \boldsymbol{\kappa}_i \text{ for all } i, \quad (59)$$
$$\hat{\boldsymbol{\eta}}_i \to \boldsymbol{\eta}_i \text{ for all } i, \quad (60)$$
$$\boldsymbol{\eta}_i^*(l) \to \boldsymbol{\eta}_i \text{ for all } i, \quad (61)$$
$$\liminf_{n \to \infty} \frac{Z_{lm}(n)}{n^\beta} > 0. \quad (62)$$

Proof: We prove the statements one after another.

(i) Proof of (59): This follows from Lemma 12 and the Borel-Cantelli lemma, since the series involving the upper bound in Lemma 12 is summable.

(ii) Proof for (60): follows from (59) and the continuity of the function $\boldsymbol{\kappa} \mapsto \boldsymbol{\eta}(\boldsymbol{\kappa})$.

(iii) Proof for (61): Since $\overline{\boldsymbol{\eta}} \in \Theta_l$, an open set, by (60), $\hat{\overline{\boldsymbol{\eta}}} \in \Theta_l$ for all sufficiently large n. From (32), it follows $\overline{\boldsymbol{\eta}}^*(l) = \hat{\overline{\boldsymbol{\eta}}}$ for all sufficiently large n, and hence (60) implies (61).

(iv) Proof for (62): Consider the expression for $Z_{lm}(n)$ in (41). See (63)-(66), as shown at the bottom of the page. Note that $\boldsymbol{\eta}_i(\boldsymbol{\kappa}_i)$ optimises the function $\boldsymbol{\eta}_i' \mapsto (\boldsymbol{\eta}')^T\boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i')$, for each $i = 1, \ldots, K$, over the set Θ_l, since $\overline{\boldsymbol{\eta}} \in \Theta_l$. We now leverage this.

Write $B_r(\overline{\boldsymbol{\eta}}(\overline{\boldsymbol{\kappa}}))$ for the open Euclidean ball of radius r around $\overline{\boldsymbol{\eta}}(\overline{\boldsymbol{\kappa}})$. Fix $\epsilon > 0$. There is then an $\delta > 0$ and a $C_\delta > 0$ such that, almost surely, for all sufficiently large n and all $\overline{\boldsymbol{\eta}}' \in B_\delta(\overline{\boldsymbol{\eta}}(\overline{\boldsymbol{\kappa}}))$, we have:

$$||\hat{\boldsymbol{\kappa}}_i - \boldsymbol{\kappa}_i||_\infty \leq \varepsilon$$

and since $\boldsymbol{\eta}_i^*(m)$ is bounded (see footnote 2),

$$\begin{aligned}
\left|\left(\left(\boldsymbol{\eta}_i' - \boldsymbol{\eta}_i^*(m)\right)^T(\hat{\boldsymbol{\kappa}}_i - \boldsymbol{\kappa}_i) + (\boldsymbol{\eta}_i')^T\frac{\boldsymbol{\Upsilon}_i}{N_i^n}\right)\right| &< C_\delta\varepsilon \\
|n_{0i}\mathcal{A}_i(\boldsymbol{\eta}_i')| &\leq C_\delta \\
\left|(\boldsymbol{\eta}_i')^T\boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i') - (\boldsymbol{\eta}_i(\boldsymbol{\kappa}_i))^T\boldsymbol{\kappa}_i + \mathcal{A}_i(\boldsymbol{\eta}_i(\boldsymbol{\kappa}_i))\right| &\leq \varepsilon.
\end{aligned}$$

Furthermore, we can lower bound the integral in (66) by restricting the integral to the set $\Theta_l \cap B_\delta(\overline{\boldsymbol{\eta}}(\overline{\boldsymbol{\kappa}}))$. Putting these ideas together, we get that (66) is lower bounded as in (67)-(70), as shown at the bottom of the next page, where the last inequality in (70) holds because the Lebesgue measure $\mathsf{Leb}\left(\overline{\boldsymbol{\eta}}' \in \Theta_l \cap B_\delta(\overline{\boldsymbol{\eta}}(\overline{\boldsymbol{\kappa}}))\right) > 0$. Continuing with the inequality in (70), by virtue of our sampling rule and by choosing ε sufficiently small, almost surely, for some constant $a > 0$, the lower bound becomes

$$\begin{aligned}
&\geq a\left(\inf_{\boldsymbol{\eta}' \in \Theta_m}\sum_{i=1}^K D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i') - K(1 + C_\delta)\varepsilon\right) &(71)\\
&> 0, &(72)
\end{aligned}$$

where the last strict inequality holds because

$$\inf_{\overline{\boldsymbol{\eta}}' \in \Theta_m}\sum_{i=1}^K D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i') > 0, \quad \forall m \neq l,$$

a fact that comes from the assumptions that Θ_l and Θ_m are disjoint and open. This completes the proof of the Proposition. ■

B. Proof of Proposition 5

Proof: The following inequalities hold almost surely:

$$\begin{aligned}
&\tau(\pi_{SMF}(L, \gamma, \beta)) \\
&\quad \leq \tau(\pi^l_{SMF}(L, \gamma, \beta)) \\
&\quad = \inf\{n \geq 1 | Z_l(n) > \log((M-1)L)\} \\
&\quad \leq \inf\{n \geq 1 | Z_{lm}(n') > \log((M-1)L), \forall n' \geq n, \forall m \neq l\} \\
&\quad < \infty,
\end{aligned}$$

where the last inequality follows because of (62) in Proposition 13. ■

$$\begin{aligned}
\liminf_{n\to\infty}\frac{Z_{lm}(n)}{n^\beta} &= \liminf_{n\to\infty}\left(\frac{1}{n^\beta}\log\mathcal{H}_l(\overline{\boldsymbol{\Upsilon}}, \mathbf{n_0}) + \frac{1}{n^\beta}\log\left\{\int_{\overline{\boldsymbol{\eta}}' \in \Theta_l}\exp\left\{\sum_{i=1}^K\left[(\boldsymbol{\eta}_i')^T(\mathbf{Y}_i^n + \boldsymbol{\Upsilon}_i) - (N_i^n + n_{0i})\mathcal{A}_i(\boldsymbol{\eta}_i')\right]\right\}d\overline{\boldsymbol{\eta}}'\right\}\right. \\
&\qquad \left. - \frac{1}{n^\beta}\cdot n\sum_{i=1}^K w_i\left\{\boldsymbol{\eta}_i^{*T}(m)\hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\boldsymbol{\eta}_i^*(m))\right\}\right) &(63)\\
&\geq \liminf_{n\to\infty}\left(\frac{1}{n^\beta}\log\left\{\int_{\overline{\boldsymbol{\eta}}' \in \Theta_l}\exp\left\{\sum_{i=1}^K\left[(\boldsymbol{\eta}_i')^T(\mathbf{Y}_i^n + \boldsymbol{\Upsilon}_i) - (N_i^n + n_{0i})\mathcal{A}_i(\boldsymbol{\eta}_i')\right]\right\}d\overline{\boldsymbol{\eta}}'\right\}\right. \\
&\qquad \left. - \frac{1}{n^\beta}\cdot n\sum_{i=1}^K w_i\left\{\boldsymbol{\eta}_i^{*T}(m)\hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\boldsymbol{\eta}_i^*(m))\right\}\right) \text{(since the first term in (63) is inconsequential)} &(64)\\
&= \liminf_{n\to\infty}\left(\frac{1}{n^\beta}\log\int_{\overline{\boldsymbol{\eta}}' \in \Theta_l}\exp\left\{n\sum_{i=1}^K\left[\frac{N_i^n}{n}(\boldsymbol{\eta}_i')^T\left(\frac{\mathbf{Y}_i^n}{N_i^n} + \frac{\boldsymbol{\Upsilon}_i}{N_i^n}\right) - \frac{(N_i^n + n_{0i})}{n}\mathcal{A}_i(\boldsymbol{\eta}_i')\right]\right\}d\overline{\boldsymbol{\eta}}'\right. \\
&\qquad \left. - \frac{1}{n^\beta}\log\exp\left\{n\sum_{i=1}^K\frac{N_i^n}{n}\left\{(\boldsymbol{\eta}_i^*(m))^T\hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\boldsymbol{\eta}_i^*(m))\right\}\right\}\right) &(65)\\
&= \liminf_{n\to\infty}\frac{1}{n^\beta}\log\int_{\overline{\boldsymbol{\eta}}' \in \Theta_l}\exp\left\{n\sum_{i=1}^K\frac{N_i^n}{n}\left((\boldsymbol{\eta}_i' - \boldsymbol{\eta}_i^*(m))^T\boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i') + \mathcal{A}_i(\boldsymbol{\eta}_i^*(m))\right)\right. \\
&\qquad \left. + n\sum_{i=1}^K\left[\frac{N_i^n}{n}\left(\left(\boldsymbol{\eta}_i' - \boldsymbol{\eta}_i^*(m)\right)^T(\hat{\boldsymbol{\kappa}}_i - \boldsymbol{\kappa}_i) + (\boldsymbol{\eta}_i')^T\frac{\boldsymbol{\Upsilon}_i}{N_i^n}\right) - \frac{n_{0i}}{n}\mathcal{A}_i(\boldsymbol{\eta}_i')\right\}\right]d\overline{\boldsymbol{\eta}}'. &(66)
\end{aligned}$$

APPENDIX D
PROOF OF PROPOSITION 6 ON ADMISSIBILITY

Proof: Fix $\overline{\boldsymbol{\eta}} \in \Theta_l$ as the true configuration. We begin with

$$\begin{aligned} & P(\delta \neq l|\overline{\boldsymbol{\eta}}) \\ &= \sum_{m\neq l} P(\delta = m|\overline{\boldsymbol{\eta}}) + P(\tau(\pi_{SMF}(L,\gamma,\beta)) = \infty|\overline{\boldsymbol{\eta}}) \\ &= \sum_{m\neq l} P(\delta = m|\overline{\boldsymbol{\eta}}), \end{aligned} \tag{73}$$

where (73) follows from Proposition 5. Let $\Delta_m^n =$

$$\{(x^n,a^n) : \tau(\pi_{SMF}(L,\gamma,\beta)(x^n,a^n) = n, \delta(x^n,a^n) = m\}$$

denote the sample paths for which the decision maker stops sampling after n time slots and decides in favour of $H = m$. The decision region in favour of m is denoted $\Delta_m := \bigcup_{n\geq 1} \Delta_m^n$. Note that

$$\Delta_m^n \cap \Delta_m^k = \emptyset \text{ for all } k \neq n. \tag{74}$$

We then have

$$\begin{aligned} & P(\delta \neq l|\overline{\boldsymbol{\eta}}) \\ &= \sum_{m\neq l} P(\delta = m|\overline{\boldsymbol{\eta}}) && (75) \\ &= \sum_{m\neq l}\sum_{n\geq 1} \int_{(x^n,a^n)\in\Delta_m^n} dP((x^n,a^n)|\overline{\boldsymbol{\eta}}) && (76) \\ &= \sum_{m\neq l}\sum_{n\geq 1} \int_{(x^n,a^n)\in\Delta_m^n} \prod_{t=1}^n \Big[P\left(a_t|a^{t-1},x^{t-1}\right) \\ &\qquad \cdot f(x_t|a_t,\boldsymbol{\eta}_{a_t})\Big] d(x^n,a^n) && (77) \\ &= \sum_{m\neq l}\sum_{n\geq 1} \int_{(x^n,a^n)\in\Delta_m^n} \prod_{t=1}^n f(x_t|a_t,\boldsymbol{\eta}_{a_t}) \\ &\qquad \cdot \left[\prod_{t=1}^n P\left(a_t|a^{t-1},x^{t-1}\right)\right] d(x^n,a^n) && (78) \\ &\leq \sum_{m\neq l}\sum_{n\geq 1} \int_{(x^n,a^n)\in\Delta_m^n} \frac{\hat{f}(x^n|a^n,\tilde{\overline{\boldsymbol{\eta}}}\in\Theta_l)}{\tilde{f}_m(x^n|a^n)} \tilde{f}_m(x^n|a^n) \\ &\qquad \cdot \left[\prod_{t=1}^n P\left(a_t|a^{t-1},x^{t-1}\right)\right] d(x^n,a^n) && (79) \\ &\leq \sum_{m\neq l} \frac{1}{(M-1)L} \sum_{n\geq 1} \int_{(x^n,a^n)\in\Delta_m^n} \tilde{f}_m(x^n|a^n) \\ &\qquad \prod_{t=1}^n P\left(a_t|a^{t-1},x^{t-1}\right) d(x^n,a^n) && (80) \\ &\leq \sum_{m\neq l} \frac{1}{(M-1)L} \cdot \tilde{P}(\delta = m|\tilde{\overline{\boldsymbol{\eta}}}\in\Theta_m) && (81) \\ &\leq \frac{1}{L}. && (82) \end{aligned}$$

In (77), the term $P\left(a_t|a^{t-1},x^{t-1}\right)$ indicates the probability of choosing arm a_t, with the convention that at time $t = 1$, the term represents $P(a_1)$. Inequality (79) follows from the definition of maximum likelihood function, in particular $\prod_{t=1}^n f(x_t|a_t,\tilde{\boldsymbol{\eta}}_{a_t}) = f(x^n|a^n,\tilde{\overline{\boldsymbol{\eta}}}) \leq \hat{f}(x^n|a^n,\tilde{\overline{\boldsymbol{\eta}}}\in\Theta_l)$. In (80), we have used

$$\frac{\hat{f}(x^n|a^n,\tilde{\overline{\boldsymbol{\eta}}}\in\Theta_l)}{\tilde{f}_m(x^n|a^n)} \leq \frac{1}{(M-1)L}$$

for $(x^n,a^n) \in \Delta_m^n$. In (81), $\tilde{P}$ is the probability under Θ_m when the prior on $\overline{\boldsymbol{\eta}}$ is $f(\tilde{\overline{\boldsymbol{\eta}}}|\tilde{\overline{\boldsymbol{\eta}}}\in\Theta_m)$. Inequality (82) follows from $\tilde{P}(\delta = m|\tilde{\overline{\boldsymbol{\eta}}}\in\Theta_m) \leq 1$ and the union bound. Choosing $L = 1/\alpha$ completes the proof. ■

$$\begin{aligned} &\geq \liminf_{n\to\infty} \frac{1}{n^\beta} \log \int_{\overline{\boldsymbol{\eta}}'\in\Theta_l\cap B_\delta(\overline{\boldsymbol{\eta}}(\overline{\boldsymbol{\kappa}}))} \exp\left\{ n\sum_{i=1}^K \frac{N_i^n}{n}\left(\left(\boldsymbol{\eta}_i' - \boldsymbol{\eta}_i^*(m)\right)^T \boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i') + \mathcal{A}_i(\boldsymbol{\eta}_i^*(m))\right)\right. && (67) \\ &\qquad \left. + n\sum_{i=1}^K \left[\frac{N_i^n}{n}\left(\left(\boldsymbol{\eta}_i' - \boldsymbol{\eta}_i^*(m)\right)^T (\hat{\boldsymbol{\kappa}}_i - \boldsymbol{\kappa}_i) + (\boldsymbol{\eta}_i')^T \frac{\boldsymbol{\Upsilon}_i}{N_i^n}\right) - \frac{n_{0i}}{n}\mathcal{A}_i(\boldsymbol{\eta}_i')\right]\right\} d\overline{\boldsymbol{\eta}}' \\ &\geq \liminf_{n\to\infty} \frac{1}{n^\beta} \log \exp\left\{ n\sum_{i=1}^K \frac{N_i^n}{n}\Big(\left(\boldsymbol{\eta}_i(\boldsymbol{\kappa}_i) - \boldsymbol{\eta}_i^*(m)\right)^T \boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i(\boldsymbol{\kappa}_i)) + \mathcal{A}_i(\boldsymbol{\eta}_i^*(m))\Big)\right\} && (68) \\ &\qquad + \liminf_{n\to\infty} \frac{1}{n^\beta} \log \int_{\overline{\boldsymbol{\eta}}'\in\Theta_l\cap B_\delta(\overline{\boldsymbol{\eta}}(\overline{\boldsymbol{\kappa}}))} \exp\left\{ n\sum_{i=1}^K \frac{N_i^n}{n}(-\varepsilon) + n\sum_{i=1}^K \frac{N_i^n}{n}(-C_\delta\varepsilon) - C_\delta\right\} d\overline{\boldsymbol{\eta}'} \\ &\geq \liminf_{n\to\infty} \sum_{i=1}^K \frac{N_i^n}{n^\beta} D(\boldsymbol{\eta}_i(\boldsymbol{\kappa}_i) \parallel \boldsymbol{\eta}_i^*(m)) + \liminf_{n\to\infty} \frac{1}{n^\beta} \log\left(\mathsf{Leb}\left(\overline{\boldsymbol{\eta}}' \in \Theta_l \cap B_\delta(\overline{\boldsymbol{\eta}}(\overline{\boldsymbol{\kappa}}))\right)\right) \\ &\qquad - \limsup_{n\to\infty} \left(\sum_{i=1}^K \frac{N_i^n}{n^\beta}(\varepsilon) + \sum_{i=1}^K \frac{N_i^n}{n^\beta}(C_\delta\varepsilon) + \frac{C_\delta}{n^\beta}\right) && (69) \\ &\geq \liminf_{n\to\infty} \sum_{i=1}^K \frac{N_i^n}{n^\beta}\left(D(\boldsymbol{\eta}_i(\boldsymbol{\kappa}_i) \parallel \boldsymbol{\eta}_i^*(m)) - (1+C_\delta)\varepsilon\right), && (70) \end{aligned}$$

APPENDIX E
ACHIEVABILITY

We first show some preliminary results before we get to achievability. In the following Proposition, we assert that several statements hold almost surely. We show that the hypothesis $l^*(n)$ chosen by the policy is eventually the correct one. In addition, we show that the parameters $\boldsymbol{\eta}_i^*(l^*(n))$ chosen by the policy converge to the true or the actual parameters. Furthermore, we will strengthen (62) to show that $Z_{lm}(n)$ is linear in n and that the drift is at least $D^*(\overline{\boldsymbol{\eta}})$.

Proposition 14: Let Assumption A hold. Let $\overline{\boldsymbol{\eta}} \in \Theta_l$ be the true configuration. Consider the non-stopping policy $\tilde{\pi}_{SMF}(\gamma)$. Then, the following convergences hold almost surely:

$$l^*(n) \to l, \tag{83}$$

$$\boldsymbol{\eta}_i^*(l^*(n)) \to \boldsymbol{\eta}_i \text{ for all } i, \tag{84}$$

$$\lambda_i^*(\overline{\boldsymbol{\eta}}^*(l^*(n)) \to \lambda_i^*(\overline{\boldsymbol{\eta}}) \text{ for all } i, \tag{85}$$

$$\frac{N_i^{n,a}}{n^a} \to \lambda_i^*(\overline{\boldsymbol{\eta}}) \text{ for all } i \tag{86}$$

$$\frac{N_i^n}{n} \to \lambda_i^*(\overline{\boldsymbol{\eta}}) \text{ for all } i \tag{87}$$

$$\liminf_{n\to\infty} \frac{Z_l(n)}{n} \geq D^*(\overline{\boldsymbol{\eta}}). \tag{88}$$

Proof: From (62), we have

$$\liminf_{n\to\infty} Z_l(n) = \liminf_{n\to\infty} Z_{lm}(n) > 0 \text{ almost surely.} \tag{89}$$

Fix $m \neq l$. Then, the following inequalites hold almost surely:

$$\begin{aligned}
\limsup_{n\to\infty} Z_m(n) &= \limsup_{n\to\infty} \min_{p\neq m} Z_{mp}(n) && (90)\\
&\leq \limsup_{n\to\infty} Z_{ml}(n) && (91)\\
&\leq \limsup_{n\to\infty} -Z_{lm}(n) && (92)\\
&\quad \text{(a property of the modified GLR)}\\
&= -\liminf_{n\to\infty} Z_{lm}(n) && (93)\\
&\leq -\liminf_{n\to\infty} \min_{p\neq l} Z_{lp}(n) && (94)\\
&= -\liminf_{n\to\infty} Z_l(n) && (95)\\
&< 0. && (96)
\end{aligned}$$

This further implies that a.s. $l^*(n) = \max_p Z_p(n) = l$, for all sufficiently large n. This completes the proof for (83).

For (84), we use (83) to get, a.s.,

$$\boldsymbol{\eta}_i^*(l^*(n)) = \boldsymbol{\eta}_i^*(l) \tag{97}$$

for all sufficiently large n, and then Proposition 13 to get $\boldsymbol{\eta}_i^*(l) \to \boldsymbol{\eta}_i$, which then yields $\boldsymbol{\eta}_i^*(l^*(n)) \to \boldsymbol{\eta}_i$.

The convergence in (85) follows from (84) and Assumption A.

The convergence in (86) follows from (85) and Lemma 11.

Proof of (87): Let $\{V_1, V_2, \ldots, V_{n^a}\}$ be such that V_k is the number of sluggish instants plus one active instance corresponding to the kth active instance, $k = 1, 2, \ldots, n^a$. Then V_t's are independent and identical random variables with the geometric distribution of parameter γ. Additionally, to make the total of n arm pulls at time instant n, the last 'sluggish run' should also be accounted. We do this by re-writing the expression in (22) as

$$N_i^n = \sum_{t=1}^{n^a} V_t 1_{\{a_t=i\}} + \overline{V}_i \tag{98}$$

where $\overline{V}_i$ is nonzero for at most for one i and corresponds to the latest sluggish run at time instant n. To study the limit of N_i^n/n, it suffices to study

$$\frac{1}{n}\sum_{t=1}^{n^a} V_t 1_{\{a_t=i\}} = \frac{n^a}{n} \cdot \frac{N_i^{n,a}}{n^a} \cdot \frac{1}{N_i^{n,a}} \sum_{t=1}^{n^a} V_t 1_{\{a_t=i\}}. \tag{99}$$

We consider each term on the right-hand side of (99) in detail. Note that $n^a/n \to \gamma$ and from (86) we get $N_i^{n,a}/n^a \to \lambda_i^*(\overline{\boldsymbol{\eta}})$. Also by Lemma 11 we have $N_i^{n,a} \to \infty$ as $n \to \infty$. Note that the summation in (99) has $N_i^{n,a}$ terms, and hence the sample mean converges to the expected value of V_t which is $1/\gamma$. Hence,we get, almost surely,

$$\lim_{n\to\infty} \frac{N_i^n}{n} = \gamma \cdot \lambda_i^*(\overline{\boldsymbol{\eta}}) \cdot \frac{1}{\gamma} = \lambda_i^*(\overline{\boldsymbol{\eta}}). \tag{100}$$

This concludes the proof of (87).

Proof of (88): In the proof of (62), instead of scaling by $1/n^\beta$, rescale by $1/n$, and arrive at

$$\liminf_{n\to\infty} \frac{Z_l(n)}{n} \geq \inf_{\boldsymbol{\eta}' \in \Theta_m} \liminf_{n\to\infty} \sum_{i=1}^{K} \frac{N_i^n}{n} D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i'),$$

which is the equivalent of (70) with an additional infimum over all $\overline{\boldsymbol{\eta}}' \in \Theta_m$. The result in (88) then follows from (87). ■

In the next three subsections we prove each of the three claims in Proposition 7.

A. Proof of (45)

Proof: We begin by proving that, as the probability of false detection constraint goes to zero, the stopping time of the policy goes to infinity (Lemma 15). We then combine this result with Proposition 14 to complete the required proof.

Lemma 15: Let $\overline{\boldsymbol{\eta}} \in \Theta_l$ be the true configuration. Consider the policy $\pi_{SMF}(L, \gamma, \beta)$. Then,

$$\liminf_{L\to\infty} \tau(\pi_{SMF}(L, \gamma, \beta)) \to \infty \text{ a.s.} \tag{101}$$

Proof: It suffices to show that, as $L \to \infty$,

$$P(\tau(\pi_{SMF}(L, \gamma, \beta)) < n) \to 0 \text{ for all } n. \tag{102}$$

Fix some $\boldsymbol{\eta}'(m) \in \Theta_{-m}$, for each m. We begin with

$$\begin{aligned}
&\limsup_{L\to\infty} P(\tau(\pi_{SMF}(L,\gamma,\beta)) < n) \\
&= \limsup_{L\to\infty} P\Big(\max_{1\le t\le n} Z_m(t) > \log((M-1)L) \\
&\qquad \text{for some } m\Big) && (103)\\
&\le \limsup_{L\to\infty} \sum_{m=1}^{M}\sum_{t=1}^{n} P\left(Z_m(t) > \log((M-1)L)\right) && (104)\\
&\le \limsup_{L\to\infty} \frac{1}{\log((M-1)L)} \cdot \\
&\qquad \sum_{m=1}^{M}\sum_{t=1}^{n} E\left[\sum_{i=1}^{K} N_i^t D(\hat{\boldsymbol{\eta}}_i \parallel \boldsymbol{\eta}_i'(m))\right] && (105)\\
&\le \limsup_{L\to\infty} \frac{1}{\log((M-1)L)} \sum_{m=1}^{M}\sum_{t=1}^{n}\sum_{i=1}^{K} E\Big[t(\hat{\boldsymbol{\eta}}_i^T\hat{\boldsymbol{\kappa}}_i \\
&\qquad -\mathcal{A}_i(\hat{\boldsymbol{\eta}}_i) - \boldsymbol{\eta}_i'(m)^T\hat{\boldsymbol{\kappa}}_i + \mathcal{A}_i(\boldsymbol{\eta}_i'(m)))\Big] && (106)\\
&\le \limsup_{L\to\infty} \frac{1}{\log((M-1)L)} \sum_{m=1}^{M}\sum_{t=1}^{n}\sum_{i=1}^{K} t\Big(E\left[\hat{\boldsymbol{\eta}}_i^T\hat{\boldsymbol{\kappa}}_i\right] \\
&\qquad -\mathcal{A}_i(\boldsymbol{\eta}_i) - \boldsymbol{\eta}_i'(m)^T\boldsymbol{\kappa}_i + \mathcal{A}_i(\boldsymbol{\eta}_i'(m))\Big) && (107)\\
&= 0. && (108)
\end{aligned}$$

Inequality in (104) follows from union bound. Inequality (105) is discussed below. Equality in (106) is obtained using the expression for relative entropy and the fact that $N_i^t \le t$. The inequality in (107) is obtained from the observation that $\mathcal{A}_i(\cdot)$ is convex and by an application of Jensen's inequality. The final equality in (108) follows because $E\left[\hat{\boldsymbol{\eta}}_i^T\hat{\boldsymbol{\kappa}}_i\right]$ is finite. Inequality in (105) is obtained using the following result

$$\begin{aligned}
&Z_m(t) \\
&= \min_{p\neq m} \log \frac{\tilde{f}_m(x^t|a^t)}{\hat{f}(x^t|x^t, \overline{\boldsymbol{\eta}} \in \Theta_p)} \\
&= \min_{p\neq m} \inf_{\overline{\boldsymbol{\eta}}'\in\Theta_p} \log \frac{\tilde{f}_m(x^t|a^t)}{f(x^t|a^t,\overline{\boldsymbol{\eta}}')} \\
&\le \log \frac{\tilde{f}_m(x^t|a^t)}{f(x^t|a^t,\overline{\boldsymbol{\eta}}'(m))}, \text{ using the chosen } \overline{\boldsymbol{\eta}}'(m) \in \Theta_{-m} \\
&\le \log \frac{\sup_{\tilde{\overline{\boldsymbol{\eta}}}\in\Omega} f(x^t|a^t,\tilde{\overline{\boldsymbol{\eta}}})}{f(x^t|a^t,\overline{\boldsymbol{\eta}}'(m))} \\
&= \sum_{i=1}^{K} N_i^t\left[\hat{\boldsymbol{\eta}}_i^T\hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\hat{\boldsymbol{\eta}}_i) - \boldsymbol{\eta}_i'(m)^T\hat{\boldsymbol{\kappa}}_i + \mathcal{A}_i(\boldsymbol{\eta}_i'(m))\right] \\
&= \sum_{i=1}^{K} N_i^t D(\hat{\boldsymbol{\eta}}_i \parallel \boldsymbol{\eta}_i'(m)).
\end{aligned}$$

The last quantity being positive, we can now apply the Markov inequality and (105) follows. This finishes the proof of the lemma. ■

Lemma 16: Let Assumption A hold. Let $\overline{\boldsymbol{\eta}} \in \Theta_l$ be the true configuration. Consider the policy $\pi_{SMF}(L,\gamma,\beta)$. We then have

$$\begin{aligned}
\liminf_{L\to\infty} \frac{Z_l(\tau(\pi_{SMF}(L,\gamma,\beta)))}{\tau(\pi_{SMF}(L,\gamma,\beta))} &\ge D^*(\overline{\boldsymbol{\eta}}) \text{ a.s.}, && (109)\\
\liminf_{L\to\infty} \frac{Z_l(\tau(\pi_{SMF}(L,\gamma,\beta))-1)}{\tau(\pi_{SMF}(L,\gamma,\beta))-1} &\ge D^*(\overline{\boldsymbol{\eta}}) \text{ a.s.} && (110)
\end{aligned}$$

Proof: The proofs of the two statements follow by focusing on sample paths that satisfy (88) of Proposition 14 and (101) of Lemma 15. The argument goes as follows. For any such sample path ω and any $\epsilon > 0$, there is an $N(\omega,\epsilon)$, independent of L, such that $Z_l(n)/n \ge D^*(\overline{\boldsymbol{\eta}}) - \epsilon$ for all $n \ge N(\omega,\epsilon)$. Now take L to infinity and employ Lemma 15 to get that $\tau(\pi_{SMF}(L,\gamma,\beta))$ is eventually bigger than $N(\omega,\epsilon)+1$, and so $\tau(\pi_{SMF}(L,\gamma,\beta)) - 1 \ge N(\omega,\epsilon)$. So both (109) and (110) hold. ■

We now begin the proof for (45). Using the definition for $\tau(\pi_{SMF}(L,\gamma,\beta))$, at the time slot prior to stoppage, we must have $Z_l(\tau(\pi_{SMF}(L,\gamma,\beta))-1) < \log((M-1)L)$. So,

$$\begin{aligned}
&\limsup_{L\to\infty} \frac{Z_l(\tau(\pi_{SMF}(L,\gamma,\beta))-1)}{\log(L)} \\
&\quad\le \limsup_{L\to\infty} \frac{\log((M-1)L)}{\log(L)} = 1. && (111)
\end{aligned}$$

Thus

$$\begin{aligned}
1 &\ge \limsup_{L\to\infty} \frac{Z_l(\tau(\pi_{SMF}(L,\gamma,\beta))-1)}{\log(L)} && (112)\\
&\ge \liminf_{L\to\infty} \frac{Z_l(\tau(\pi_{SMF}(L,\gamma,\beta))-1)}{\tau(\pi_{SMF}(L,\gamma,\beta))-1} \\
&\qquad \cdot \limsup_{L\to\infty} \frac{\tau(\pi_{SMF}(L,\gamma,\beta))-1}{\log L} && (113)\\
&\ge D^*(\overline{\boldsymbol{\eta}}) \cdot \limsup_{L\to\infty} \frac{\tau(\pi_{SMF}(L,\gamma,\beta))-1}{\log L}, && (114)
\end{aligned}$$

where in the last inequality, we have used (109). Finally,

$$\begin{aligned}
&\limsup_{L\to\infty} \frac{\tau(\pi_{SMF}(L,\gamma,\beta))}{\log(L)} \\
&\quad= \limsup_{L\to\infty} \frac{\tau(\pi_{SMF}(L,\gamma,\beta))-1}{\log(L)} && (115)\\
&\quad\le \frac{1}{D^*(\overline{\boldsymbol{\eta}})} \text{ a.s..} && (116)
\end{aligned}$$

This completes the proof of (45). ■

B. Proof of (46)

We begin with a couple of lemmas.

Lemma 17: For every l, for every i, the mapping

$$\Theta_l \ni \boldsymbol{\eta}_i \mapsto \inf_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i')$$

is continuous.

Proof: Write $\mathcal{G}_l(\boldsymbol{\eta}_i) := \inf_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i')$. Observe that $\mathcal{G}_l \ge 0$. Fix $\epsilon > 0$.

(i) Consider a sequence $\boldsymbol{\eta}_i(n) \to \boldsymbol{\eta}_i$ with all of them being in Θ_l. By the definition of $\mathcal{G}_l(\boldsymbol{\eta}_i)$, there exists $\boldsymbol{\eta}_i' \in \Theta_{-l}$ such

that $D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i') \leq \mathcal{G}_l(\boldsymbol{\eta}_i) + \epsilon$. We then have

$$\begin{aligned}
&\mathcal{G}_l(\boldsymbol{\eta}_i(n)) \\
&\quad \leq \; D(\boldsymbol{\eta}_i(n) \parallel \boldsymbol{\eta}_i') \quad \text{(since the left-side is an infimum)} \\
&\quad \leq \; D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i') + \epsilon \text{ (for all sufficiently large } n, \\
&\qquad\qquad \text{using continuity of relative entropy)} \\
&\quad \leq \; \mathcal{G}_l(\boldsymbol{\eta}_i) + 2\epsilon \quad \text{(by the choice of } \boldsymbol{\eta}_i').
\end{aligned}$$

Thus $\limsup_{n\to\infty} \mathcal{G}_l(\boldsymbol{\eta}_i(n)) \leq \mathcal{G}_l(\boldsymbol{\eta}_i)$.

(ii) Since $\mathcal{G}_l(\boldsymbol{\eta}_i(n))$ is bounded, by Lemma 8 (see footnote 2) there exists a convergent sequence of $\boldsymbol{\eta}_i'(n) \to \boldsymbol{\eta}_i'$. By the same argument that led to (57), using the bounded of the $\boldsymbol{\eta}_i'(n)$ sequence, for all sufficiently large n, we have

$$\mathcal{G}_l(\boldsymbol{\eta}_i(n)) \geq D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i'(n)) - 2\epsilon \geq \mathcal{G}_l(\boldsymbol{\eta}_i) - 2\epsilon.$$

This establishes that $\liminf_{n\to\infty} \mathcal{G}_l(\boldsymbol{\eta}_i(n)) \geq \mathcal{G}_l(\boldsymbol{\eta}_i)$.

Together, (i) and (ii) establish the continuity of $\mathcal{G}_l(\cdot)$. ■

The next lemma provides an estimate of the probability that the likelihood for the correct hypothesis is small.

Lemma 18: Let Assumption A hold. Fix $L > 1$. Let $\overline{\boldsymbol{\eta}} \in \Theta_l$ be the true configuration. Then there exists a constant $0 < B < \infty$ and a constant N_0, both independent of L, such that for all $n \geq \max\{2\log((M-1)L)/D^*(\boldsymbol{\eta}), \mathcal{N}_0\}$, we have

$$P(Z_l(n) < \log((M-1)L)) < \frac{B}{n^3}. \tag{117}$$

Proof: Before we start with the proof, let us note that

$$\begin{aligned}
&D^*(\overline{\boldsymbol{\eta}}) \\
&\quad = \; \inf_{\overline{\boldsymbol{\eta}}' \in \Theta_{-l}} \sum_{i=1}^{K} \lambda_i^* D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i') \qquad (118) \\
&\quad = \; \inf_{\overline{\boldsymbol{\eta}}' \in \Theta_{-l}} \sum_{i=1}^{K} \lambda_i^* \left[(\boldsymbol{\eta}_i - \boldsymbol{\eta}_i')^T \boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i) + \mathcal{A}_i(\boldsymbol{\eta}_i') \right] \qquad (119) \\
&\quad = \; \sum_{i=1}^{K} \lambda_i^* \left(\boldsymbol{\eta}_i^T \boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i) \right) \\
&\qquad - \sup_{\overline{\boldsymbol{\eta}}' \in \Theta_{-l}} \sum_{i=1}^{K} \lambda_i^* \left(\boldsymbol{\eta}_i'^T \boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i') \right). \qquad (120)
\end{aligned}$$

Let us now turn to the probability of interest. Observe that $Z_l(n) = \min_{m \neq l} Z_{lm}(n)$. Using (41), we obtain the following inequality:

$$\begin{aligned}
&P(Z_l(n) < \log((M-1)L)) \\
&\quad \leq \; P\left(\log \mathcal{H}_l(\overline{\boldsymbol{\Upsilon}}, \mathbf{n_0}) < -\epsilon' n\right) \\
&\qquad + P\Big(-\log \mathcal{H}_l(\mathbf{Y} + \overline{\boldsymbol{\Upsilon}}, \mathbf{N} + \mathbf{n_0}) \\
&\qquad - n \sum_{i=1}^{K} \lambda_i^* [\boldsymbol{\eta}_i^T \boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i)] < -\epsilon' n \Big) \\
&\qquad + P\left(- \sup_{\overline{\boldsymbol{\eta}}' \in \Theta_{-l}} \sum_{i=1}^{K} N_i^n \left(\boldsymbol{\eta}_i'^T \hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\boldsymbol{\eta}_i') \right) \right. \\
&\qquad \left. + \sup_{\overline{\boldsymbol{\eta}}' \in \Theta_{-l}} n \sum_{i=1}^{K} \lambda_i^* \left(\boldsymbol{\eta}_i'^T \boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i') \right) < -\epsilon' n \right) \\
&\qquad + P\left(nD^*(\overline{\boldsymbol{\eta}}) - 3\epsilon' n < \log((M-1)L)\right); \qquad (121)
\end{aligned}$$

the inequality in (121) is obtained using union bound together with adding and subtracting $D^*(\overline{\boldsymbol{\eta}})$. Our goal is now to show that each of the terms on the right-hand side above is either 0 or $O(n^{-3})$.

(i) We begin with the last term in (121). Let

$$\epsilon = \frac{D^*(\overline{\boldsymbol{\eta}})}{D^*(\overline{\boldsymbol{\eta}}) - 3\epsilon'} - 1, \tag{122}$$

and

$$n_0 = \frac{2\log((M-1)L)}{D^*(\overline{\boldsymbol{\eta}})} > \frac{(1+\epsilon)\log((M-1)L)}{D^*(\overline{\boldsymbol{\eta}})}. \tag{123}$$

This is n_0 is one of the values that n must exceed in the statement of the lemma. Then for $n > n_0$, we have

$$\begin{aligned}
n(D^*(\overline{\boldsymbol{\eta}}) - 3\epsilon') \;&>\; \frac{(1+\epsilon)\log((M-1)L)}{D^*(\overline{\boldsymbol{\eta}})} [D^*(\overline{\boldsymbol{\eta}}) - 3\epsilon'] \\
&=\; \log((M-1)L). \qquad (124)
\end{aligned}$$

Hence, we get that for $n > n_0$,

$$P\left(nD^*(\overline{\boldsymbol{\eta}}) - 3\epsilon' n < \log((M-1)L)\right) = 0. \tag{125}$$

(ii) Consider next the first term in (121):

$$P\left(\log \mathcal{H}_l(\overline{\boldsymbol{\Upsilon}}, \mathbf{n_0}) < -\epsilon' n\right). \tag{126}$$

The right-hand side inside the probability goes to negative infinity whereas, the left-hand side is a constant. Hence, the probability of the event under study is zero for all sufficiently large n (independent of L).

(iii) Next consider the second term in (121). For convenience define $\mathcal{F}_i(\boldsymbol{\kappa}_i) := \boldsymbol{\eta}_i^T \boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i)$, the Fenchel dual of $\mathcal{A}_i$ evaluated at $\boldsymbol{\kappa}_i$. We then have

$$\begin{aligned}
&P\left(-\frac{1}{n} \log \mathcal{H}_l(\mathbf{Y} + \overline{\boldsymbol{\Upsilon}}, \mathbf{N} + \mathbf{n_0}) - \sum_{i=1}^{K} \lambda_i^* \mathcal{F}_i(\boldsymbol{\kappa}_i) < -\epsilon' \right) \\
&\leq \; P\left(-\frac{1}{n} \log \mathcal{H}_l(\mathbf{Y} + \overline{\boldsymbol{\Upsilon}}, \mathbf{N} + \mathbf{n_0}) - \sum_{i=1}^{K} \lambda_i^* \mathcal{F}_i(\boldsymbol{\kappa}_i) < -\epsilon', \right. \\
&\qquad \left. \left| \frac{N_{i'}^n}{n} - \lambda_{i'}^* \right| \leq \epsilon_1, \left\| \frac{\mathbf{Y}_{i'}^n}{N_{i'}^n} - \boldsymbol{\kappa}_{i'} \right\|_\infty \leq \epsilon_2, \forall i' \right) \\
&\qquad + \sum_{i'} P\left(\left| \frac{N_{i'}^n}{n} - \lambda_{i'}^* \right| > \epsilon_1 \right) \\
&\qquad + \sum_{i'} P\left(\left\| \frac{\mathbf{Y}_{i'}^n}{N_{i'}^n} - \boldsymbol{\kappa}_{i'} \right\|_\infty > \epsilon_2 \right), \qquad (127)
\end{aligned}$$

where ϵ_1, ϵ_2 are suitable constants that will be specified soon. Under the conditions

$$\left| \frac{N_{i'}^n}{n} - \lambda_{i'}^* \right| < \epsilon_1 \text{ and } \left\| \frac{\mathbf{Y}_{i'}^n}{N_{i'}^n} - \boldsymbol{\kappa}_{i'} \right\|_\infty \leq \epsilon_2,$$

we will follow the steps leading to (69) to lower bound $-\frac{1}{n}\log \mathcal{H}_l(\mathbf{Y} + \overline{\boldsymbol{\Upsilon}}, \mathbf{N} + \mathbf{n_0})$. First observe that (128)-(129), as shown at the bottom of the next page, hold.

Note that $\boldsymbol{\eta}_i(\boldsymbol{\kappa}_i)$ optimises the function $\boldsymbol{\eta}_i' \mapsto (\boldsymbol{\eta}')^T \boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i')$, for each $i = 1, \ldots, K$, over the set Θ_l, since $\overline{\boldsymbol{\eta}} \in \Theta_l$. As before, we leverage this.

Fix $\delta > 0$. Almost surely, there is a $C_\delta > 0$ such that for all sufficiently large n, $||\hat{\boldsymbol{\kappa}}_i - \boldsymbol{\kappa}_i||_\infty \leq \epsilon_2$, and further, for all $\overline{\boldsymbol{\eta}}' \in B_\delta(\overline{\boldsymbol{\eta}})$, we have:

$$\left|\sum_{i=1}^{K}\left(\frac{N_i^n}{n} - \lambda_i^*\right)\left((\boldsymbol{\eta}_i')^T\boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i')\right)\right| \leq C_\delta\epsilon_1$$

$$\left|(\boldsymbol{\eta}_i')^T\boldsymbol{\Upsilon}_i - n_{0i}\mathcal{A}_i(\boldsymbol{\eta}_i')\right| \leq C_\delta$$

$$\left|\sum_{i=1}^{K}\frac{N_i^n}{n}(\boldsymbol{\eta}_i')^T(\hat{\boldsymbol{\kappa}}_i - \boldsymbol{\kappa}_i)\right| \leq C_\delta\epsilon_2$$

$$\left|\sum_{i=1}^{K}\left((\boldsymbol{\eta}_i')^T\boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i') - \left(\boldsymbol{\eta}_i^T\boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i)\right)\right)\right|$$

$$= \left|\sum_{i=1}^{K}\left((\boldsymbol{\eta}_i')^T\boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i') - \mathcal{F}(\boldsymbol{\kappa}_i)\right)\right| \leq \tau(\delta)$$

where in the last inequality $\tau(\delta) \to 0$ as $\delta \to 0$ due to the continuity of $\mathcal{A}_i(\cdot)$. Further, we can lower bound the integral in (129) by restricting the integral to the set $\Theta_l \cap B_\delta(\overline{\boldsymbol{\eta}})$. Putting these ideas together, we get that (129) is lower bounded by (130)-(133), as shown at the bottom of the page. Using this lower bound, we can now upper bound the first term in (127) with

$$P\left(\frac{1}{n}\log\left(\mathsf{Leb}\left(\overline{\boldsymbol{\eta}}' \in \Theta_l \cap B_\delta(\overline{\boldsymbol{\eta}}(\overline{\boldsymbol{\kappa}}))\right)\right) - \tau(\delta) - C_\delta\epsilon_1 - KC_\delta\epsilon_2 - \frac{KC_\delta}{n} < -\epsilon'\right); \quad (134)$$

the event inside the probability argument on the right-hand side of the above inequality will not occur, via suitable choices of δ, ϵ_1 and ϵ_2, for all sufficiently large n dependent on $\delta, \epsilon_1, \epsilon_2, \epsilon'$ but independent of L. For all such n, the probability on the left-hand side of (134) is zero.

The third term in (127) is upper bounded by C/n^3 for some constant C by Lemma 12 (in fact it decays faster, $O(1/n^4)$). The constant C is independent of L.

We now argue that there is an $\mathcal{N}_0$ independent of L such that, for all $n \geq \mathcal{N}_0$, the second term in (127) is upper bounded by C_1/n^3 for a suitable constant C_1 which is also independent of L.

Let us use the notation $n^a(m)$ to be the number of active samplings up to time m. First observe that, by triangle inequality,

$$\left|\frac{N_{i'}^n}{n} - \lambda_{i'}^*\right| > \epsilon_1 \Rightarrow \left|\frac{N_{i'}^{n,a}}{n^a(n)} - \lambda_{i'}^*\right| > \frac{\epsilon_1}{2} \quad \text{or} \quad \left|\frac{N_{i'}^n}{n} - \frac{N_{i'}^{n,a}}{n^a(n)}\right| > \frac{\epsilon_1}{2}. \quad (135)$$

Choose a sufficiently small ϵ_3, and a sufficiently small ϵ_2 (a new ϵ_2 that may depend on ϵ_3), so that the following hold:

- $3(K-1)\epsilon_3 < \epsilon_1/2$;
- for every $\overline{\boldsymbol{\kappa}}'$ with $\max_i \|\boldsymbol{\kappa}_i' - \boldsymbol{\kappa}_i\|_\infty < \epsilon_2$, we have $\overline{\boldsymbol{\eta}}(\boldsymbol{\kappa}') \in \Theta_l$ (due to the openness of Θ_l and continuity of the $\overline{\boldsymbol{\eta}}(\cdot)$ mapping);

$$-\frac{1}{n}\log\mathcal{H}_l(\mathbf{Y} + \overline{\boldsymbol{\Upsilon}}, \mathbf{N} + \mathbf{n_0}) = \frac{1}{n}\log\left\{\int_{\overline{\boldsymbol{\eta}}' \in \Theta_l}\exp\left\{\sum_{i=1}^{K}\left((\boldsymbol{\eta}_i')^T(\mathbf{Y}_i^n + \boldsymbol{\Upsilon}_i) - (N_i^n + n_{0i})\mathcal{A}_i(\boldsymbol{\eta}_i')\right)\right\}d\overline{\boldsymbol{\eta}}'\right\} \quad (128)$$

$$= \frac{1}{n}\log\int_{\overline{\boldsymbol{\eta}}' \in \Theta_l}\exp\left\{n\sum_{i=1}^{K}\left(\frac{N_i^n}{n}\left((\boldsymbol{\eta}_i')^T\frac{\mathbf{Y}_i^n}{N_i^n} - \mathcal{A}_i(\boldsymbol{\eta}_i')\right) + (\boldsymbol{\eta}_i')^T\frac{\boldsymbol{\Upsilon}_i}{n} - \frac{n_{0i}}{n}\mathcal{A}_i(\boldsymbol{\eta}_i')\right)\right\}d\overline{\boldsymbol{\eta}}'. \quad (129)$$

$$-\frac{1}{n}\log\mathcal{H}_l(\mathbf{Y} + \overline{\boldsymbol{\Upsilon}}, \mathbf{N} + \mathbf{n_0})$$

$$\geq \frac{1}{n}\log\int_{\overline{\boldsymbol{\eta}}' \in \Theta_l \cap B_\delta(\overline{\boldsymbol{\eta}})}\exp\left\{n\sum_{i=1}^{K}\frac{N_i^n}{n}\left((\boldsymbol{\eta}_i')^T\frac{\mathbf{Y}_i^n}{N_i^n} - \mathcal{A}_i(\boldsymbol{\eta}_i')\right) - KC_\delta\right\}d\overline{\boldsymbol{\eta}}' \quad (130)$$

$$= \frac{1}{n}\log\int_{\overline{\boldsymbol{\eta}}' \in \Theta_l \cap B_\delta(\overline{\boldsymbol{\eta}})}\exp\left\{n\sum_{i=1}^{K}\lambda_i^*\left((\boldsymbol{\eta}_i')^T\boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i')\right)\right.$$
$$\left. + n\sum_{i=1}^{K}\left(\frac{N_i^n}{n} - \lambda_i^*\right)\left((\boldsymbol{\eta}_i')^T\boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i')\right) + n\sum_{i=1}^{K}\frac{N_i^n}{n}(\boldsymbol{\eta}_i')^T(\hat{\boldsymbol{\kappa}}_i - \boldsymbol{\kappa}_i) - KC_\delta\right\}d\overline{\boldsymbol{\eta}}' \quad (131)$$

$$= \frac{1}{n}\log\int_{\overline{\boldsymbol{\eta}}' \in \Theta_l \cap B_\delta(\overline{\boldsymbol{\eta}})}\exp\left\{n\sum_{i=1}^{K}\lambda_i^*\mathcal{F}(\boldsymbol{\kappa}_i) + n\sum_{i=1}^{K}\lambda_i^*\left((\boldsymbol{\eta}_i')^T\boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i') - \mathcal{F}(\boldsymbol{\kappa}_i)\right)\right.$$
$$\left. + n\sum_{i=1}^{K}\left(\frac{N_i^n}{n} - \lambda_i^*\right)\left((\boldsymbol{\eta}_i')^T\boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i')\right) + n\sum_{i=1}^{K}\frac{N_i^n}{n}(\boldsymbol{\eta}_i')^T(\hat{\boldsymbol{\kappa}}_i - \boldsymbol{\kappa}_i) - KC_\delta\right\}d\overline{\boldsymbol{\eta}}' \quad (132)$$

$$\geq \sum_{i=1}^{K}\lambda_i^*\mathcal{F}(\boldsymbol{\kappa}_i) + \frac{1}{n}\log\left(\mathsf{Leb}\left(\overline{\boldsymbol{\eta}}' \in \Theta_l \cap B_\delta(\overline{\boldsymbol{\eta}})\right)\right) - \tau(\delta) - C_\delta\epsilon_1 - C_\delta\epsilon_2 - \frac{KC_\delta}{n}. \quad (133)$$

- for every $\overline{\boldsymbol{\kappa}}'$ with $\max_i \|\boldsymbol{\kappa}_i' - \boldsymbol{\kappa}_i\|_\infty < \epsilon_2$, we have $\max_i |\lambda_i^*(\overline{\boldsymbol{\eta}}(\overline{\boldsymbol{\kappa}})) - \lambda_i^*| < \epsilon_3$ (due to the continuities of the $\boldsymbol{\lambda}^*(\cdot)$ and the $\overline{\boldsymbol{\eta}}(\cdot)$ mappings).

Consider the conditions

$$\sup_{m\geq n}\left|\frac{n^a(m)}{m}-\gamma\right| \leq \gamma\epsilon_1 \text{ and}$$
$$\sup_{m:n^a(m)\geq 3\epsilon_3\gamma(1-\epsilon_1)n} \max_{i'}\left\|\frac{\mathbf{Y}_{i'}^m}{N_{i'}^m}-\boldsymbol{\kappa}_{i'}\right\|_\infty \leq \epsilon_2. \quad (136)$$

Under these conditions, apply Lemma 11 with $\epsilon=\epsilon_3$, $n_0 = 3\epsilon_3\gamma(1-\epsilon_1)n$, and $n_{\epsilon_3} = \gamma(1-\epsilon_1)n$ for $n \geq \mathcal{N}_0 := 1/(3\epsilon_3^2(1-\epsilon_1)\gamma)$, use the second and the third bullets above, and we get

$$\sup_{m:n^a(m)\geq\gamma(1-\epsilon_1)n}\left|\frac{N_{i'}^{m,a}}{n^a(m)}-\lambda_{i'}^*\right| \leq 3(K-1)\epsilon_3.$$

In particular, since we are operating under $n^a(m)/m \geq \gamma(1-\epsilon_1)$ for all $m \geq n$, taking $m=n$ in the above displayed equation, we get

$$\left|\frac{N_{i'}^{n,a}}{n^a(n)}-\lambda_{i'}^*\right| \leq 3(K-1)\epsilon_3,$$

a condition which is incompatible with $\left|\frac{N_{i'}^{n,a}}{n^a(n)}-\lambda_{i'}^*\right| > \epsilon_1/2$ in view of the first bullet above. This contradiction, along with (135), (136), shows that, for $n\geq\mathcal{N}_0$,

$$\left|\frac{N_{i'}^n}{n}-\lambda_{i'}^*\right| > \epsilon_1$$
$$\Rightarrow \sup_{m\geq n}\left|\frac{n^a(m)}{m}-\gamma\right| > \gamma\epsilon_1 \quad (137)$$
$$\text{or}\Big[\sup_{m\geq n}\frac{n^a(m)}{m} \leq \gamma(1+\epsilon_1) \quad \text{and}$$
$$\sup_{m:n^a(m)\geq 3\epsilon_3\gamma(1-\epsilon_1)n}\max_{i'}\left\|\frac{\mathbf{Y}_{i'}^m}{N_{i'}^m}-\boldsymbol{\kappa}_{i'}\right\|_\infty > \epsilon_2\Big] \quad (138)$$
$$\text{or}\ \left|\frac{N_{i'}^n}{n}-\frac{N_{i'}^{n,a}}{n^a(n)}\right| > \frac{\epsilon_1}{2}. \quad (139)$$

(a) By Bernstein inequality and the union bound, the first event has probability decaying exponentially in n.

(b) Using $n^a(m) \leq \gamma m(1+\epsilon_1)$, if the second event above holds, then we have

$$\sup_{m:m\geq 3\ n\epsilon_3(1-\epsilon_1)/(1+\epsilon_1)}\max_{i'}\left\|\frac{\mathbf{Y}_{i'}^m}{N_{i'}^m}-\boldsymbol{\kappa}_{i'}\right\|_\infty > \epsilon_2;$$

by Lemma 12 and the union bound, its probability is upper bounded by

$$\sum_{m:m\geq 3\ n\epsilon_3(1-\epsilon_1)/(1+\epsilon_1)}\frac{C}{m^4} \quad \leq \quad \frac{C_2}{n^3}$$

for some suitable constant C_2.

(c) Let us now address the third term in (138). Using (99), we get

$$\left|\frac{N_{i'}^n}{n}-\frac{N_{i'}^{n,a}}{n^a(n)}\right| = \frac{N_{i'}^{n,a}}{n^a(n)}\left|\frac{n^a(n)}{n}\frac{1}{N_{i'}^{n,a}}\left(\sum_{k=1}^{N_{i'}^{n,a}}V_k^{(i')}+\overline{V}_i\right)-1\right|$$

where k runs over the indices involving the choice of i' in an active slot. Using $\frac{N_{i'}^{n,a}}{n^a(n)} \leq 1$, we get

$$\left|\frac{N_{i'}^n}{n}-\frac{N_{i'}^{n,a}}{n^a(n)}\right| > \frac{\epsilon_1}{2}$$
$$\Rightarrow \left|\frac{n^a(n)}{n}\frac{1}{N_{i'}^{n,a}}\left(\sum_{k=1}^{N_{i'}^{n,a}}V_k^{(i')}+\overline{V}_i\right)-1\right| > \frac{\epsilon_1}{2}$$
$$\Rightarrow \left|\frac{n^a(n)}{n}-\gamma\right| > \gamma\delta \quad \text{(for a } \delta \text{ to be chosen soon)}$$
$$\text{or}\ \Bigg[\left|\frac{n^a(n)}{n}-\gamma\right| \leq \gamma\delta \quad \text{and}$$
$$\left|\frac{1}{N_{i'}^{n,a}}\sum_{k=1}^{N_{i'}^{n,a}}V_k^{(i')}-\frac{1}{\gamma}\right| > \frac{\delta}{\gamma}\Bigg]$$
$$\text{or}\ \Bigg[\left|\frac{n^a(n)}{n}-\gamma\right| \leq \gamma\delta \quad \text{and}$$
$$\left|\frac{1}{N_{i'}^{n,a}}\sum_{k=1}^{N_{i'}^{n,a}}V_k^{(i')}-\frac{1}{\gamma}\right| \leq \frac{\delta}{\gamma}$$
$$\text{and}\ \frac{n^a(n)}{n}\frac{\overline{V}_i}{N_{i'}^{n,a}} > \frac{\epsilon_1}{4}\Bigg]$$
$$\text{or}\ \Bigg[\left|\frac{n^a(n)}{n}-\gamma\right| \leq \gamma\delta \quad \text{and}$$
$$\left|\frac{1}{N_{i'}^{n,a}}\sum_{k=1}^{N_{i'}^{n,a}}V_k^{(i')}-\frac{1}{\gamma}\right| \leq \frac{\delta}{\gamma}$$
$$\text{and}\quad \left|\frac{n^a(n)}{n}\frac{1}{N_{i'}^{n,a}}\sum_{k=1}^{N_{i'}^{n,a}}V_k^{(i')}-1\right| > \frac{\epsilon_1}{4}\Bigg].$$

Choose δ sufficiently small so that $(1+\delta)^2 < 1+\epsilon_1/4$ and $(1-\delta)^2 > 1-\epsilon_1/4$. The first of these events has exponentially (in n) small probability for all n (Bernstein inequality). By Lemma 11, for all $n\geq\mathcal{N}_0$, we have $N_{i'}^{n,a} \geq (n^a(n))^\beta/2$. The Chernoff bound then gives that the second event too has exponentially (in n) small probability. The random variable $\overline{V}_i$ is stochastically dominated by a geometric random variable and hence the third event has exponentially small probability for all $n\geq\mathcal{N}_0$ and Lemma 11. Finally, by the choice of δ, the fourth event cannot occur.

The above arguments (a)-(c) establish that the probability of the event $\left|\frac{N_{i'}^n}{n}-\lambda_{i'}^*\right| > \epsilon_1$, i.e., the second term in (127), is also upper bounded by C_1/n^3 for some constant C_1 independent of L, for all $n\geq\mathcal{N}_0$.

Note that, with the above, we have also established that the second term in (121) is upper bounded by C/n^3, for some C_3, for all $n\geq\mathcal{N}_0$.

(iv) Finally, consider the third term in (121). The chain of inequalities (140)-(143), as shown at the bottom of the next page, are self-evident: Following the approach that led to the bound for (127), since $\boldsymbol{\eta}_i \mapsto \inf_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i')$ is a continuous function by Lemma 17, we obtain that (143) is also bounded by C'/n^3. This establishes Lemma 18. ■

Proof of result in (46): A sufficient condition to establish the convergence of expected stopping time is to show the second moment condition:

$$\limsup_{L\to\infty} E\left[\left(\frac{\tau(\pi_{SMF}(L,\gamma,\beta))}{\log(L)}\right)^2\right] < \infty.$$

We now proceed to establish this. Define

$$u(L) := \left(\frac{(1+\epsilon)\log((M-1)L)}{D^*(\overline{\boldsymbol{\eta}})\log(L)} + \frac{1}{\log(L)}\right)^2.$$

We then have

$$\begin{aligned}
&\limsup_{L\to\infty} E\left[\left(\frac{\tau(\pi_{SMF}(L,\gamma,\beta))}{\log(L)}\right)^2\right] \\
&= \limsup_{L\to\infty} \int_{x\geq 0} P\left(\frac{\tau(\pi_{SMF}(L,\gamma,\beta))}{\log(L)} > \sqrt{x}\right) dx \\
&\leq \limsup_{L\to\infty} \int_{x\geq 0} P\left(\tau^l(\pi_{SMF}(L,\gamma,\beta)) > \lfloor\sqrt{x}\log(L)\rfloor\right) dx,
\end{aligned}$$

which is upperbounded in inequalities (144)-(146), as shown at the bottom of the next page. The inequality in (145) is obtained using the fact that $P\left(\tau^l(\pi_{SMF}(L,\gamma,\beta)) > \lfloor\sqrt{x}\log(L)\rfloor\right)$ is constant in the interval

$$x \in \left[\left(\frac{n}{\log(L)}\right)^2, \left(\frac{n+1}{\log(L)}\right)^2\right];$$

for inequality (146), we have from Lemma 18 that

$$\text{for all } n \geq \frac{(1+\epsilon)\log((M-1)L)}{D^*(\overline{\boldsymbol{\eta}})}, \tag{147}$$

$P(Z_l(n) < \log((M-1)L)) < B/n^3$. This completes the proof.

C. Proof of (47)

To prove this, observe that

$$\begin{aligned}
&E[C(\pi_{SMF}(L,\gamma))] \\
&= E\Big[\tau(\pi_{SMF}(L,\gamma)) + \sum_{l=1}^{\tau(\pi_{SMF}(L,\gamma))-1} g(A_l, A_{l+1})\Big] \\
&\leq E[\tau(\pi_{SMF}(L,\gamma))] \\
&\qquad + g_{\max} E\Big[\sum_{l=1}^{\tau(\pi_{SMF}(L,\gamma))-1} 1_{\{A_l \neq A_{l+1}\}}\Big] \\
&\leq E[\tau(\pi_{SMF}(L,\gamma))] + g_{\max} E\Big[\sum_{l=1}^{\tau(\pi_{SMF}(L,\gamma))} U_{l+1}\Big] \\
&= E[\tau(\pi_{SMF}(L,\gamma))] + g_{\max}\gamma E[\tau(\pi_{SMF}(L,\gamma))] \\
&= E[\tau(\pi_{SMF}(L,\gamma))](1 + g_{\max}\gamma),
\end{aligned}$$

where in the penultimate equality, we have used Doob's optional stopping theorem. Divide by $\log L$ and let $L \to \infty$ to get the required result. This completes the proof of (47), completes the proof of all three results in the proposition, and thus finishes the proof of Proposition 7.

$$\begin{aligned}
&P\left(-\sup_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} \sum_{i=1}^K N_i^n \left(\boldsymbol{\eta}_i'^T\hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\boldsymbol{\eta}_i')\right) + \sup_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} n\sum_{i=1}^K \lambda_i^* \left(\boldsymbol{\eta}_i'^T\boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i')\right) < -\epsilon' n\right) \\
&= P\left(\sup_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} \sum_{i=1}^K \frac{N_i^n}{n}\left(\boldsymbol{\eta}_i'^T\hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\boldsymbol{\eta}_i')\right) - \sup_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} \sum_{i=1}^K \lambda_i^*\left(\boldsymbol{\eta}_i'^T\boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i')\right) > \epsilon'\right) \\
&= P\left(\sup_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} \sum_{i=1}^K \frac{N_i^n}{n}\left(-D(\hat{\boldsymbol{\eta}}_i \parallel \boldsymbol{\eta}_i') + \hat{\boldsymbol{\eta}}_i^T\hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\hat{\boldsymbol{\eta}}_i)\right) - \sup_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} \sum_{i=1}^K \lambda_i^*\left(-D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i') + \boldsymbol{\eta}_i^T\boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i)\right) > \epsilon'\right) \qquad (140) \\
&= P\left(\sum_{i=1}^K \frac{N_i^n}{n}\left(-\inf_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} D(\hat{\boldsymbol{\eta}}_i \parallel \boldsymbol{\eta}_i') + \hat{\boldsymbol{\eta}}_i^T\hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\hat{\boldsymbol{\eta}}_i)\right) - \sum_{i=1}^K \lambda_i^*\left(-\inf_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i') + \boldsymbol{\eta}_i^T\boldsymbol{\kappa}_i - \mathcal{A}_i(\boldsymbol{\eta}_i)\right) > \epsilon'\right) \qquad (141) \\
&= P\left(\sum_{i=1}^K \left(\frac{N_i^n}{n} - \lambda_i^*\right)\left(-\inf_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} D(\hat{\boldsymbol{\eta}}_i \parallel \boldsymbol{\eta}_i') + \hat{\boldsymbol{\eta}}_i^T\hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\hat{\boldsymbol{\eta}}_i)\right) - \sum_{i=1}^K \lambda_i^*\left(\inf_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} D(\hat{\boldsymbol{\eta}}_i \parallel \boldsymbol{\eta}_i') - \inf_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i')\right)\right. \\
&\qquad \left. + \sum_{i=1}^K \lambda_i^*\left(\hat{\boldsymbol{\eta}}_i^T\hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\hat{\boldsymbol{\eta}}_i) - \boldsymbol{\eta}_i^T\boldsymbol{\kappa}_i + \mathcal{A}_i(\boldsymbol{\eta}_i)\right) > \epsilon'\right) \qquad (142) \\
&\leq P\left(\sum_{i=1}^K \left(\frac{N_i^n}{n} - \lambda_i^*\right)\left(-\inf_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} D(\hat{\boldsymbol{\eta}}_i \parallel \boldsymbol{\eta}_i') + \hat{\boldsymbol{\eta}}_i^T\hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\hat{\boldsymbol{\eta}}_i)\right) - \sum_{i=1}^K \lambda_i^*\left(\inf_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} D(\hat{\boldsymbol{\eta}}_i \parallel \boldsymbol{\eta}_i') - \inf_{\overline{\boldsymbol{\eta}}'\in\Theta_{-l}} D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i')\right)\right. \\
&\qquad \left. + \sum_{i=1}^K \lambda_i^*\left(\hat{\boldsymbol{\eta}}_i^T\hat{\boldsymbol{\kappa}}_i - \mathcal{A}_i(\hat{\boldsymbol{\eta}}_i) - \boldsymbol{\eta}_i^T\boldsymbol{\kappa}_i + \mathcal{A}_i(\boldsymbol{\eta}_i)\right) > \epsilon', \left|\frac{N_{i'}^n}{n} - \lambda_{i'}^*\right| \leq \epsilon_1, \left\|\frac{\mathbf{Y}_{i'}^n}{N_{i'}^n} - \boldsymbol{\kappa}_{i'}\right\|_\infty \leq \epsilon_2, \forall i'\right) \\
&\qquad + \sum_{i'} P\left(\left|\frac{N_{i'}^n}{n} - \lambda_{i'}^*\right| > \epsilon_1\right) + \sum_{i'} P\left(\left\|\frac{\mathbf{Y}_{i'}^n}{N_{i'}^n} - \boldsymbol{\kappa}_{i'}\right\|_\infty > \epsilon_2\right). \qquad (143)
\end{aligned}$$

APPENDIX F
VERIFICATION OF CONTINUOUS SELECTION IN SOME EXAMPLES

In this section we shall show that the odd arm identification problem and the best arm identification problem admit continuous selections.

A. Odd Arm Identification Problem

In order to show that the correspondence $\boldsymbol{\lambda} \mapsto \boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}})$ admits a continuous selection, we show that the function $F(\cdot, \boldsymbol{\eta})$ defined in (21) is strictly concave for each $\boldsymbol{\eta} \in \Theta_l$. We begin with

$$\boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}}) = \underset{\boldsymbol{\lambda} \in \mathcal{P}(K)}{\operatorname{argmax}} F(\boldsymbol{\lambda}, \boldsymbol{\eta}) = \underset{\boldsymbol{\lambda} \in \mathcal{P}(K)}{\operatorname{argmax}} \inf_{\overline{\boldsymbol{\eta}}' \in \Theta_{-l}} \sum_{i=1}^{K} \lambda_i D(\boldsymbol{\eta}_i \parallel \boldsymbol{\eta}_i'). \tag{148}$$

Recall the example discussed in Section (III-B). We observe from [10] that $\boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}})$ is of the form

$$\left[\frac{1-\lambda_l}{K-1}, \cdots \frac{1-\lambda_l}{K-1}, \lambda_l, \frac{1-\lambda_l}{K-1}, \cdots, \frac{1-\lambda_l}{K-1}\right]$$

and the expression for $D^*(\cdot)$ in (18) can be reduced to

$$D^*(\overline{\boldsymbol{\eta}}) = \max_{\lambda_l \in [0,1]} \lambda_l D(\boldsymbol{\eta}_1 || \tilde{\boldsymbol{\eta}}) + (1-\lambda_l) \frac{K-2}{K-1} D(\boldsymbol{\eta}_2 || \tilde{\boldsymbol{\eta}}) \tag{149}$$

where $\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta}(\tilde{\boldsymbol{\kappa}})$ and

$$\tilde{\boldsymbol{\kappa}} = \frac{\lambda_l \boldsymbol{\kappa}_1 + (1-\lambda_l)\frac{K-2}{K-1}\boldsymbol{\kappa}_2}{\lambda_l + (1-\lambda_l)\frac{K-2}{K-1}}. \tag{150}$$

To establish the strict concavity, we show that the second derivative of the objective function in (148) is strictly negative for all values of λ_l. Using the result in (149), we redefine the objective function in (148) as

$$\Phi(\lambda_l) = \lambda_l D(\boldsymbol{\eta}_1 || \tilde{\boldsymbol{\eta}}) + (1-\lambda_l) \frac{K-2}{K-1} D(\boldsymbol{\eta}_2 || \tilde{\boldsymbol{\eta}}). \tag{151}$$

Taking the first derivative,

$$\begin{aligned} \frac{d\Phi}{d\lambda_l} &= D(\boldsymbol{\eta}_1 || \tilde{\boldsymbol{\eta}}) - \frac{K-2}{K-1} D(\boldsymbol{\eta}_2 || \tilde{\boldsymbol{\eta}}) \\ &\quad + \left[\lambda_l \nabla_{\tilde{\boldsymbol{\eta}}} D(\boldsymbol{\eta}_1 || \tilde{\boldsymbol{\eta}}) + (1-\lambda_l) \frac{K-2}{K-1} \nabla_{\tilde{\boldsymbol{\eta}}} D(\boldsymbol{\eta}_2 || \tilde{\boldsymbol{\eta}})\right]^T \frac{d\tilde{\boldsymbol{\eta}}}{d\lambda_l} \\ &= D(\boldsymbol{\eta}_1 || \tilde{\boldsymbol{\eta}}) - \frac{K-2}{K-1} D(\boldsymbol{\eta}_2 || \tilde{\boldsymbol{\eta}}). \end{aligned} \tag{152}$$

The equality in (152) follows from the fact that the $\boldsymbol{\eta}'$ that attains the infimum in (148) is $\tilde{\boldsymbol{\eta}}$. Differentiating again by applying chain rule and using the result that $\nabla_{\boldsymbol{\eta}_2} D(\boldsymbol{\eta}_1 || \boldsymbol{\eta}_2) = \boldsymbol{\kappa}_2 - \boldsymbol{\kappa}_1$ we get

$$\frac{d^2\Phi}{d\lambda_l^2} = \left[(\tilde{\boldsymbol{\kappa}} - \boldsymbol{\kappa}_1) - \frac{K-2}{K-1}(\tilde{\boldsymbol{\kappa}} - \boldsymbol{\kappa}_2)\right]^T \frac{d\tilde{\boldsymbol{\eta}}}{d\lambda_l}. \tag{153}$$

Observe that

$$\frac{d\tilde{\boldsymbol{\eta}}}{d\lambda_l} = D_{\tilde{\boldsymbol{\kappa}}}\tilde{\boldsymbol{\eta}} \cdot \frac{d\tilde{\boldsymbol{\kappa}}}{d\lambda_l} \tag{154}$$

$$= \mathsf{Hess}(\mathcal{F}(\tilde{\boldsymbol{\kappa}})) \cdot \frac{-1}{\lambda_l + \frac{K-2}{K-1}(1-\lambda_l)} \cdot \left((\tilde{\boldsymbol{\kappa}} - \boldsymbol{\kappa}_1) - \frac{K-2}{K-1}(\tilde{\boldsymbol{\kappa}} - \boldsymbol{\kappa}_2)\right). \tag{155}$$

Equality in (154) is obtained using chain rule for differentiation and $D_{\tilde{\boldsymbol{\kappa}}}\tilde{\boldsymbol{\eta}}$ is the matrix $\left(\frac{\partial}{\partial \tilde{\boldsymbol{\kappa}}_j}\tilde{\boldsymbol{\eta}}_i\right)_{1 \le i,j \le d}$. From (7), we recognise that $D_{\tilde{\boldsymbol{\kappa}}}\tilde{\boldsymbol{\eta}} = \mathsf{Hess}(F(\tilde{\boldsymbol{\kappa}}))$, the Hessian of the function $F(\boldsymbol{\kappa})$ with respect to $\boldsymbol{\kappa}$ evaluated at $\tilde{\boldsymbol{\kappa}}$. Using this and a straightforward calculation of the derivative $d\tilde{\boldsymbol{\kappa}}/d\lambda_l$, we get (155).

Substituting (155) in (153) and using the fact that the Hessian of the strictly convex function $\mathcal{F}(\cdot)$ is positive definite, we get the required inequality as

$$\frac{d^2\Phi}{d\lambda_l^2} < 0 \tag{156}$$

$$\begin{aligned} &\limsup_{L\to\infty} E\left[\left(\frac{\tau(\pi_{SMF}(L,\gamma,\beta))}{\log(L)}\right)^2\right] \\ &\le \limsup_{L\to\infty}\left[u(L) + \int_{x \ge u(L)} P\left(\tau^l(\pi_{SMF}(L,\gamma,\beta)) > \lfloor \sqrt{x}\log(L)\rfloor\right) dx\right] && (144) \\ &\le \left(\frac{1+2\epsilon}{D^*(\overline{\boldsymbol{\eta}})}\right)^2 + \limsup_{L\to\infty} \sum_{n \ge \lfloor\sqrt{u(L)}\log(L)\rfloor} \left(\left(\frac{n+1}{\log(L)}\right)^2 - \left(\frac{n}{\log(L)}\right)^2\right) P\left(\tau^l(\pi_{SMF}(L,\gamma,\beta)) > n\right) && (145) \\ &\le \left(\frac{1+2\epsilon}{D^*(\overline{\boldsymbol{\eta}})}\right)^2 + \limsup_{L\to\infty} \sum_{n \ge \lfloor\sqrt{u(L)}\log(L)\rfloor} \left(\frac{2n+1}{(\log(L))^2}\right) P\left(Z_l(n) < \log((M-1)L)\right) \\ &\le \left(\frac{1+2\epsilon}{D^*(\overline{\boldsymbol{\eta}})}\right)^2 + \limsup_{L\to\infty} \sum_{n \ge \lfloor\sqrt{u(L)}\log(L)\rfloor} \left(\frac{2n+1}{(\log(L))^2}\right) \frac{B}{n^3} \quad \text{(for } L \text{ sufficiently large)} && (146) \\ &< \infty; \end{aligned}$$

thereby completing the proof of strict concavity. Using this and the first sufficient condition for Assumption A to hold (see the first bullet after Assumption A), the correspondence $\boldsymbol{\lambda} \mapsto \boldsymbol{\lambda}^*(\overline{\boldsymbol{\eta}})$ admits a continuous selection.

B. Best Arm Identification Problem

That the continuous selection assumption, Assumption A, holds for this problem has been proved by Garivier and Kaufmann [9, Prop. 6.2].

ACKNOWLEDGMENT

The authors acknowledge fruitful discussions with Aditya O. Deshmukh.

REFERENCES

[1] H. Chernoff, "Sequential design of experiments," *Ann. Math. Statist.*, vol. 30, no. 3, pp. 755–770, 1959.
[2] A. E. Albert, "The sequential design of experiments for infinitely many states of nature," *Ann. Math. Statist.*, vol. 32, no. 3, pp. 774–799, Sep. 1961.
[3] N. K. Vaidhiyan and R. Sundaresan, "Learning to detect an oddball target," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 831–852, Feb. 2018.
[4] N. K. Vaidhiyan, S. P. Arun, and R. Sundaresan, "Neural dissimilarity indices that predict oddball detection in behaviour," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 4778–4796, Aug. 2017.
[5] D. Chen, Q. Huang, H. Feng, Q. Zhao, and B. Hu, "Active anomaly detection with switching cost," in *Proc. ICASSP—IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5346–5350.
[6] O. Dekel, J. Ding, T. Koren, and Y. Peres, "Bandits with switching costs: $T^{2/3}$ regret," in *Proc. 46th Annu. ACM Symp. Theory Comput.*, 2014, pp. 459–467.
[7] E. Kaufmann, O. Cappé, and A. Garivier, "On the complexity of best arm identification in multi-armed bandit models," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1–42, 2016.
[8] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. New York, NY, USA: Springer-Verlag, 1994.
[9] A. Garivier and E. Kaufmann, "Optimal best arm identification with fixed confidence," in *Proc. Conf. Learn. Theory*, 2016, pp. 998–1027.
[10] G. R. Prabhu, S. Bhashyam, A. Gopalan, and R. Sundaresan, "Learning to detect an anomalous target with observations from an exponential family," in *Proc. IEEE Data Sci. Workshop (DSW)*, Jun. 2019, pp. 88–92.
[11] G. R Prabhu, S. Bhashyam, A. Gopalan, and R. Sundaresan, "Learning to detect an oddball target with observations from an exponential family," 2017, *arXiv:1712.03682*.
[12] A. Deshmukh, S. Bhashyam, and V. V. Veeravalli, "Controlled sensing for composite multihypothesis testing with application to anomaly detection," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018, pp. 2109–2113.
[13] A. Deshmukh, V. V. Veeravalli, and S. Bhashyam, "Sequential controlled sensing for composite multihypothesis testing," *Sequential Anal.*, vol. 40, no. 2, pp. 259–289, Apr. 2021.
[14] S. Juneja and S. Krishnasamy, "Sample complexity of partition identification using multi-armed bandits," in *Proc. Conf. Learn. Theory*, A. Beygelzimer and D. Hsu, Eds., Phoenix, AZ, USA, vol. 99, Jun. 2019, pp. 1824–1852. [Online]. Available: http://proceedings.mlr.press/v99/juneja19a.html
[15] N. K. Vaidhiyan and R. Sundaresan, "Active search with a cost for switching actions," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2015, pp. 17–24.
[16] S. Krishnasamy, P. T. Akhil, A. Arapostathis, S. Shakkottai, and R. Sundaresan, "Augmenting max-weight with explicit learning for wireless scheduling with switching costs," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2017, pp. 352–360.
[17] M. I. Jordan and M. J. Wainwright, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, nos. 1–2, pp. 1–305, 2008.
[18] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[19] E. Kaufmann and S. Kalyanakrishnan, "Information complexity in bandit subset selection," in *Proc. Conf. Learn. Theory*, 2013, pp. 228–251.
[20] R. Degenne, W. M. Koolen, and P. Ménard, "Non-asymptotic pure exploration by solving games," 2019, *arXiv:1906.10431*.
[21] J. Scarlett, I. Bogunovic, and V. Cevher, "Overlapping multi-bandit best arm identification," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 2544–2548.
[22] R. Huang, M. M. Ajallooeian, C. Szepesvári, and M. Müller, "Structured best arm identification with fixed confidence," in *Proc. 28th Int. Conf. Algorithmic Learn. Theory*, S. Hanneke and L. Reyzin, Eds., vol. 76, Oct. 2017, pp. 593–616. [Online]. Available: https://proceedings.mlr.press/v76/huang17a.html
[23] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal sequential outlier hypothesis testing," *Sequential Anal.*, vol. 36, no. 3, pp. 309–344, Jul. 2017.
[24] A. Tajer, V. V. Veeravalli, and H. V. Poor, "Outlying sequence detection in large data sets: A data-driven approach," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 44–56, Sep. 2014.
[25] H. Zhuang, C. Wang, and Y. Wang, "Identifying outlier arms in multi-armed bandit," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5204–5213.
[26] A. Garivier, P. Ménard, and G. Stoltz, "Explore first, exploit next: The true shape of regret in bandit problems," *Math. Oper. Res.*, vol. 44, no. 2, pp. 377–399, May 2019.
[27] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*. Boston, MA, USA: Birkhaüser, 1990.
[28] R. K. Sundaram, *A First Course in Optimization Theory*. Cambridge, U.K.: Cambridge Univ. Press, 1996.
[29] Y. S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales* (Springer Texts in Statistics), 3rd ed. New York, NY, USA: Springer, 2003.

Gayathri R. Prabhu received the B.Tech. degree in electronics and communication engineering from the Amrita School of Engineering, India, in 2012, and the M.Tech. degree in digital signal processing from the Indian Institute of Space Science and Technology, India, in 2014. She is currently a Research Scholar with the Department of Electrical Engineering, IIT Madras, India. Her research interests include communication and information theory, signal processing, and detection and estimation theory.

Srikrishna Bhashyam (Senior Member, IEEE) received the B.Tech. degree in electronics and communication engineering from IIT Madras, India, in 1996, and the M.S. and Ph.D. degrees in electrical and computer engineering from Rice University, Houston, TX, USA, in 1998 and 2001, respectively. He was a Senior Engineer with Qualcomm Inc., Campbell, CA, USA, from 2001 to 2003, where he was involved in wideband code division multiple access modem design. Since 2003, he has been with IIT Madras. He is currently a Professor with the Department of Electrical Engineering. His research interests include communication and information theory, statistical signal processing, and wireless networks. He served as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2009 to 2014. He has been an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS since 2017.

Aditya Gopalan received the Ph.D. degree in electrical engineering from The University of Texas at Austin. He was a Post-Doctoral Fellow at the Technion—Israel Institute of Technology. He is currently an Associate Professor with the Department of Electrical Communication Engineering, Indian Institute of Science. His research interests include machine learning and statistical inference, control, and network algorithms.

Rajesh Sundaresan (Senior Member, IEEE) received the B.Tech. degree in electronics and communication from IIT Madras in 1994 and the M.A. and Ph.D. degrees in electrical engineering from Princeton University in 1996 and 1999, respectively. From 1999 to 2005, he has worked with Qualcomm Inc., on the design of communication algorithms for wireless modems. Since 2005, he has been with the Indian Institute of Science, where he is currently a Professor with the Department of Electrical Communication Engineering and an Associate Faculty with the Robert Bosch Centre for Cyber-Physical Systems. He has held visiting positions at Qualcomm Inc., the Coordinated Sciences Laboratory of the University of Illinois at Urbana–Champaign, the Toulouse Mathematical Institute, Strand Life Sciences, and the Indian Statistical Institute's Bengaluru Centre. His research interests include decision theory, communication, computation, and control over networks, cyber-social systems, and more recently data-driven decision frameworks for public health responses. He was an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY from 2012 to 2015.