Adv. Appl. Prob. **27**, 1054–1078 (1995)
Printed in N. Ireland
© *Applied Probability Trust* 1995

SAMPLE MEAN BASED INDEX POLICIES WITH $O(\log n)$ REGRET FOR THE MULTI-ARMED BANDIT PROBLEM

RAJEEV AGRAWAL, **University of Wisconsin-Madison*

Abstract

We consider a non-Bayesian infinite horizon version of the multi-armed bandit problem with the objective of designing simple policies whose *regret* increases slowly with time. In their seminal work on this problem, Lai and Robbins had obtained a $O(\log n)$ lower bound on the regret with a constant that depends on the Kullback–Leibler number. They also constructed policies for some specific families of probability distributions (including exponential families) that achieved the lower bound. In this paper we construct index policies that depend on the rewards from each arm only through their sample mean. These policies are computationally much simpler and are also applicable much more generally. They achieve a $O(\log n)$ regret with a constant that is also based on the Kullback–Leibler number. This constant turns out to be optimal for one-parameter exponential families; however, in general it is derived from the optimal one via a 'contraction' principle. Our results rely entirely on a few key lemmas from the theory of large deviations.

UPPER CONFIDENCE BOUNDS; ASYMPTOTICALLY EFFICIENT; LARGE DEVIATIONS; STOCHASTIC ADAPTIVE CONTROL

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 90B50

1. Introduction

A multi-armed bandit refers to an imagined slot machine with two or more arms. Successive plays of each arm j yield i.i.d. rewards with an unknown probability measure ρ_j. Rewards are independent across arms. The problem is to design a policy that, at each stage, chooses to play one of the arms with a view to maximize the expected sum of rewards. Equivalently, we may attempt to minimize the *regret* or *learning loss* which is defined as the difference between the maximum expected sum of rewards when the probability measure of each arm is known, and the expected sum of rewards actually obtained by a particular policy. In this paper, we consider a non-Bayesian infinite horizon version of the problem with the objective of designing policies whose regret increases slowly with time. This version of the multi-armed bandit problem was introduced by Lai and Robbins in [14], [15]. The regret is a much stronger objective function (than the average reward per unit time) and reveals the rôle of certain fundamental quantities such as the Kullback–Leibler number in this problem. While the multi-armed bandit problem is important in its

Received 13 August 1993; revision received 25 July 1994.

* Postal address: Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI 53706-1691, U.S.A. E-mail: agrawal@engr.wisc.edu

Research supported by NSF Grant No. ECS-8919818.

own right, it is also prototypical of a much larger class of stochastic adaptive control problems which have an intrinsic conflict between learning and control. The main ideas of Lai and Robbins' seminal papers, [14], [15], have since been exploited to address a host of related and some not so (apparently) related problems: see references [1]–[8].

The starting point of [15] is an asymptotic lower bound on the regret of $O(\log n)$ with a constant that depends on the Kullback–Leibler number. Thereafter, Lai and Robbins construct policies for certain families of distributions which achieve the lower bound. The policies they construct are of *index* type, i.e. at each time the policy computes an index for each arm based on the observations from that arm only and plays the arm with the highest index. Their indices make use of certain functions called *upper confidence bounds*. There are two major drawbacks of the upper confidence bounds constructed in the literature so far. They have been designed for only a limited class of families of probability distributions (mostly for one parameter exponential families, with some exceptions—([15], Example 5.4, [7]). Second, they are generally quite complicated and typically rely on the entire sequence of observations from the corresponding arm. This may impose severe computational and memory requirements on the policy, thereby making it hard, if not impossible, to implement it in real time. In an attempt to address this problem Agrawal and Teneketzis [4] constructed simple certainty equivalence with forcing policies. These policies were based only on sample means and an *a priori* specified sequence of forcing times, and achieved a regret of $o(f_n \log n)$ for any sequence $f_n \to \infty$ as $n \to \infty$. In this paper, we construct some simple index policies that depend on the rewards obtained from an arm only through their sample mean, and yet achieve a $O(\log n)$ regret.

In Section 2 we formulate the problem precisely and present the lower bound of Lai and Robbins (Theorem 2.1). Our point of departure is Theorem 2.2, which gives a general upper bound on the regret associated with a certain class of index policies that we refer to as upper confidence bound policies. This result essentially follows from the corresponding results of Lai and Robbins. However, it should be noted that the policies originally proposed in [14] and [15] were more complicated than the ones considered here which are closer to the ones used by Agrawal *et al.* [2]. Hence, the proof of the general upper bound is an adaptation of [2], Theorem 4.1.

In Section 3 we present some results from the theory of large deviations. The probabilistic component of all subsequent results of this paper is based entirely on these large deviations results.

In Section 4 we construct a class of index functions that depend on the sequence of observations from an arm only through the sample mean. In particular, they inflate the sample mean by an amount that depends on two quantities: an *a priori* specified sequence of numbers $\boldsymbol{a}$, and a function J that is derived from a 'contraction' of the Kullback–Leibler number. We establish that this class of indices are indeed 'upper confidence bounds' (Lemmas 4.11, 4.12), and we identify their corresponding logarithmic constant (Corollary 4.14). This allows us to use the general upper bound theorem obtained earlier to get an $O(\log n)$ upper bound on the regret (Theorem 4.10). While this is the best order as established by the general lower bound, it may not have the best constant. Thus the policies constructed in this paper may not be asymptotically optimal in general; but the loss in the constant may be a small price for getting a computationally easier policy.

In Section 5 we focus on three particular choices of the function J. The first of these turns out to be optimal for one-parameter exponential families. The second is inspired by one-parameter shifted families for which the first and second choice coincide. The third is based on a crude variance bound. We compute these functions and the corresponding upper confidence bounds for several examples.

2. The problem

There are $p \geqq 2$ arms. Successive plays of arm j $(j = 1, 2, \cdots, p)$, yield i.i.d. rewards $\{X_i^j\}_{i=1}^{\infty}$, with an unknown probability measure ρ_j. It is given that $\rho_j \in \mathcal{N} \subset \mathcal{M}$, where $\mathcal{M}$ is the set of all Borel probability measures on $\mathbb{R}$, and $\mathcal{N}$ is a known subset of probability measures with finite means. Rewards are independent across arms. Let $\boldsymbol{\rho} := (\rho_1, \cdots, \rho_p)$ denote the vector of probability measures of the p arms. A *policy* ϕ consists of a sequence of $\{1, \cdots, p\}$-valued random variables $\{\phi_n\}_{n=1}^{\infty}$, indicating which arm has been selected for play at stage n on the basis of all the past actions and past observations. That is, ϕ_n is a function of only the past actions $\phi_1, \cdots, \phi_{n-1}$ and the past rewards $Z_1, \cdots, Z_{n-1}$. Let

$$T_n(j) := \sum_{i=1}^{n} 1\{\phi_i = j\}$$

be the number of times arm j has been used up to stage n. Then, Z_n, the reward collected at stage n, can be expressed in terms of the individual reward sequences, $\{X_i^j\}_{i=1}^{\infty}$, from each arm j $(j = 1, 2, \cdots, p)$, as

$$Z_n := \sum_{j=1}^{p} 1\{\phi_n = j\} X^j_{T_n(j)} = \sum_{j=1}^{p} (T_n(j) - T_{n-1}(j)) X^j_{T_n(j)}.$$

Let

$$J_n := \sum_{i=1}^{n} Z_i = \sum_{j=1}^{p} \sum_{i=1}^{T_n(j)} X_i^j$$

be the sum of rewards collected up to stage n. Then, by Wald's equation (see [9], Problem 22.9),

$$E_{\boldsymbol{\rho}}^{\phi} J_n = \sum_{j=1}^{p} m_{\rho_j} E_{\boldsymbol{\rho}} T_n(j),$$

where

$$m_{\rho_j} := \int x \rho_j(dx)$$

is the mean reward from arm j. The problem is to find a policy ϕ which maximizes, in some sense, $E_{\boldsymbol{\rho}}^{\phi} J_n$ as $n \to \infty$. Let

$$m^*(\boldsymbol{\rho}) := \max_{j=1,\cdots,p} m_{\rho_j}$$

be the maximum mean reward of the p arms. Clearly, if the vector of probability measures $\boldsymbol{\rho}$ were known, then the optimal policy would be to always use an arm $j^*(\boldsymbol{\rho})$ such that $m_{j^*(\boldsymbol{\rho})} = m^*(\boldsymbol{\rho})$, for which

$$E_{\boldsymbol{\rho}}^{j^*(\boldsymbol{\rho})} J_n = nm^*(\boldsymbol{\rho}).$$

In the absence of the knowledge of $\boldsymbol{\rho}$, it is desirable to approach this performance as closely as possible. For this purpose define the *regret* (or *learning loss*) as

$$R_n^{\phi}(\boldsymbol{\rho}) := nm^*(\boldsymbol{\rho}) - E_{\boldsymbol{\rho}}^{\phi} J_n = \sum_{j=1}^{p} (m^*(\boldsymbol{\rho}) - m_{\rho_j}) E_{\boldsymbol{\rho}}^{\phi} T_n(j). \tag{2.1}$$

The objective is to design policies for which the regret increases slowly.

The Kullback–Leibler number and related quantities. The Kullback–Leibler number is a measure of the distance between two probability measures, which plays a fundamental role in the multi-armed bandit problem as formulated above. Recall that $\mathcal{M}$ is the set of all Borel probability measures on $\mathbb{R}$, and that $\mathcal{N} \subset \mathcal{M}$ is a subset of probability measures (with finite means) from which the arms are drawn. Let $\rho, \nu \in \mathcal{M}$. The Kullback–Leibler number is defined as

$$\mathcal{I}(\nu, \rho) := \begin{cases} \displaystyle\int \log \frac{d\nu}{d\rho} d\nu, & \text{if } \nu \ll \rho \text{ and } \displaystyle\int \left| \log \frac{d\nu}{d\rho} \right| d\nu < \infty, \\ \infty, & \text{otherwise.} \end{cases} \tag{2.2}$$

We also introduce several quantities related to the Kullback–Leibler number $\mathcal{I}$, which will be used for the rest of the paper. First let

$$m_\rho := \int x \rho(dx)$$

denote the mean corresponding to the measure ρ. Now for any $\rho, \nu \in \mathcal{M}$ and $x, z \in \mathbb{R}$, define the following

$$I(x, \rho) := \inf_{\substack{\nu \in \mathcal{M} \\ m_\nu = x}} \mathcal{I}(\nu, \rho) \tag{2.3}$$

$$I(\nu, z) := \inf_{\substack{\rho \in \mathcal{N} \\ m_\rho = z}} \mathcal{I}(\nu, \rho) \tag{2.4}$$

$$i(x, z) := \inf_{\substack{\rho \in \mathcal{N} \\ m_\rho = z}} \inf_{\substack{\nu \in \mathcal{M} \\ m_\nu = x}} \mathcal{I}(\nu, \rho) = \inf_{\substack{\rho \in \mathcal{N} \\ m_\rho = z}} I(x, \rho) = \inf_{\substack{\nu \in \mathcal{M} \\ m_\nu = x}} I(\nu, z) \tag{2.5}$$

where $\inf \varnothing := \infty$. Note that we have used I to denote two different functions, which we will usually distinguish by their arguments.

Lower bound.

Theorem 2.1 (Lai and Robbins [14], Theorem 5). *Let ϕ be any uniformly good policy, i.e. its regret satisfies*

$$R_n^\phi(\boldsymbol{\rho}) = o(n^a) \qquad \forall a > 0, \boldsymbol{\rho} \in \mathcal{N}^p. \tag{2.6}$$

Then

$$\liminf_{n \to \infty} \frac{R_n^\phi(\boldsymbol{\rho})}{\log n} \geqq \sum_{\substack{j = 1, \cdots, p \\ m_{\rho_j} < m^*(\boldsymbol{\rho})}} \frac{(m^*(\boldsymbol{\rho}) - m_{\rho_j})}{I(\rho_j, m^*(\boldsymbol{\rho})^+)} \qquad \forall \boldsymbol{\rho} \in \mathcal{N}^p$$

where $I(\nu, z^+) := \inf_{y > z} I(\nu, y)$.

Proof. Follows by an easy modification of the proof of Theorem 2 of Lai and Robbins [15].

Policies whose regret have the same asymptotic upper bound as (2.7) are called *asymptotically efficient.*

Index policies. Let $\boldsymbol{g} := \{g_{ni} : \mathbb{R}^i \to \mathbb{R} \cup \{\infty\},\ n = 1, 2, \cdots, ;\ i = 1, \cdots, n\}$ be a collection of Borel-measurable functions. In the first p stages, use each arm once. For $n \geqq p$, let $X_1^j, \cdots, X_{T_n(j)}^j$ be the sequence of rewards obtained from arm j up to stage n. Based on the rewards from each arm, compute an index $g_{n, T_n(j)}(X_1^j, \cdots, X_{T_n(j)}^j)$ and use the arm with the highest index at stage $n + 1$. Any such policy $\phi^{\boldsymbol{g}}$ is specified by the collection of *index* functions $\boldsymbol{g}$, and will be called an *index* policy. The nice feature of index policies is that they separate the observations from the various arms. However, it is not *a priori* clear what we are sacrificing by restricting attention to such policies.

Lai and Robbins [15], [14] showed that under certain assumptions on the set of probability measures $\mathcal{N}$, there exist asymptotically efficient index policies, and constructed several such candidates. Central to their results are the following conditions on the index functions $\boldsymbol{g}$ which give them the interpretation of *upper confidence bounds* (UCB).

A1. g_{ni} is non-decreasing in $n \geqq i$ for each fixed $i = 1, 2, \cdots$.

A2. Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables with common probability measure $\rho \in \mathcal{N}$. Then, for any $z < m_\rho$,

$$P_\rho(g_{ni}(X_1, \cdots, X_i) < z \text{ for some } i \leqq n) = o(n^{-1}). \tag{2.8}$$

We shall refer to any index policy, $\phi^{\boldsymbol{g}}$, whose index functions, $\boldsymbol{g}$, satisfy A1–A2 as upper confidence bound (UCB) policies. The following theorem provides an upper bound on the regret of any UCB policy.

Theorem 2.2. *Let $\phi^{\boldsymbol{g}}$ be any index policy with* $\boldsymbol{g}$ *satisfying conditions* A1–A2. *Then its regret satisfies*

$$\limsup_{n\to\infty} \frac{R_n^{\phi^{\boldsymbol{g}}}(\boldsymbol{\rho})}{\log n} \leqq \sum_{\substack{j=1,\cdots,p \\ m_{\rho_j} < m^*(\boldsymbol{\rho})}} \frac{(m^*(\boldsymbol{\rho}) - m_{\rho_j})}{K^{\boldsymbol{g}}(\rho_j, m^*(\boldsymbol{\rho}))} \tag{2.9}$$

for all $\boldsymbol{\rho} \in \mathcal{N}^p$, *and where* $K^{\boldsymbol{g}}(\rho, z) \in [0, \infty]$ *is defined as*

$$\frac{1}{K^{\boldsymbol{g}}(\rho, z)} := \inf_{\varepsilon > 0} \limsup_{n\to\infty} \frac{E_\rho[\sup\{1 \leqq i \leqq n : g_{ni}(X_1, \cdots, X_i) \geqq z - \varepsilon\}]}{\log n}. \tag{2.10}$$

Proof. Follows by an easy modification of the proof of Theorem 4.1 of Agrawal et al. [2]. In light of (2.1), it suffices to show that $\limsup_{n\to\infty} E_{\boldsymbol{\rho}}^{\phi} T_n(j)/\log n \leqq 1/K^{\boldsymbol{g}}(\rho_j, m^*(\boldsymbol{\rho}))$ for any $j = 1, \cdots, p$ with $m_{\rho_j} < m^*(\boldsymbol{\rho})$, $\boldsymbol{\rho} \in \mathcal{N}^p$. To this end we have

$$\begin{aligned}
T_n(j) &= 1 + \sum_{k=p+1}^{n} 1\{\phi_k^{\boldsymbol{g}} = j\} \\
&\leqq 1 + \sum_{k=p}^{n-1} 1\{g_{k,T_k(j)}(X_1^j, \cdots, X_{T_k(j)}^j) \geqq \max_{j'=1,\cdots,p} g_{k,T_k(j')}(X_1^{j'}, \cdots, X_{T_k(j')}^{j'})\} \\
&\leqq 1 + \sum_{k=p}^{n-1} 1\{g_{k,T_k(j)}(X_1^j, \cdots, X_{T_k(j)}^j) \geqq g_{k,T_k(j^*)}(X_1^{j^*}, \cdots, X_{T_k(j^*)}^{j^*})\} \\
&\leqq 1 + \sum_{k=p}^{n-1} 1\{g_{k,T_k(j)}(X_1^j, \cdots, X_{T_k(j)}^j \geqq m^*(\boldsymbol{\rho}) - \varepsilon\} \\
&\quad + \sum_{k=p}^{n-1} 1\{g_{k,T_k(j^*)}(X_1^{j^*}, \cdots, X_{T_k(j^*)}^{j^*}) < m^*(\boldsymbol{\rho}) - \varepsilon\} \\
&\leqq 1 + \sum_{k=p}^{n-1} 1\{g_{n,T_k(j)}(X_1^j, \cdots, X_{T_k(j)}^j) \geqq m^*(\boldsymbol{\rho}) - \varepsilon\} \\
&\quad + \sum_{k=p}^{n-1} 1\{g_{k,T_k(j^*)}(X_1^{j^*}, \cdots, X_{T_k(j^*)}^{j^*}) < m^*(\boldsymbol{\rho}) - \varepsilon\}
\end{aligned}$$

$$\leqq 1+\sup\{1\leqq i\leqq n: g_{ni}(X_1^j,\cdots,X_i^j)\geqq m^*(\boldsymbol{\rho})-\varepsilon\}$$
$$+\sum_{k=1}^{n} 1\{g_{ki}(X_1^{j^*},\cdots,X_i^{j^*})<m^*(\boldsymbol{\rho})-\varepsilon \text{ for some } 1\leqq i\leqq k\}.$$

Note that we have used assumption A1 in the fourth inequality. The theorem now follows by taking expectation, dividing by log n, taking the lim sup as $n\to\infty$, and infimum over $\varepsilon>0$. Observe that the last term above vanishes by A2.

Lai and Robbins considered only those upper confidence bounds $\boldsymbol{g}$ and families of probability measures $\mathcal{N}$ for which the logarithmic rate function $K^{\boldsymbol{g}}$ satisfies

$$K^{\boldsymbol{g}}(\rho, m_{\rho^*})=I(\rho, m_{\rho^*})=I(\rho, m_{\rho^*}^+) \tag{2.11}$$

for all $\rho,\rho^*\in\mathcal{N}$ with $m_\rho<m_{\rho^*}$, where $I(\rho, z^+):=\inf_{y>z} I(\rho, y)$. In that case, by Theorems 2.1 and 2.2 it follows that the corresponding UCB policy, $\phi^{\boldsymbol{g}}$, is asymptotically efficient.

In this paper, we focus on identifying certain classes of index functions that are easy to compute. In particular we will restrict attention to those functions that depend on the observations only through the sample mean. Thus, with some abuse of notation we may write $g_{ni}(X_1,\cdots,X_i)=g_{ni}(\bar{X}_i)$ where $\bar{X}_i:=(X_1+\cdots+X_i)/i$.

3. Large deviation preliminaries

In this section we present some large deviation results that will be used crucially in the following sections. We begin with some necessary definitions. For any probability measure $\rho\in\mathcal{M}$, let

$$m_\rho:=\int x\rho(dx), \tag{3.1}$$

$$c_\rho(\theta):=\log\int e^{\theta x}\rho(dx), \tag{3.2}$$

$$N_\rho:=\{\theta\in\mathbb{R};\ c_\rho(\theta)<\infty\}, \tag{3.3}$$

$$I_\rho(x):=\sup_{\theta\in\mathbb{R}}\{\theta x-c_\rho(\theta)\}, \tag{3.4}$$

$$S_\rho:=\{x\in\mathbb{R}:\rho((x-\varepsilon,x+\varepsilon))>0,\ \forall\varepsilon>0\}. \tag{3.5}$$

Thus, m_ρ is the mean, c_ρ the *cumulant generating function*, N_ρ the domain of c_ρ, I_ρ the *convex-conjugate* (or *Legendre–Fenchel transform*) of c_ρ, and S_ρ is the *support* of ρ. Let int N_ρ denote the interior of N_ρ. Also, let convS_ρ denote the convex hull of S_ρ and int(conv S_ρ) denote the interior of the convex hull. For $\theta\in N_\rho$, define the *twisted* measure ρ_θ as

$$\rho_\theta(dx):=\exp\{\theta x-c_\rho(\theta)\}\rho(dx). \tag{3.6}$$

It is a well-known fact that N_ρ is a convex set and c_ρ is convex on N_ρ (see [10], Theorem 1.13, p. 19), and that all derivatives of c_ρ exist on $\operatorname{int} N_\rho$ (see [10], Theorem 2.2, p. 34). Let $\dot{c}_\rho$ and $\ddot{c}_\rho$ denote its first and second derivatives, respectively. The cumulant generating function, c_ρ, is called *steep* if for any endpoint $\lambda \in N_\rho - \operatorname{int} N_\rho$, $|\dot{c}_\rho(\theta)| \to \infty$ as $\theta \to \lambda$, $\theta \in \operatorname{int} N_\rho$. The Kullback–Leibler number also plays a fundamental role in the theory of large deviations, and in keeping with the notation in that literature we let

$$(3.7) \qquad \mathcal{I}_\rho(\nu) := \mathcal{I}(\nu, \rho) \qquad \forall \rho, \nu \in \mathcal{M}.$$

Lemma 3.1 (contraction principle). *Assume that c_ρ is steep. Then the following hold*:

1. *For each point $x \in$* int (conv S_ρ), *there exists a unique point $\theta \in$* int N_ρ *such that $m_{\rho_\theta} = \dot{c}_\rho(\theta) = x$. $\mathcal{I}_\rho(\nu)$ attains its infimum over the set $\{\nu \in \mathcal{M},\ m_\nu = x\}$ at the unique measure ρ_θ and*

$$(3.8) \qquad I_\rho(x) = \mathcal{I}_\rho(\rho_\theta) = \inf\{\mathcal{I}_\rho(\nu) : \nu \in M,\ m_\nu = x\} < \infty.$$

2. *For $x \notin$* conv S_ρ,

$$(3.9) \qquad I_\rho(x) = \inf\{\mathcal{I}_\rho(\nu) : \nu \in \mathcal{M},\ m_\nu = x\} = \infty.$$

3. *Suppose that* convS_ρ *has a finite endpoint a. If ρ has an atom at a $(p(\{a\}) > 0)$, then for $x = a$* (3.8) *is valid with ρ_θ replaced by δ_a, the Dirac measures at a. If ρ does not have an atom at a, then* (3.9) *is valid for $x = a$.*

Proof. Follows from a modification of the proof of Theorem VII.3.1, p. 254, Ellis [12]. That theorem requires $N_\rho = \mathbb{R}$. However, as pointed out on p. 265 of [12], this condition can be relaxed. Theorem 3.6, p. 75, Brown [10] provides the requisite result to do this. In particular, it establishes that $\dot{c}_\rho : \operatorname{int} N_\rho \to \operatorname{int}(\operatorname{conv} S_\rho)$ is one-to-one and onto.

Lemma 3.2 (large deviation upper bound). *Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables in $\mathbb{R}$, and let ρ be their common distribution. Let $S_n = X_1 + \cdots + X_n$. Then,*

$$(3.10) \qquad P_\rho\left(\frac{S_n}{n} \in F\right) \leqq 2 \exp\{-n \inf_{x \in F} I_\rho(x)\}$$

for all closed $F \subset \mathbb{R}$.

Proof. See Theorem 2.2.3, p. 27, Dembo and Zeitouni [11].

Lemma 3.3. *$I_\rho(x) \geqq 0$ is non-increasing in $x \leqq m_\rho$ with $I_\rho(m_\rho) = 0$. Hence, $i(x, z) \geqq 0$ defined in* (2.5) *is non-increasing in $x \leqq z$ with $i(z, z) = 0$ for each fixed $z \in \mathbb{R}$.*

Proof. See Lemma 2.2.5, p. 27, Dembo and Zeitouni [11].

4. Sample mean based upper confidence bounds

In this section we construct a family of index functions g that depend on the reward sequence only through their sample mean and show that they are *upper confidence bounds,* in that they satisfy conditions A1–A2. We also identify their corresponding logarithmic rate function $K^g(\rho, z)$. These functions will be based on two quantities a and J that are introduced below.

Let $J:\mathbb{R}\times\mathbb{R}\rightarrow[0,\infty]$ be any function satisfying the following three properties:

P1 $J(x, z)\leqq\inf_{y>z} i(x, y)$ for all $x\leqq z\in\mathbb{R}$, where $i(x, z)$ is defined in (2.5).

P2 $J(x, z)$ is non-increasing in $x\leqq z$ for each fixed $z\in\mathbb{R}$.

P3 $J(x, z)$ is non-decreasing in $z\geqq x$ for each fixed $x\in\mathbb{R}$.

Below we give three examples of functions that satisfy P1–P3.

Example 4.1

$$J^1(x, z):=\inf_{y>z} i(x, y).$$

Example 4.2

$$J^2(x, z):=J^2(z-x):=\inf_{v-u=z-x} J^1(u, v).$$

Example 4.3

$$J^3(x, z):=J^3(z-x):=\begin{cases}(z-x)^2/2\sigma^2, & (z-x)\leqq D,\\ \infty, & \textit{otherwise,}\end{cases}$$

where

$$D:=\sup_{\rho\in\mathcal{N}}(m_\rho-\inf S_\rho), \tag{4.1}$$

$$\sigma^2:=\sup_{\rho\in\mathcal{N}}\sup_{t\leqq 0}\ddot{c}_\rho(t). \tag{4.2}$$

Note that we have abused the notation J^2 and J^3 in Examples 4.2 and 4.3. Also note that properties P2 and P3 are equivalent for those two examples. Based on Lemma 3.3, it is easy to check that Examples 4.1 and 4.2 satisfy conditions P1-P3. That Example 4.3 also satisfies P1–P3, follows from Proposition 4.4 below.

Proposition 4.4. *Let J^2, J^3 be the functions defined in Examples* 4.2, 4.3 *above. Then, for all $\delta\geqq 0$, $J^2(\delta)\geqq J^3(\delta)$.*

Proof.

$$J^2(\delta) := \inf_{z-x=\delta}\ \inf_{\rho\in\mathcal{N}, m_\rho>z}\ \inf_{\nu\in\mathcal{M}, m_\nu=x} \mathcal{I}(\nu,\rho)$$

$$= \inf_{\rho\in\mathcal{N}}\ \inf_{\nu\in\mathcal{M}, m_\nu<m_\rho-\delta} \mathcal{I}(\nu,\rho)$$

$$\geqq \inf_{\rho\in\mathcal{N}} I_\rho(m_\rho-\delta).$$

By the contraction principle (Lemma 3.1), $I_\rho(m_\rho-\delta)=\infty$ for $\delta>m_\rho-\inf S_\rho$. Otherwise,

$$I_\rho(m_\rho-\delta) = \sup_{t\in\mathbb{R}} \{(m_\rho-\delta)t - c_\rho(t)\}$$

$$= \sup_{t\leqq 0} \{(m_\rho-\delta)t - c_\rho(t)\}$$

$$\geqq \sup_{t\leqq 0} \{(m_\rho-\delta)t - [c_\rho(0)+\dot{c}_\rho(0)t+\sigma^2t^2/2]\}$$

$$= \sup_{t\leqq 0} \{-\delta t+\sigma^2t^2/2\}$$

$$= \delta^2/2\sigma^2.$$

The last inequality above follows from Taylor's theorem and the definition of σ^2 as an upper bound on $\ddot{c}_\rho$. The theorem now follows by combining the above.

The significance of the first two examples is brought out by Theorems 5.1, 5.3 and Corollary 5.4. As will be seen from Theorem 4.10, the regret of corresponding upper confidence bound policies is inversely proportional to the function J. Therefore, it will be desirable to choose the largest J that satisfies P3. In this sense J^1 is the best. However, in many instances it may be preferable to choose a smaller J, if the corresponding upper confidence bounds are computationally simpler. For this reason we consider J^2 and J^3 in Section 5.1.

Also, let $\boldsymbol{a} := \{a_{ni} : n=1,2,\cdots;\ i=1,\cdots,n\}$ be a collection of non-negative numbers satisfying the following conditions:

C1 a_{ni} is non-decreasing in $n\geqq i$, for each fixed $i=1,2,\cdots$.

C2 For any $c>0$,

$$\sum_{i=1}^{c\log n} \exp\{-ia_{ni}\} = o(n^{-1}).$$

C3 For any $I\in[0,\infty]$, let

$$l_n(I) := \sup\{1\leqq i\leqq n : a_{ni}>I\}.$$

Then,

$$\limsup_{n\to\infty} \frac{l_n(I)}{\log n} \leqq \frac{1}{I}.$$

The following examples of $\boldsymbol{a}$ satisfy the above conditions (see Appendix).

Example 4.5

$$a_{ni} := \frac{\log n + \log\log n + b_n}{i}$$

for any sequence $\{b_n\}_{n=1}^{\infty}$ *such that* $b_n \nearrow \infty$ *and* $b_n/\log n \to 0$ *as* $n \to \infty$.

Example 4.6

$$a_{ni} := \frac{\log n}{i} + \left(\frac{b_n}{i}\right)^{\alpha}$$

for any $0 < \alpha < 1$, *and for any sequence* $\{b_n\}_{n=1}^{\infty}$ *of non-negative numbers such that* $b_n \nearrow \infty$ *and* $b_n/\log n \to 0$ *as* $n \to \infty$.

Remark. From the Appendix, we can in fact observe that the stronger condition $\sum_{i=1}^{n} \exp\{-ia_{ni}\} = o(n^{-1})$ holds, instead of C2, for Example 4.6.

Definition 4.7. *Let* $\boldsymbol{a} = \{a_{ni} : n = 1, 2, \cdots; \ i = 1, \cdots, n\}$ *be any sequence of non-negative numbers and let* $J : \mathbb{R}^2 \to [0, \infty]$ *be any Borel-measurable function. Define the family of functions* $\boldsymbol{g}^{\boldsymbol{a},J} := \{g_{ni}^{\boldsymbol{a},J} : \mathbb{R} \to \mathbb{R} \cup \{\infty\} : n = 1, 2, \cdots; \ i = 1, \cdots, n\}$ by

$$g_{ni}^{\boldsymbol{a},J}(x) := \inf\{z \geqq x : J(x, z) \geqq a_{ni}\} \tag{4.3}$$

where $\inf \varnothing := \infty$. *Throughout the sequel we will assume that* J *satisfies* P1–P3 *and* $\boldsymbol{a}$ *satisfies* C1–C3. *The sample mean based policy with* $\boldsymbol{g}^{\boldsymbol{a},J}$ *as the index functions will be denoted by* $\phi^{\boldsymbol{a},J}$.

Remark 4.8. *For* $J = J^2$ *of Example* 4.2, $\boldsymbol{g}^{\boldsymbol{a},J^2}$ *reduces to*

$$g_{ni}^{\boldsymbol{a},J^2}(x) = x + \inf\{y \geqq 0 : J^2(y) \geqq a_{ni}\}.$$

Remark 4.9. *For* $J = J^3$ *of Example* 4.3, $\boldsymbol{g}^{\boldsymbol{a},J^3}$ *reduces to*

$$g_{ni}^{\boldsymbol{a},J^3}(x) = x + \inf\{y \geqq 0 : J^3(y) \geqq a_{ni}\} = x + \min\{(2a_{ni})^{\frac{1}{2}}\sigma, D\}.$$

Theorem 4.10. *The sample mean based indices* $g^{\boldsymbol{a},J}$ *(with* J *satisfying* P1–P3 *and* $\boldsymbol{a}$ *satisfying* C1–C3*) are upper confidence bounds, i.e. they satisfy conditions* A1–A2. *Moreover, their corresponding logarithmic rate function* $K^{\boldsymbol{a},J}(\rho, z)$ *is bounded by*

$$K^{\boldsymbol{a},J}(\rho, z) \geqq J(m_\rho^+, z^-) := \sup_{m_\rho < w < y < z} J(w, y) = \lim_{w \searrow m_\rho, \, y \nearrow z} J(w, y) \tag{4.4}$$

for all $\rho \in \mathcal{N}$ and $z > m_\rho$. Consequently, the regret of any sample mean based index policy $\phi^{a,J}$ has the following asymptotic upper bound:

$$\limsup_{n\to\infty} \frac{R_n^{\phi^{a,J}}(\boldsymbol{\rho})}{\log n} \leqq \sum_{j=1}^{p} \frac{(m_{\rho_{j^*(\rho)}} - m_{\rho_j})}{J(m_{\rho_j}^{+}, m_{\rho_{j^*(\rho)}}^{-})}. \tag{4.5}$$

Proof. The proof follows immediately from Theorem 2.2 and Lemmas 4.11, 4.12, and Corollary 4.14.

Lemma 4.11. *$g_{ni}^{a,J}(x)$ is non-decreasing in $n \geqq i$ for each fixed $i = 1, 2, \cdots$ and $x \in \mathbb{R}$.*

Proof. Follows easily from the definition of the function $g_{ni}^{a,J}$ and condition C1 on the numbers a_{ni}.

Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables in $\mathbb{R}$, and let $\rho \in \mathcal{N}$ be their common distribution. let $\bar{X}_i := (X_1 + \cdots + X_i)/i$.

Lemma 4.12. *Let J satisfy P1–P3 and let $\boldsymbol{a}$ satisfy C1–C3. Then, the sample mean based indices $\boldsymbol{g}^{a,J}$ satisfy*

$$P_\rho(g_{ni}^{a,J}(\bar{X}_i) < z \text{ for some } i \leqq n) = o(n^{-1}) \tag{4.6}$$

for any $z < m_\rho$, $\rho \in \mathcal{N}$.

Proof. By the definition of the functions $g_{ni}^{a,J}$ and property P3 of J, it follows that

$$\begin{aligned}
&P_\rho(g_{ni}^{a,J}(\bar{X}_i) < z \text{ for some } i \leqq n)\\
&\quad \leqq \boldsymbol{P}_\rho(\bar{X}_i < z \text{ and } J(\bar{X}_i, z) \geqq a_{ni} \text{ for some } i \leqq n)\\
&\quad \leqq \sum_{i=1}^{n} \boldsymbol{P}_\rho(\bar{X}_i < z \text{ and } J(\bar{X}_i, z) \geqq a_{ni})\\
&\quad \leqq \sum_{i=1}^{c\log n} \boldsymbol{P}_\rho(\bar{X}_i < z \text{ and } J(\bar{X}_i, z) \geqq a_{ni}) + \sum_{i=c\log n+1}^{\infty} \boldsymbol{P}_\rho(\bar{X}_i < z)\\
&\quad \leqq \sum_{i=1}^{c\log n} \boldsymbol{P}_\rho(\bar{X}_i < z \text{ and } i(\bar{X}_i, m_\rho) \geqq a_{ni}) + \sum_{i=c\log n+1}^{\infty} \boldsymbol{P}_\rho(\bar{X}_i < z) \quad \text{(by P1)}\\
&\quad \leqq \sum_{i=1}^{c\log n} \boldsymbol{P}_\rho(i(\bar{X}_i, m_\rho) \geqq a_{ni}) + \sum_{i=c\log n+1}^{\infty} \boldsymbol{P}_\rho(\bar{X}_i < z)\\
&\quad \leqq \sum_{i=1}^{c\log n} \boldsymbol{P}_\rho(I_\rho(\bar{X}_i) \geqq a_{ni}) + \sum_{i=c\log n+1}^{\infty} \boldsymbol{P}_\rho(\bar{X}_i < z) \quad \text{(by (2.5))}\\
&\quad \leqq \sum_{i=1}^{c\log n} 2\exp\{-ia_{ni}\} + \sum_{i=c\log n+1}^{\infty} 2\exp\{-iI_\rho(z)\} \quad \text{(by Lemmas 3.2, 3.3)}
\end{aligned}$$

$$\begin{aligned}
&\leqq 2\sum_{i=1}^{c\log n}\exp\{-ia_{ni}\}+2\frac{\exp\{-(c\log n+1)I_\rho(z)\}}{1-\exp\{-I_\rho(z)\}}\\
&=2\sum_{i=1}^{c\log n}\exp\{-ia_{ni}\}+2n^{-cI_\rho(z)}\frac{\exp\{-I_\rho(z)\}}{1-\exp\{-I_\rho(z)\}}\\
&=o(n^{-1})\qquad \forall c>1/I_\rho(z)\qquad \text{(by C2).}
\end{aligned}$$

Lemma 4.13. *Let J satisfy P1–P3 and let $\boldsymbol{a}$ satisfy C1–C3. Then, the sample mean based indices $\mathbf{g}^{a,J}$ satisfy*

$$\limsup_{n\to\infty}\frac{E_\rho[\sup\{1\leqq i\leqq n: g_{ni}^{a,J}(\bar{X}_i)>z\}]}{\log n}\leqq\frac{1}{J(m_\rho^+,z)}\tag{4.7}$$

for any $z>m_\rho$, where $J(x^+,z):=\sup_{x<y<z}J(y,z)=\lim_{y\searrow x}J(y,z)$ for $x<z$.

Proof. Let

$$L_n(z):=\sup\{1\leqq i\leqq n: g_{ni}^{a,J}(\bar{X}_i)>z\}.$$

Then, by the definition of g_{ni}'s, it follows that for any $y\leqq z$,

$$\begin{aligned}
L_n(z)&\leqq\sup\{1\leqq i\leqq n:\bar{X}_i>z \text{ or } J(\bar{X}_i,z)<a_{ni}\}\\
&\leqq\sup\{1\leqq i\leqq n:\bar{X}_i>y\}+\sup\{1\leqq i\leqq n:\bar{X}_i\leqq y\text{ and } J(\bar{X}_i,z)<a_{ni}\}\\
&\leqq\sup\{1\leqq i:\bar{X}_i>y\}+\sup\{1\leqq i\leqq n:\bar{X}_i\leqq y\text{ and } J(y,z)<a_{ni}\}\\
&\leqq\sup\{1\leqq i:\bar{X}_i>y\}+\sup\{1\leqq i\leqq n:J(y,z)<a_{ni}\}\\
&=L(y)\text{ (say)}+l_n(J(y,z)).
\end{aligned}$$

The third inequality follows from property P2 of the function J. Now by the Lemma 3.2 (large deviation upper bound) and Lemma 3.3, it follows that for any $y>m_\rho$,

$$\begin{aligned}
E_\rho[L(y)]\leqq\sum_{i=1}^{\infty}P_\rho(\bar{X}_i\geqq y)\leqq\sum_{i=1}^{\infty}2\exp\left\{-i\inf_{x\geqq y}I_\rho(x)\right\}\\
=\sum_{i=1}^{\infty}2\exp\{-iI_\rho(y)\}\\
<\infty.
\end{aligned}$$

The second term, $l_n(J(y,z))$, is deterministic, and by condition C3 we have

$$\limsup_{n\to\infty}\frac{E_\rho[L_n(z)]}{\log n}\leqq\frac{1}{J(y,z)}.$$

The lemma now follows by taking the infimum over $m_\rho<y<z$. Note that $J(x^+,z):=\sup_{x<y<z}J(y,z)=\lim_{y\searrow x}J(y,z)$ for $x<z$ by property P2.

Corollary 4.14. *Let J satisfy P1–P3 and let $\mathbf{a}$ satisfy C1–C3. Then, the sample mean based indices $\mathbf{g}^{a,J}$ satisfy*

$$\inf_{\varepsilon>0} \limsup_{n\to\infty} \frac{E_\rho[\sup\{1 \leqq i \leqq n : g_{ni}^{a,J}(\bar{X}_i) \geqq z - \varepsilon\}]}{\log n} \leqq \frac{1}{J(m_\rho^+, z^-)} \tag{4.8}$$

for any $z > m_\rho$, where $J(x^+, z^-) := \sup_{x<w<y<z} J(w, y) = \lim_{w\searrow x, y\nearrow z} J(w, y)$.

5. One-parameter families of probability measures

In this section, we investigate the sample mean based indices $g^{a,J}$ and the corresponding policies $\phi^{a,J}$ for the three specific choices of the functions J given in Examples 4.1, 4.2, and 4.3. For this we need to specify the set of probability measures $\mathcal{N}$ from which the arms come. We will be particularly interested in one-parameter exponential families, i.e. when $\mathcal{N} = \mathcal{N}_\rho^e(\Theta)$ is given by the following. Let $\rho \in \mathcal{M}$ be some Borel probability measure on $\mathbb{R}$, and let $c = c_\rho$ be its cumulant generation function (3.2) with N_ρ (3.3) as its domain. Let ρ_θ be the exponentially twisted measures defined in (3.6). Then,

$$\mathcal{N}_\rho^e(\Theta) := \{\rho_\theta : \theta \in \Theta\} \tag{5.1}$$

where $\Theta \subset \operatorname{int} N_\rho$ is the parameter set of interest. When dealing with such a parametric family, it will be notationally convenient to identify the probability measure ρ_θ corresponding to a parameter $\theta \in \Theta$ with that parameter itself. We will need the following facts about exponential families, some of which were also mentioned in Section 3. N_ρ is an interval, and for all $\theta, \lambda \in \operatorname{int} N_\rho$,

$$m_\theta := m_{\rho_\theta} = \dot{c}(\theta) \tag{5.2}$$

$$\operatorname{var}_\theta := \operatorname{var}_{\rho_\theta} := \int (x - m_\theta)^2 \rho_\theta(dx) = \ddot{c}(\theta) > 0 \tag{5.3}$$

$$c_\theta(t) := c_{\rho_\theta}(t) = c(t + \theta) - c(\theta) \tag{5.4}$$

$$\mathcal{I}(\theta, \lambda) := \mathcal{I}(\rho_\theta, \rho_\lambda) = (\theta - \lambda)\dot{c}(\theta) - (c(\theta) - c(\lambda)) \tag{5.5}$$

$$= \int_\theta^\lambda (\lambda - t)\ddot{c}(t)dt \tag{5.6}$$

where $\dot{c}$ and $\ddot{c}$ are the first and second derivatives of c. That N_ρ is an interval and (5.2) and (5.3) are well known (see Brown [10], Theorems 1.13, 2.2); (5.4) and (5.5) follow easily from their definitions and the above. Note that since $\ddot{c}(\theta) = \operatorname{var}_\theta > 0$, it follows that the mapping $\theta \mapsto m_\theta = \dot{c}(\theta)$ is strictly increasing and hence one-to-one on $\operatorname{int} N_\rho$. Let $\dot{c}^{-1}$ denote its inverse on the range $\dot{c}(\operatorname{int} N_\rho)$. Recall from the proof of Lemma 3.1, that the range $\dot{c}(\operatorname{int} N_\rho) = \operatorname{int}(\operatorname{conv} S_\rho)$, where S_ρ is the support of ρ, if c_ρ is steep.

Theorem 5.1. *Let ρ be such that c_ρ is steep. Let $\mathcal{N}_\rho^e(\Theta)$ be a one-parameter* *exponential family with parameter set $\Theta \subset \operatorname{int} N_\rho$ which is dense from the right, i.e. it satisfies*:

$$(5.7) \qquad \forall \theta \in \Theta, \qquad \exists \lambda_n \in \Theta \cap (\theta, \infty),\, n = 1, 2, \cdots \quad \ni \lim_{n\to\infty} \lambda_n = \theta.$$

Let J^1 be given by Example 4.1. *Then, $J^1(x, z) = \infty$ if $z \geqq \sup \dot{c}(\Theta)$. Otherwise, for $z < \sup \dot{c}(\Theta)$ let $z^+ := \inf((z, \infty) \cap \dot{c}(\Theta))$. Then for $x < z$,*

$$(5.8) \quad J^1(x, z) = \begin{cases} \mathcal{I}(\dot{c}^{-1}(x), \dot{c}^{-1}(z^+)) & x \in \dot{c}(\operatorname{int} N_\rho), \\ c(\dot{c}^{-1}(z^+)) - \dot{c}^{-1}(z^+)x - \log \rho(\{x\}) & x = \inf(\dot{c}(\operatorname{int} N_\rho)) \textit{ and } \rho(\{x\}) > 0, \\ \infty & \textit{otherwise.} \end{cases}$$

*Moreover, for any **a** satisfying conditions C*1–*C*3,

$$(5.9) \quad K^{a,J^1}(\mu, m_{\mu^*}) = J^1(m_\mu^+, m_{\mu^*}^-) = J^1(m_\mu, m_{\mu^*}) = J^1(m_\mu, m_{\mu^*}^+) = I(\mu, m_{\mu^*}^+)$$

for all $\mu, \mu^* \in \mathcal{N}_\rho^e(\Theta)$ such that $m_\mu < m_{\mu^*}$. Consequently, the sample mean based index policy ϕ^{a,J^1} is asymptotically efficient.

Proof. Recall that

$$J^1(x, z) = \inf_{\nu \in \mathcal{N}, m_\nu > z} \; \inf_{\pi \in \mathcal{M}, m_\pi = x} \mathcal{I}(\pi, \nu),$$

with the convention that $\inf \varnothing = \infty$. Hence, for $z \geqq \sup \dot{c}(\Theta)$, $\{\nu \in \mathcal{N}_\rho^e(\Theta) : m_\nu > z\} = \varnothing$, and consequently $J^1(x, z) = \infty$. Now assume that $z < \sup \dot{c}(\Theta)$. Then by the contraction principle (Lemma 3.1), for any $\nu = \rho_\lambda \in \mathcal{N}_\rho^e(\Theta)$, we have

$$\inf_{\pi \in \mathcal{M}, m_\pi = x} \mathcal{I}(\pi, \nu) = \begin{cases} \mathcal{I}(\rho_{\dot{c}^{-1}(x)}, \rho_\lambda) & \text{if } x \in \dot{c}(\operatorname{int} N_\rho), \\ \mathcal{I}(\delta_x, \rho_\lambda) & \text{if } x = \inf(\dot{c}(\operatorname{int} N_\rho)) \text{ and } \rho(\{x\}) > 0, \\ \infty & \text{if } x = \inf(\dot{c}(\operatorname{int} N_\rho)) \text{ and } \rho(\{x\}) = 0, \\ \infty & \text{if } x < \inf(\dot{c}(\operatorname{int} N_\rho)). \end{cases}$$

Thus, for $z > x \in \dot{c}(\operatorname{int} N_\rho)$,

$$
\begin{aligned}
J^1(x, z) &= \inf_{\lambda \in \Theta, \dot{c}(\lambda) > z} \mathscr{I}(\dot{c}^{-1}(x), \lambda) \\
&= \inf_{\lambda \in \Theta, \dot{c}(\lambda) > z} \int_{\dot{c}^{-1}(x)}^{\lambda} (\lambda - t)\ddot{c}(t)\, dt \\
&= \int_{\dot{c}^{-1}(x)}^{\dot{c}^{-1}(z^+)} (\dot{c}^{-1}(z^+) - t)\ddot{c}(t)\, dt \qquad (5.10) \\
&= \mathscr{I}(\dot{c}^{-1}(x), \dot{c}^{-1}(z^+)).
\end{aligned}
$$

Similarly, for $z > x = \inf(\dot{c}(\operatorname{int} N_\rho))$ with $\rho(\{x\}) > 0$,

$$
\begin{aligned}
J^1(x, z) &= \inf_{\lambda \in \Theta, \dot{c}(\lambda) > z} \mathscr{I}(\delta_x, \rho_\lambda) \\
&= \inf_{\lambda \in \Theta, \dot{c}(\lambda) > z} (-\log \rho_\lambda(\{x\})) \\
&= \inf_{\lambda \in \Theta, \dot{c}(\lambda) > z} (-\log \rho(\{x\}) - \lambda x + c(\lambda)) \\
&= -\log \rho(\{x\}) - \dot{c}^{-1}(z^+) + c(\dot{c}^{-1}(z^+)) \\
&= \mathscr{I}(\delta_x, \rho_{\dot{c}^{-1}}(z^+)).
\end{aligned}
$$

This establishes (5.8).

Next note that for $\mu, \mu^* \in \mathscr{N}^e_\rho(\Theta)$ with $x = m_\mu < m_{\mu^*} = z$ we are in the regime $\sup \dot{c}(\Theta) > z > x \in \dot{c}(\operatorname{int} N_\rho)$ and therefore we have the integral expression (5.10) for $J^1(x, z)$. Moreover, it is easy to check that the mapping $z \mapsto z^+$ is continuous because of the condition (5.7) (dense from right) on Θ. Therefore, (5.9) follows by the continuity of $\dot{c}^{-1}$.

Finally, in light of Theorems 2.1, 2.2 and 4.10, it follows that ϕ^{a,J^1} is asymptotically efficient.

The index g^{a,J^1} and the index policy ϕ^{a,J^1} obtained above with J^1 given by (5.8) are very similar to those developed by Lai [13] for a finite-horizon Bayes version of the multi-armed bandit problem where the arms are drawn from an exponential family.

From (5.2) and (5.3), we can also obtain the following expressions for the constants in Example 4.3 for certain exponential families.

Remark 5.2. *Let $\mathscr{N}^e_\rho(\Theta)$ be a one-parameter exponential family with parameter set*

$\Theta \subset N_\rho$ such that $\Theta = (-\infty, a)$ (or $\Theta = (-\infty, a]$) for some $a \in (-\infty, \infty]$. Then, for $J = J^3$ of Example 4.3,

$$D = \dot{c}(a) - \inf S_\rho, \tag{5.11}$$

$$\sigma^2 = \sup_{\theta \in \Theta} \ddot{c}(\theta). \tag{5.12}$$

We shall also consider one-parameter *shifted* families

$$\mathcal{N}_\rho^s(\Theta) = \{\rho^\theta : \theta \in \Theta\} \tag{5.13}$$

where ρ^θ are the shifted measures given by

$$\rho^\theta(A) = \rho(A - \theta) \qquad \text{for all Borel measurable } A, \tag{5.14}$$

and $\Theta \subset \mathbb{R}$ is the parameter set of interest. Without loss of generality, we assume that $m_\rho = 0$. Then, for all $\theta, \lambda \in \mathbb{R}$,

$$m^\theta := m_{\rho^\theta} = \theta \tag{5.15}$$

$$c^\theta(t) := c_{\rho^\theta}(t) = t\theta + c(t) \tag{5.16}$$

$$\mathcal{I}(\theta, \lambda) := \mathcal{I}(\rho^\theta, \rho^\lambda) = \mathcal{I}(\rho, \rho^{\lambda-\theta}) =: \mathcal{I}(0, \lambda - \theta) \tag{5.17}$$

$$= \mathcal{I}(\rho^{\theta-\lambda}, \rho) =: \mathcal{I}(\theta - \lambda, 0) \tag{5.18}$$

$$I_\theta(x) := I_{\rho^\theta}(x) = I_\rho(x - \theta) =: I_0(x - \theta). \tag{5.19}$$

Note that for the multi-armed bandit problem with $\mathcal{N} = \mathcal{N}_\rho^s(\Theta)$, the sequence of rewards $\{X_i^j\}$ from any arm j can be written as $X_i^j = \theta_j + W_i$ where $\rho_j = \rho^{\theta_j}$ for some $\theta_j \in \Theta$ and $\{W_i\}$ is an i.d.d. noise sequence with common distribution ρ (regardless of the arm j). This corresponds to the case where the rewards for each arm are obtained as a common additive noise about their respective mean rewards.

Theorem 5.3. *Let $\mathcal{N} = \mathcal{N}_\rho^s(\mathbb{R})$ be a one-parameter shifted family of probability measures. Let J^1 and J^2 be given by Examples* 4.1 *and* 4.2 *respectively. Then, for all $\delta > 0$*

$$J^1(x, x + \delta) = J^2(\delta) = I_\rho(-\delta^+) := \inf_{\varepsilon > \delta} I_\rho(-\varepsilon). \tag{5.20}$$

Proof. For $\delta > 0$,

$$\begin{aligned} J^1(x, x+\delta) &= \inf_{\nu \in \mathcal{N}, m_\nu > x+\delta} \; \inf_{\pi \in \mathcal{M}, m_\pi = x} \mathcal{I}(\pi, \nu) \\ &= \inf_{\nu \in \mathcal{N}, m_\nu > x+\delta} I_\nu(x) \\ &= \inf_{\theta > x+\delta} I_{\rho^\theta}(x) \end{aligned}$$

$$= \inf_{\theta > x+\delta} I_\rho(x-\theta)$$

$$= \inf_{\varepsilon > \delta} I_\rho(-\varepsilon)$$

$$=: I_\rho(-\delta^+).$$

Since $J^1(x, x+\delta)$ depends on δ only, $J^1(x, x+\delta) = J^2(\delta)$, and the theorem follows.

Note that for the case of a one-parameter shifted family of probability measures $\mathcal{N} = \mathcal{N}^s_\rho(\mathbb{R})$, $\mathcal{I}(\rho^x, \rho^z) = \mathcal{I}(\rho, \rho^{z-x}) = \mathcal{I}(\rho^{x-z}, \rho)$ also depends on $z - x$ only. However, in general $I_\rho(x-z) \leqq \mathcal{I}(\rho^{x-z}, \rho)$, with equality holding iff $\mathcal{N}$ is a family of Gaussian distributions parametrized by their means (with constant variance).

Corollary 5.4. *Let $J = J^2$ be given by Example* 4.2, *and let* $\boldsymbol{a}$ *satisfy conditions* C1–C3. *Then the sample mean based index policy $\phi^{\boldsymbol{a},J^2}$ is asymptotically optimal if $\mathcal{N}$ is a family of Gaussian distributions parametrized by their means* (*with constant variance*).

Proof. Follows from Theorems 5.1 and 5.3 by simply observing that a family of Gaussian distributions parametrized by their means (with constant variance) is a one-parameter steep exponential family as well as a one-parameter shifted family with parameter set $\Theta = \mathbb{R}$.

5.1. *Examples.* In this subsection we consider several examples of the space of probability measures $\mathcal{N}$, and derive the corresponding 'rate' functions J^1, J^2, and J^3 (given by Examples 4.1, 4.2, and 4.3, respectively), as well as the corresponding upper confidence bound functions $\boldsymbol{g}^{\boldsymbol{a},J^1}$, $\boldsymbol{g}^{\boldsymbol{a},J^2}$, and $\boldsymbol{g}^{\boldsymbol{a},J^3}$ (given by Definition 4.7, Remark 4.8, and Remark 4.9, respectively). Examples 5.5–5.8 will be of one-parameter exponential families, whereas Examples 5.5, 5.9 will be of one-parameter shifted families. We will need to compute c, N_ρ, $\dot{c}$, $\dot{c}^{-1}$, and I, which can be obtained from (3.2), (3.3) (or (5.13)), and (5.5) (or (5.17), (5.19)). For exponential families, J^1 is given by (5.8), whereas for shifted families $J^1 = J^2$ is given by (5.20). We also make use of Remark 5.2, to compute J^3 and the corresponding upper confidence bounds for the examples of exponential families.

Example 5.5 (Gaussian). Let $\mathcal{N} = \mathcal{N}^e_\rho(\Theta)$ be a one-parameter exponential family with

$$\rho(dx) = \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\}\, dx.$$

For this example,

$$c(\theta) = \theta^2/2, \qquad N_\rho = \mathbb{R},$$

$$\dot{c}(\theta) = \theta, \qquad \dot{c}^{-1}(x) = x,$$

$$\mathcal{I}(\theta, \lambda) = (\theta - \lambda)^2/2.$$

Note that $\mathcal{N}$ is a family of Gaussian distributions (with constant variance) which is also a one-parameter shifted family of probability measures. Hence, Theorems 5.1, 5.3, Corollary 5.4, and Proposition 5.2 apply. Thus, for $\Theta = N_\rho$,

$$J^1(x, z) = J^2(z - x) = J^3(z - x) = (x - z)^2/2.$$

Consequently, the corresponding upper confidence bounds are identical, and by Remark 4.9 are given by

$$g_{ni}^{a,J^1}(x) = g_{ni}^{a,J^2}(x) = g_{ni}^{a,J^3}(x) = x + (2a_{ni})^{\frac{1}{2}}$$

Example 5.6 (exponential). Let $\mathcal{N} = \mathcal{N}_\rho^e(\Theta)$ be a one-parameter exponential family with

$$\rho(dx) = e^{-x}1\{x \geqq 0\}\, dx.$$

For this example,

$$c(\theta) = -\log(1 - \theta), \qquad N_\rho = (-\infty, 1),$$
$$\dot{c}(\theta) = (1 - \theta)^{-1}, \qquad \dot{c}^{-1}(x) = (x - 1)/x,$$
$$\mathcal{I}(\theta, \lambda) = \frac{\theta - \lambda}{1 - \theta} - \log\frac{1 - \lambda}{1 - \theta}.$$

Thus, for $\Theta = N_\rho$, and $x < z$, we have

$$J^1(x, z) = \begin{cases} \mathcal{I}(\dot{c}^{-1}(x), \dot{c}^{-1}(z)) = \dfrac{x}{z} - 1 - \log\dfrac{x}{z}, & x > 0, \\ \infty, & \text{otherwise.} \end{cases}$$

So for $x \geqq 0$,

$$g_{ni}^{a,J^1}(x) = \inf\left\{z \geqq x : \frac{x}{z} - 1 - \log\frac{x}{z} \geqq a_{ni}\right\}$$
$$= xg(a_{ni})$$

where for any $a \geqq 0$,

$$g(a) := \inf\left\{z \geqq 1 : \frac{1}{z} - 1 - \log\frac{1}{z} \geqq a\right\}.$$

Thus, the upper confidence bound is a 'scaling' of the sample mean by a factor

$g(a_{ni})$. For $\Theta = N_\rho$, $J^2 \equiv 0$. Thus, we now consider $\Theta = (-\infty, a) \subset N_\rho$ with $a < 1$. For this choice of Θ, J^1 obtained above is modified to ∞ when $z \geqq (1-a)^{-1}$. Thus,

$$\begin{aligned} J^2(\delta) &= \inf_{x \in (0,(1-a)^{-1}-\delta)} J^1(x, x+\delta) \\ &= \inf_{x \in (0,(1-a)^{-1}-\delta)} \left(\frac{x}{x+\delta} - 1 - \log \frac{x}{x+\delta}\right) \\ &= \begin{cases} -\log(1-(1-a)\delta) - (1-a)\delta, & 0 < \delta < (1-a)^{-1}, \\ \infty, & \text{otherwise.} \end{cases} \end{aligned}$$

For the same choice of Θ, we also have

$$J^2(\delta) \geqq J^3(\delta) = \begin{cases} (1-a)^2\delta^2/2, & 0 \leqq \delta \leqq (1-a)^{-1}, \\ \infty, & \text{otherwise.} \end{cases}$$

The corresponding upper confidence bound is given by

$$g_{ni}^{a,J^3}(x) = x + (1-a)^{-1} \min\{(2a_{ni})^{\frac{1}{2}}, 1\}.$$

Example 5.7 (Bernoulli). Let $\mathcal{N} = \mathcal{N}_\rho^e(\Theta)$ be a one-parameter exponential family with

$$\rho(dx) = \frac{1}{2}(\delta_0(dx) + \delta_1(dx))$$

where δ_0 and δ_1 are the Dirac delta measures at 0 and 1, respectively. For this example,

$$c(\theta) = \log \frac{1+e^\theta}{2}, \qquad N_\rho = \mathbb{R},$$

$$\dot{c}(\theta) = \frac{e^\theta}{1+e^\theta}, \qquad \dot{c}^{-1}(x) = \log \frac{x}{1-x},$$

$$\mathcal{I}(\theta, \lambda) = (\theta - \lambda)\frac{e^\theta}{1+e^\theta} - \log \frac{1+e^\theta}{1+e^\lambda}.$$

Thus, for $\Theta = N_\rho$, and $x < z$, we have

$$J^1(x, z) = \begin{cases} x \log \dfrac{x}{z} + (1-x) \log \dfrac{1-x}{1-z}, & 0 < x < z < 1, \\ -\log(1-z), & 0 = x < z < 1, \\ \infty, & \text{otherwise.} \end{cases}$$

Thus,

$$
\begin{aligned}
g_{ni}^{a,J^1}(x) &= \inf\{z \geqq x : J^1(x, z) \geqq a_{ni}\} \\
&= \begin{cases} \inf\left\{x \leqq z < 1 : x \log \dfrac{x}{z} + (1-x) \log \dfrac{1-x}{1-z} \geqq a_{ni}\right\}, & x > 0, \\ 1 - \exp\{-a_{ni}\}, & x = 0. \end{cases}
\end{aligned}
$$

Also for $\Theta = N_\rho$,

$$
\begin{aligned}
J^2(\delta) &= \inf_{x \in (0, 1-\delta)} J^1(x, x+\delta) \\
&= \begin{cases} \displaystyle\inf_{x \in (0,1-\delta)} \left(x \log \frac{x}{x+\delta} + (1-x) \log \frac{1-x}{1-x-\delta}\right), & 0 \leqq \delta < 1, \\ \infty, & \text{otherwise.} \end{cases}
\end{aligned}
$$

For the same choice of Θ, we also have

$$
J^2(\delta) \geqq J^3(\delta) = \begin{cases} 2\delta^2, & 0 \leqq \delta \leqq 1, \\ \infty, & \text{otherwise.} \end{cases}
$$

The corresponding upper confidence bound is given by

$$
g_{ni}^{a,J^3}(x) = x + \min\{(2a_{ni})^{\frac{1}{2}}/2, 1\}.
$$

Example 5.8 (Poisson). Let $\mathcal{N} = \mathcal{N}_\rho^e(\Theta)$ be a one-parameter exponential family with

$$
\rho(dx) = \sum_{j=0}^{\infty} \frac{e^{-1}}{j!} \delta_j(dx)
$$

where δ_j is the Dirac delta measure at j ($j \geqq 0$). For this example,

$$
\begin{gathered}
c(\theta) = e^\theta - 1, \qquad N_\rho = \mathbb{R}, \\
\dot{c}(\theta) = e^\theta, \qquad \dot{c}^{-1}(x) = \log x, \\
\mathcal{I}(\theta, \lambda) = (\theta - \lambda)e^\theta - e^\theta + e^\lambda.
\end{gathered}
$$

Thus, for $\Theta = N_\rho$, and $x < z$, we have

$$
J^1(x, z) = \begin{cases} x\left(\dfrac{z}{x} - 1 - \log \dfrac{z}{x}\right), & x > 0, \\ z, & x = 0, \\ \infty, & \text{otherwise.} \end{cases}
$$

Thus, for $x = 0$, $g_{ni}^{a,J^1}(x) = \inf\{z \geqq x : z \geqq a_{ni}\} = a_{ni}$, and for $x > 0$,

$$
\begin{aligned}
g_{ni}^{a,J^1}(x) &= \inf\left\{z \geqq x : x\left(\frac{z}{x} - 1 - \log \frac{z}{x}\right) \geqq a_{ni}\right\} \\
&= x g(a_{ni}/x)
\end{aligned}
$$

where for any $a > 0$,

$$g(a) := \inf\{z \geqq 1 : z - 1 - \log z \geqq a\}.$$

For $\Theta = N_\rho$, $J^2 \equiv 0$. Thus, we now consider $\Theta = (-\infty, a) \subset N_\rho$ with $a < \infty$. For this choice of Θ, J^1 obtained above is modified to ∞ when $z \geqq e^a$. Thus,

$$\begin{aligned} J^2(\delta) &= \inf_{x \in (0, e^a - \delta)} J^1(x, x + \delta) \\ &= \inf_{x \in (0, e^a - \delta)} (\delta - x \log(1 + \delta/x)) \\ &= \begin{cases} \delta - (e^a - \delta)\log(1 + \delta/(e^a - \delta)), & 0 < \delta < e^a, \\ \infty, & \text{otherwise.} \end{cases} \end{aligned}$$

For the same choice of Θ, we also have

$$J^2(\delta) \geqq J^3(\delta) = \begin{cases} \delta^2/2e^a, & 0 \leqq \delta \leqq e^a, \\ \infty, & \text{otherwise.} \end{cases}$$

The corresponding upper confidence bound is given by

$$g_{ni}^{a, J^3}(x) = x + \min\{(2a_{ni}e^a)^{\frac{1}{2}}/2, e^a\}.$$

Example 5.9 (Laplacian). Let $\mathcal{N} = \mathcal{N}_\rho^s(\Theta)$ be a one-parameter shifted family with

$$\rho(dx) = \frac{1}{2} e^{-|x|}\, dx.$$

For this example,

$$\begin{aligned} c(\theta) &= -\log(1 - \theta^2), \quad \theta \in (-1, 1), \\ \dot{c}(\theta) &= \frac{2\theta}{1 - \theta^2}, \qquad \dot{c}^{-1}(x) = \frac{(1 + x^2)^{\frac{1}{2}} - 1}{x}, \\ I_\rho(\delta) &= \delta \dot{c}^{-1}(\delta) - c(\dot{c}^{-1}(\delta)) \\ &= \delta \frac{(1 + \delta^2)^{\frac{1}{2}} - 1}{\delta} + \log\left(1 - \left(\frac{(1 + \delta^2)^{\frac{1}{2}} - 1}{\delta}\right)^2\right) \\ &= (1 + \delta^2)^{\frac{1}{2}} - 1 + \log((1 + \delta^2)^{\frac{1}{2}} - 1) - \log(\delta^2/2). \end{aligned}$$

Thus, for $\Theta = \mathbb{R}$, we have

$$J^1(x, x + \delta) = J^2(\delta) = I_\rho(-\delta^+) = I_\rho(-\delta) = I_\rho(\delta),$$

and

$$g_{ni}^{a, J^1}(x) = g_{ni}^{a, J^2}(x) = x + \inf\{\delta \geqq 0 : J^2(\delta) \geqq a_{ni}\}.$$

6. Conclusion

In this paper we constructed index policies for the multi-armed bandit problem. The indices for an arm depended on the observations from that arm only through their sample mean. The index functions make use of a 'contraction' of the Kullback–Leibler number and an *a priori* defined sequence of numbers. We show that these indices are upper confidence bounds and obtain an $O(\log n)$ upper bound on the regret associated with the corresponding index policies. We also show that the policies based on the function J^1 achieve the best constant for one-parameter exponential families. While these policies may not achieve the best constant in general, they are atrractive because of their simplicity. Moreover, the upper confidence bounds constructed in this paper can easily be generalized to handle other (additive) functionals besides the sample mean. In particular, if we use the empirical measure, then we would recover the optimal constant (modulo some continuity conditions).

The upper confidence bounds (indices) constructed in this paper can also be used to construct policies for the problem with switching costs (see [2]) as well as for the problem with multiple plays (see [7], [3]).

Appendix

Proof that Example 4.5 satisfies C1–C3. That C1 is satisfied is obvious from the condition that b_n is non-decreasing with n. To obtain C2, observe that

$$\begin{aligned}\sum_{i=1}^{c\log n} \exp\{-ia_{ni}\} &= \sum_{i=1}^{c\log n} \exp\{-(\log n + \log\log n + b_n)\}\\ &= c\log n(n\log n\cdot\exp\{b_n\})^{-1} = o(n^{-1}).\end{aligned}$$

Finally, note that

$$\begin{aligned}l_n(I) :&= \sup\{1 \leqq i \leqq n : a_{ni} > I\}\\ &= \sup\left\{1 \leqq i \leqq n : \frac{\log n + \log\log n + b_n}{i} > I\right\}\\ &= \sup\left\{1 \leqq i \leqq n : \frac{\log n + \log\log n + b_n}{I} > i\right\}\\ &\leqq \frac{\log n + \log\log n + b_n}{I}.\end{aligned}$$

C3 now follows on dividing by $\log n$ and taking limits as $n \to \infty$.

Proof that Example 4.6 satisfies C1–C3. That C1 is satisfied is obvious from the condition that b_n is non-decreasing with n. To obtain C2, observe that

$$\sum_{i=1}^{c\log n} \exp\{-ia_{ni}\} \leqq \sum_{i=1}^{\infty} \exp\{-ia_{ni}\} = \sum_{i=1}^{\infty} \exp\{-(\log n + b_n^{\alpha} i^{1-\alpha})\} = n^{-1}\sum_{i=1}^{\infty} \exp\{-b_n^{\alpha} i^{1-\alpha}\}.$$

Moreover,

$$\begin{aligned}\sum_{i=1}^{\infty} \exp\{-b_n^{\alpha} i^{1-\alpha}\} &\leqq \int_0^{\infty} \exp\{-b_n^{\alpha} x^{1-\alpha}\}\,dx \\ &= b_n^{-\alpha/(1-\alpha)}(1-\alpha)^{-1}\int_0^{\infty} e^{-t} t^{(1-\alpha)^{-1}-1}\,dt \\ &= b_n^{-\alpha/(1-\alpha)}(1-\alpha)^{-1}\Gamma((1-\alpha)^{-1}) \\ &= b_n^{-\alpha/(1-\alpha)}\Gamma((1-\alpha)^{-1}+1),\end{aligned}$$

where Γ is the gamma function. C2 now follows from the condition that $b_n \nearrow \infty$ as $n \to \infty$. Finally, note that

$$\begin{aligned} l_n(I) &:= \sup\{1 \leqq i \leqq n : a_{ni} > I\} \\ &= \sup\left\{1 \leqq i \leqq n : \frac{\log n}{i} + \left(\frac{b_n}{i}\right)^{\alpha} > I\right\} \\ &\leqq c_n b_n + \sup\left\{c_n b_n \leqq i \leqq n : \frac{\log n}{i} + \left(\frac{b_n}{i}\right)^{\alpha} > I\right\} \\ &\leqq c_n b_n + \sup\left\{c_n b_n \leqq i \leqq n : \frac{\log n}{i} + c_n^{-\alpha} > I\right\} \\ &= c_n b_n + \sup\left\{c_n b_n \leqq i \leqq n : i < \frac{\log n}{I - c_n^{-\alpha}}\right\} \\ &\leqq c_n b_n + \frac{\log n}{I - c_n^{-\alpha}} \end{aligned}$$

for all $c_n > 0$. Now choose $c_n = ((\log n)/b_n)^{\frac{1}{2}}$. Dividing by $\log n$ we get

$$l_n(I) \leqq \left(\frac{b_n}{\log n}\right)^{\frac{1}{2}} + \frac{1}{I - (b_n/\log n)^{\alpha/2}}.$$

C3 now follows from the condition that $b_n/\log n \to 0$ as $n \to \infty$.

References

[1] Agrawal, R. (1991) Minimizing the learning loss in adaptive control of Markov chains under the weak accessibility condition. *J. Appl. Prob.* **28,** 779–790.

[2] Agrawal, R., Hegde, M. and Teneketzis, D. (1988) Asymptotically efficient adaptive allocation rules for the multi-armed bandit problem with switching cost. *IEEE Trans. Autom. Control.* **33,** 899–906.

[3] AGRAWAL, R., HEGDE, M. AND TENEKETZIS, D. (1990) Multi-armed bandit problems with multiple plays and switching cost. *Stoch. Stoch. Reports* **29,** 437–459.

[4] AGRAWAL, R. AND TENEKETZIS, D. (1989) Certainty equivalence control with forcing: Revisited. *Syst. Contr. Lett.* **13,** 405–412.

[5] AGRAWAL, R., TENEKETZIS, D. AND ANANTHARAM, V. (1989) Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite parameter space. *IEEE Trans. Autom. Control.,* 258–267.

[6] AGRAWAL, R., TENEKETZIS, D. AND ANANTHARAM, V. (1989) Asymptotically efficient adaptive allocation schemes for controlled Markov chains: Finite parameter space. *IEEE Trans. Autom. Contr.* **34,** 1249–1259.

[7] ANANTHARAM, V., VARAIYA, P. AND WALRAND, J. (1987) Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays; Part I: IID rewards. *IEEE Trans. Autom. Control* **32,** 968–975.

[8] ANANTHARAM, V., VARAIYA, P. AND WALRAND, J. (1987) Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays; Part II: Markovian rewards. *IEEE Trans. Autom. Control* **32,** 975–982.

[9] BILLINGSLEY, P. (1986) *Probability and Measure,* 2nd edn., Wiley, New York.

[10] BROWN, L. D. (1986) *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory.* Institute of Mathematical Statistics.

[11] DEMBO, A. AND ZEITOUNI, O. (1993) *Large Deviation Techniques and Applications.* Jones and Bartlett.

[12] ELLIS, R. S. (1985) *Entropy, Large Deviations, and Statistical Mechanics.* Springer-Verlag, Berlin.

[13] LAI, T. L. (1987) Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* **15,** 1091–1114.

[14] LAI, T. L. AND ROBBINS, H. (1984) Asymptotically optimal allocation of treatments in sequential experiments. In *Design of Experiments,* ed. T. J. Santer and A. J. Tamhane, pp. 127–142, Marcel Dekker, New York.

[15] LAI, T. L. AND ROBBINS, H. (1985) Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **6,** 4–22.