# Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem with Multiple Plays—Part I: I.I.D. Rewards

VENKATACHALAM ANANTHARAM, MEMBER, IEEE, PRAVIN VARAIYA, FELLOW, IEEE, AND JEAN WALRAND, MEMBER, IEEE

*Abstract*—**At each instant of time we are required to sample a fixed number $m \geq 1$ out of $N$ i.i.d. processes whose distributions belong to a family suitably parameterized by a real number $\theta$. The objective is to maximize the long run total expected value of the samples. Following Lai and Robbins, the learning loss of a sampling scheme corresponding to a configuration of parameters $C = (\theta_1, \cdots, \theta_N)$ is quantified by the *regret* $R_n(C)$. This is the difference between the maximum expected reward at time $n$ that could be achieved if $C$ were known and the expected reward actually obtained by the sampling scheme. We provide a lower bound for the regret associated with any uniformly good scheme, and construct a scheme which attains the lower bound for every configuration $C$. The lower bound is given explicitly in terms of the Kullback–Liebler number between pairs of distributions. Part II of this paper considers the same problem when the reward processes are Markovian.**

## I. INTRODUCTION

IN this paper we study a version of the multiarmed bandit problem with multiple plays. We are given a one-parameter family of reward distributions with densities $f(x, \theta)$ with respect to some measure $\nu$ on $R$. $\theta$ is a real valued parameter. There are $N$ arms $X_j$, $j = 1, \cdots, N$ with parameter configuration $C = (\theta_1, \cdots, \theta_N)$. When arm $j$ is played, it gives a reward with distribution $f(x, \theta_j)d\nu(x)$. Successive plays of arm $j$ produce i.i.d. rewards. At each stage we are required to play a fixed number, $m$, of the arms, $1 \leq m \leq N$.

Suppose we know the distributions of the individual rewards. To maximize the total expected reward up to any stage, one must play the arms with the $m$ highest means. However, if the parameters $\theta_j$ are unknown, we are forced to play the poorer arms in order to learn about their means from the observations. The aim is to minimize, in some sense, the *total* expected loss incurred in the process of learning for *every* possible parameter configuration.

For single plays, i.e., $m = 1$, this problem was studied by Lai and Robbins [3]–[5]. The techniques used here closely parallel their approach. However, the final results are somewhat more general even in the single play case. For multiple plays, i.e., $m > 1$, we report the first general results. In Part II of this paper we study the same problem when the reward statistics of the arms are Markovian with finite state space instead of i.i.d.

Manuscript received September 8, 1986; revised June 1, 1987. Paper recommended by Past Associate Editor, A Ephremides. This work was supported by the Air Force Office of Scientific Research under Contract F49260-87-C-0041.

V. Anantharam was with the Department of Electrical Engineering and Computer Science and the Electronics Research Laboratory, University of California, Berkeley, CA 94720. He is now with the School of Electrical Engineering, Cornell University, Ithaca, NY 14853.

P. Varaiya and J. Walrand are with the Department of Electrical Engineering and Computer Science and the Electronics Research Laboratory, University of California, Berkeley, CA 94720.

IEEE Log Number 8716539.

The actual values $\theta$ that can arise as parameters of the arms are known *a priori* to belong to a subset $\Theta \subseteq R$. In Sections II–V $\Theta$ is assumed to satisfy the denseness condition (2.4). This restriction is removed in Sections VI and VII.

Multiarmed bandit problems provide a simple framework to study resource allocation problems arising in manufacturing systems, computer systems, etc. When there is a single scarce resource, e.g., a single machine or CPU, one gets a multiarmed bandit problem with a single play with arms corresponding to jobs. When there is more than one scarce resource, one gets a problem with multiple plays, each play corresponding to one resource.

The results constitute part of the first author's dissertation.

## II. SETUP

We assume that the rewards are integrable

$$\int_{-\infty}^{\infty} |x| f(x, \theta)\, d\nu(x) < \infty \tag{2.1}$$

and the mean reward

$$\mu(\theta) = \int_{-\infty}^{\infty} x f(x, \theta)\, d\nu(x)$$

is a strictly monotone increasing function of the parameter $\theta$.

The Kullback–Liebler number,

$$I(\theta, \lambda) = \int_{-\infty}^{\infty} \log \left[\frac{f(x, \theta)}{f(x, \lambda)}\right] f(x, \theta)\, d\nu(x)$$

is a well-known measure of dissimilarity between two distributions. In general, $0 \leq I(\theta, \lambda) \leq \infty$. We assume that

$$0 < I(\theta, \lambda) < \infty \qquad \text{if } \lambda > \theta \tag{2.2}$$

and

$$I(\theta, \lambda) \text{ is continuous in } \lambda > \theta \qquad \text{for fixed } \theta. \tag{2.3}$$

In Sections II–V the following denseness condition on $\Theta$ is imposed:

for all $\lambda \in \Theta$ and $\delta > 0$,

$$\text{there is } \lambda' \in \Theta \text{ s.t. } \mu(\lambda) < \mu(\lambda') < \mu(\lambda) + \delta. \tag{2.4}$$

Let $Y_{j1}, Y_{j2}, \cdots$ denote successive rewards from arm $j$. Let $F_t(j)$ denote the $\sigma$-algebra generated by $Y_{j1}, \cdots, Y_{jt}$, let $F_\infty(j) = \vee_t F_t(j)$, and $G(j) = \vee_{i \neq j} F_\infty(i)$. An adaptive allocation rule is a rule for deciding which $m$ arms to play at time $t + 1$ based only on knowledge of the past rewards $Y_{j1}, \cdots, Y_{jT_t(j)}$, $j = 1, \cdots, N$ and the past decisions. For an adaptive allocation rule $\Phi$,

0018-9286/87/1100-0968$01.00 

the number of plays we have made of arm $j$ by time $t$, $T_t(j)$, is a stopping time of $\{F_s(j) \vee G(j), s \geq 1\}$. By Wald's lemma, (see, e.g., [1, Lemma 3.1]), if $S_t$ denotes the total reward received up to time $t$,

$$ES_t = \sum_{j=1}^{N} \mu(\theta_j) ET_t(j). \tag{2.5}$$

For a configuration $C = (\theta_1, \cdots, \theta_N)$, the loss associated to a rule is a function of the number of plays $t$ which gives the difference between the expected reward that could have been achieved with prior knowledge of the parameters and the expected reward actually achieved under the rule. Following [4], this function is called the *regret*. Let $\sigma$ be a permutation of $\{1, \cdots, N\}$ such that

$$\mu(\theta_{\sigma(1)}) \geq \theta(\mu_{\sigma(2)}) \geq \cdots \geq \mu(\theta_{\sigma(N)}).$$

Then the regret is

$$R_t(\theta_1, \cdots, \theta_N) = t \sum_{i=1}^{m} \mu(\theta_{\sigma(i)}) - ES_t. \tag{2.6}$$

The problem is to minimize the regret in some sense. Note that it is impossible to do this uniformly over all parameter configurations. For example, the rule "always play the arms $1, 2, \cdots, m$" will have zero regret when $\mu(\theta_i) > \mu(\theta_j)$ for all $1 \leq i \leq m$ and $m + 1 \leq j \leq N$. However, when a parameter configuration has $\mu(\theta_i) < \mu(\theta_j)$ for some $1 \leq i \leq m$ and $m + 1 \leq j \leq N$, this rule will have regret proportional to $t$.

We call a rule *uniformly good* if for *every* parameter configuration $R_t(\theta_1, \cdots, \theta_N) = o(t^\alpha)$ for every real $\alpha > 0$. We consider as uninteresting any rule that is not uniformly good.

## III. A Lower Bound for the Regret of a Uniformly Good Rule

Let the arms have parameter configuration $C = (\theta_1, \cdots, \theta_N)$ and let $\sigma$ be a permutation of $\{1, \cdots, N\}$ such that

$$\mu(\theta_{\sigma(1)}) \geq \cdots \geq \mu(\theta_{\sigma(N)}).$$

a) If $\mu(\theta_{\sigma(m)}) > \mu(\theta_{\sigma(m+1)})$, we call arms $\sigma(1), \cdots, \sigma(m)$ the distinctly *m-best* arms and $\sigma(m + 1), \cdots, \sigma(N)$ the distinctly *m-worst* arms.

b) If $\mu(\theta_{\sigma(m)}) = \mu(\theta_{\sigma(m+1)})$ let $0 \leq l < m$ and $m \leq n \leq N$ be such that

$$\mu(\theta_{\sigma(1)}) \geq \cdots \geq \mu(\theta_{\sigma(l)}) > \mu(\theta_{\sigma(l+1)}) = \cdots = \mu(\theta_{\sigma(m)}) = \cdots = \mu(\theta_{\sigma(n)}) > \mu(\theta_{\sigma(n+1)}) \geq \cdots \geq \mu(\theta_{\sigma(N)}).$$

Then we call arms $\sigma(1), \cdots, \sigma(l)$ the distinctly *m-best* arms, and arms $\sigma(n + 1), \cdots, \sigma(N)$ the distinctly *m-worst* arms.

c) The arms with mean equal to $\mu(\theta_{\sigma(m)})$ are called the *m-border* arms. Note that in a) $\sigma(m)$ is both a distinctly $m$-best arm and an $m$-border arm. In b) the $m$-order arms are the arms $j$, $l + 1 \leq j \leq n$.

The separation of arms into these three types will be crucial in all that follows.

Let $\Phi$ be an adaptive allocation rule. Then $\Phi$ is uniformly good iff for every distinctly $m$-best arm $j$

$$E(t - T_t(j)) = o(t^\alpha)$$

and for every distinctly $m$-worst arm $j$

$$E(T_t(j)) = o(t^\alpha)$$

for every real $\alpha > 0$.

*Theorem 3.1:* Let the family of reward distributions satisfy conditions (2.2), (2.3), and (2.4). Let $\Phi$ be a uniformly good rule. If the arms have parameter configuration $C = (\theta_1, \cdots, \theta_N)$, then for each distinctly $m$-worst arm $j$ and each $\epsilon > 0$

$$\lim_{t \to \infty} P_C \left\{ T_t(j) \geq \frac{(1 - \epsilon) \log t}{I(\theta_j, \theta_{\sigma(m)})} \right\} = 1$$

so that

$$\liminf_{t \to \infty} \frac{E_C T_t(j)}{\log t} \geq \frac{1}{I(\theta_j, \theta_{\sigma(m)})}$$

where $\sigma$ is a permutation of $\{1, \cdots, N\}$ such that

$$\mu(\theta_{\sigma(1)}) \geq \cdots \geq \mu(\theta_{\sigma(N)}).$$

Consequently,

$$\liminf_{t \to \infty} \frac{R_t(\theta_1, \cdots, \theta_N)}{\log t} \geq \sum_{j \text{ is } m-\text{worst}} \frac{[\mu(\theta_{\sigma(m)}) - \mu(\theta_j)]}{I(\theta_j, \theta_{\sigma(m)})}$$

for every configuration $C = (\theta_1, \cdots, \theta_N)$.

*Proof:* Let $j$ be an $m$-worst arm. Fix $\rho > 0$. Assumptions (2.3) and (2.4) allow us to choose a parameter value $\lambda$ satisfying

$$\mu(\lambda) > \mu(\theta_{\sigma(m)}) > \mu(\theta_j)$$

and

$$|I(\theta_j, \lambda) - I(\theta_j, \theta_{\sigma(m)})| \leq \rho I(\theta_j, \theta_{\sigma(m)}). \tag{3.1}$$

Consider the new configuration of parameters $C^* = (\theta_1, \cdots, \theta_{j-1}, \lambda, \theta_{j+1}, \cdots, \theta_N)$, i.e., replace $\theta_j$ by $\lambda$. Then arm $j$ is one of the distinctly $m$-best for the parameter configuration $C^*$. Thus, under configuration $C^*$, a uniformly good rule plays arm $j$ almost always. Even when the configuration is $C$, if the observed sequence of rewards makes it seem that the configuration is $C^*$, a uniformly good rule will tend to play arm $j$. The idea of the proof is to show that this happens with fairly high probability, by studying the likehood ratio of an observed sequence of rewards under the two configurations.

Let $Y_1, Y_2, \cdots$ denote the sequence of rewards from plays of arm $j$ under the uniformly good rule $\Phi$. Define

$$L_t = \sum_{a=1}^{t} \log \left[ \frac{f(Y_a, \theta_j)}{f(Y_a, \lambda)} \right].$$

By the strong law of large numbers $L_t/t \to I(\theta_j, \lambda)$ a.s. $[P_C]$. Hence, $1/t \max_{a \leq t} L_a \to I(\theta_j, \lambda)$ a.s. $[P_C]$. For any $K > 0$, we have

$$\lim_{t \to \infty} P_C\{L_a > K(1 + \rho) I(\theta_j, \lambda) \log t \quad \text{for some } a < K \log t\} = 0. \tag{3.2}$$

We write

$$\{T_t(j) < K \log t, L_{T_t(j)} \leq K(1 + \rho) I(\theta_j, \lambda) \log t\} = \bigcup_{a < K \log t} \{T_t(j) = a, L_a \leq K(1 + \rho) I(\theta_j, \lambda) \log t\}$$

and

$$P_{C^*}\{T_t(j) = a, L_a \leq K(1 + \rho) I(\theta_j, \lambda) \log t\} = \int_{\{T_t(j) = a, L_a \leq K(1+\rho) I(\theta_j, \lambda) \log t\}} \prod_{b=1}^{a} \frac{f(Y_b, \lambda)}{f(Y_b, \theta_j)} \, dP_C$$

$$\geq t^{-K(1+\rho) I(\theta_j, \lambda)} P_C\{T_t(j) = a, L_a \leq K(1 + \rho) I(\theta_j, \lambda) \log t\}.$$

Thus

$$P_{C^*}\{T_t(j)<K\log t,\ L_{T_t(j)}\le K(1+\rho)I(\theta_j,\lambda)\log t\}$$
$$\ge t^{-K(1+\rho)I(\theta_j,\lambda)}P_C\{T_t(j)<K\log t,$$
$$L_{T_t(j)}\le K(1+\rho)I(\theta_j,\lambda)\log t\}. \tag{3.3}$$

Since $\Phi$ is uniformly good and arm $j$ is distinctly $m$-best under $C^* = (\theta_1, \cdots, \theta_{j-1}, \lambda, \theta_{j+1}, \cdots, \theta_N)$

$$E_{C^*}(t-T_t(j))=o(t^\alpha)$$

so that

$$(t-K\log t)P_{C^*}\{T_t(j)<K\log t\}=o(t^\alpha),$$

hence,

$$P_{C^*}\{T_t(j)<K\log t\}=o(t^{\alpha-1}) \tag{3.4}$$

for every real $\alpha > 0$.

Choosing $K = 1/(1 + 2\rho)I(\theta_j, \lambda)$, we have, from (3.2)–(3.4),

$$\lim_{t\to\infty} P_C\left\{T_t(j)<\frac{\log t}{(1+2\rho)I(\theta_j,\lambda)}\right\}=0. \tag{3.5}$$

Since $(1 + \rho)I(\theta_j, \theta_{\sigma(m)}) \ge I(\theta_j, \lambda)$ by (3.1), we have

$$\lim_{t\to\infty} P_C\left\{T_t(j)<\frac{\log t}{(1+2\rho)(1+\rho)I(\theta_j,\theta_{\sigma(m)})}\right\}=0$$

for every $\rho > 0$. Writing $1/(1 + 2\rho)(1 + \rho)$ as $1 - \epsilon$ proves the first claim. Letting $\epsilon \to 0$ proves the second claim. □

## IV. Construction of Statistics

Motivated by Theorem 3.1, we call an adaptive allocation rule *asymptotically efficient* if for each configuration $C = (\theta_1, \cdots, \theta_N)$,

$$\operatorname{limsup}_{t\to\infty}\frac{R_t(\theta_1,\cdots,\theta_N)}{\log t}\le\sum_{j\text{ is }m\text{-worst}}\frac{[\mu(\theta_{\sigma(m)})-\mu(\theta_j)]}{I(\theta_j,\theta_{\sigma(m)})}.$$

To construct an asymptotically efficient rule we need a technique for deciding when we need to experiment, i.e., when to play an arm in order to learn more about its parameter value from the additional sample. At time $t$ we have $T_t(j)$ samples from arm $j$ from which we can estimate $\theta_j$ by various methods, e.g., sample mean, maximum likelihood estimate, sample median. The decision we have to make at time $t + 1$ is whether to play the $m$ arms whose estimated parameter values are the largest—"play the winners" rule—or to experiment by playing some of the apparently inferior arms. To do this we will construct a family of statistics $g_{ta}(Y_1, \cdots, Y_a)$, $1 \le a \le t$, $t = 1, 2, \cdots$, so that when $g_{tT_t(j)}$ is larger than any of the $m$ best estimated parameter values, this indicates the need to experiment with arm $j$. Such statistics are constructed in [5] for exponential families of distributions, based on results of Pollack and Siegmund [7]. We use a similar technique to construct $g_{ta}(Y_1, \cdots, Y_a)$ under the following assumptions:

$$\log f(x,\theta)\text{ is concave in }\theta\text{ for each fixed }x, \tag{4.1}$$

$$\int x^2 f(x,\theta)\,d\nu(x)<\infty \qquad \text{for each } \theta\in R. \tag{4.2}$$

The reader may wish to glance at the beginning of Section V at this point, to see how these statistics are used to construct an asymptotically efficient rule.

Lemmas 4.1 and 4.2 are needed in the proof of Theorem 4.1.

*Lemma 4.1:* Let $S_t = X_1 + \cdots + X_t$ where $X_1, X_2, \cdots$ are i.i.d., $EX_1 > 0$, and let $N = \sum_{t=1}^{\infty} 1(S_t \le 0)$, $L = \sum_{t=1}^{\infty} 1(\inf_{a\ge t} S_a \le 0)$. The following are equivalent:

a) $E(|X_1|^2 1(X_1 \le 0)) < \infty$;

b) $EN < \infty$;

c) $EL < \infty$.

*Proof:* See Hogan [2]. □

*Lemma 4.2:* Let $S_t = X_1 + \cdots + X_t$ where $X_1, X_2, \cdots$ are i.i.d., $EX_1 > 0$. Given $A > 0$, let $N_A = \sum_{t=1}^{\infty} 1(S_t \le A)$. If $E(|X_1|^2 1(X_1 \le 0)) < \infty$, then

$$\limsup_{A\to\infty}\frac{EN_A}{A}\le\frac{1}{EX_1}.$$

*Proof:* For $\epsilon > 0$

$$N_A\le\frac{A(1+\epsilon)}{EX_1}+\sum_{t=1}^{\infty}1\left(S_t\le\frac{tEX_1}{1+\epsilon}\right).$$

Let $Z_t = X_t - EX_1/(1 + \epsilon)$. Then

$$E\{|Z_1|^2 1(Z_1\le 0)\}$$
$$\le 2E\left\{\left[|X_1|^2+\left(\frac{EX_1}{1+\epsilon}\right)^2\right]1\left(X_1\le\frac{EX_1}{1+\epsilon}\right)\right\}$$
$$\le 2E|X_1|^2 1(X_1\le 0)+2E|X_1|^2 1\left(0<X_1\le\frac{EX_1}{1+\epsilon}\right)$$
$$+2\left(\frac{EX_1}{1+\epsilon}\right)^2$$
$$<\infty.$$

By Lemma 4.1, for some constant $K$ depending on $\epsilon$,

$$EN_A\le\frac{A(1+\epsilon)}{EX_1}+K,$$

so that

$$\operatorname{limsup}_{A\to\infty}\frac{EN_A}{A}\le\frac{1+\epsilon}{EX_1}.$$

Letting $\epsilon \to 0$ concludes the proof. □

*Theorem 4.1:* Let $Y_1, Y_2, \cdots$ be the sequence of rewards from an arm. Let

$$W_a(\theta)=\int_{-\infty}^{0}\prod_{b=1}^{a}\frac{f(Y_b,\theta+t)}{f(Y_b,\theta)}\,h(t)\,dt,$$

where $h: (-\infty, 0) \to R_+$ is a strictly positive continuous function with $\int_{-\infty}^{0} h(t)dt = 1$. For any $K > 0$ let

$$U(a, Y_1, \cdots, Y_a, K)=\inf\{\theta\mid W_a(\theta)\ge K\}. \tag{4.3}$$

Then for all $\lambda > \theta > \eta$,

1) $P_\theta\{\eta<U(a, Y_1, \cdots, Y_a, K)\text{ for all } a\ge 1\}\ge 1-\frac{1}{K}$,

2) $\lim_{K\to\infty}\frac{1}{\log K}\sum_{a=1}^{\infty}P_\theta\{U(a, Y_1, \cdots, Y_a, K)\ge\lambda\}=\frac{1}{I(\theta,\lambda)}$.

*Heuristics:* Having observed samples $Y_1, \cdots, Y_a$, for any $\theta \in R$, $W_a(\theta)$ is a natural statistic to test the compound hypothesis that the samples have been generated by a parameter value less than $\theta$ against the hypothesis that they have been generated by $\theta$. By the log concavity assumption (4.1), $W_a(\theta)$ is increasing in $\theta$. Therefore, for fixed $K$, for any $\theta > U(a, Y_1, \cdots, Y_a, K)$, it is

more likely that the samples have been generated by parameter values below $\theta$ than by $\theta$, whereas, for any $\theta < U(a, Y_1, \cdots, Y_a, K)$, it is more likely that the samples have been generated by $\theta$ than by parameter values below $\theta$. When we use $U(a, Y_1, \cdots, Y_a, K)$ to decide if there is a need to experiment, we choose $K$ appropriately—the larger $K$ is, the more sure we will be that the samples have been generated by parameter values below $\theta$ before we reject the possibility that they may have been generated by $\theta$. This heuristic was suggested by [7].

*Proof:* By (4.1), $W_a(\theta)$ is increasing in $\theta$, so

$$U(a, Y_1, \cdots, Y_a, K) \le \theta \Leftrightarrow W_a(\theta) \ge K.$$

Now

$$\begin{aligned}\{&U(a, Y_1, \cdots, Y_a, K)\\ &\le \eta \quad \text{for some } a \ge 1\}\\ &\subset \{U(a, Y_1, \cdots, Y_a, K) < \theta \quad \text{for some } a \ge 1\}\\ &= \{W_a(\theta) > K \quad \text{for some } a \ge 1\}.\end{aligned}$$

$W_a(\theta)$ is a nonnegative martingale under $\theta$ with mean 1. By the maximal inequality (see, e.g., [6, Lemma IV-2-9])

$$P_\theta\{W_a(\theta) \ge K \quad \text{for some } a \ge 1\} \le \frac{1}{K}$$

establishing (1).

Let $N_K = \sum_{a=1}^{\infty} 1(W_a(\lambda) < K)$. Given $\epsilon > 0$, choose $0 < \delta < \lambda - \theta$ so that

$$|I(\theta, \eta)| < \epsilon \quad \text{if } |\eta - \theta| < \delta,\ \eta > \theta.$$

Now

$$\begin{aligned}\{W_a(\lambda) < K\} &\subset \left\{\log \int_{\substack{|\eta-\theta|<\delta\\ \eta>\theta}} \prod_{b=1}^{a} \frac{f(Y_b, \eta)}{f(Y_b, \lambda)} \cdot h(\eta - \lambda)\, d\eta < \log K\right\}\\ &= \left\{\log \int_{\substack{|\eta-\theta|<\delta\\ \eta>\theta}} \prod_{b=1}^{a} \frac{f(Y_b, \eta)}{f(Y_b, \lambda)} \cdot h^o(\eta)\, d\eta < \log K - \log A\right\}\end{aligned}$$

where

$$A = \int_{\substack{|\eta-\theta|<\delta\\ \eta>\theta}} h(\eta - \lambda)\, d\eta, \quad h^o(\eta) = \frac{h(\eta - \lambda)}{A}.$$

By Jensen's inequality

$$\{W_a(\lambda) < K\} \subseteq \left\{\sum_{b=1}^{a} \int_{\substack{|\eta-\theta|<\delta\\ \eta>\theta}} \log \frac{f(Y_b, \eta)}{f(Y_b, \lambda)} \cdot h^o(\eta)\, d\eta < \log K - \log A\right\}.$$

Thus, we must examine the sum of i.i.d. variables

$$X_b = \int_{\substack{|\eta-\theta|<\delta\\ \eta>\theta}} \log \frac{f(Y_b, \eta)}{f(Y_b, \lambda)} h^o(\eta)\, d\eta$$

where $Y_b$ has distribution $f(x, \theta)$. These random variables have mean

$$EX_1 = E\left[\log \frac{f(Y_1, \theta)}{f(Y_1, \lambda)} + \int_{\substack{|\eta-\theta|<\delta\\ \eta>\theta}} \log \frac{f(Y_1, \eta)}{f(Y_1, \theta)} \cdot h^o(\eta)\, d\eta\right] \ge I(\theta, \lambda) - \epsilon > 0$$

for $\epsilon$ sufficiently small.

We proceed to verify the condition of Lemma 4.2 for the random variables $X_b$

$$0 \ge X_1 1(X_1 \le 0) \ge \int_{\substack{|\eta-\theta|<\delta\\ \eta>\theta}} \log \frac{f(Y_1, \eta)}{f(Y_1, \lambda)} 1\left(\frac{f(Y_1, \eta)}{f(Y_1, \lambda)} \le 1\right) h^o(\eta)\, d\eta,$$

$$E_\theta[X_1 1(X_1 \le 0)]^2 \le \int_{\substack{|\eta-\theta|<\delta\\ \eta>\theta}} E_\theta\left[\log \frac{f(Y_1, \eta)}{f(Y_1, \lambda)} 1\left(\frac{f(Y_1, \eta)}{f(Y_1, \lambda)} \le 1\right)\right]^2 h^o(\eta)\, d\eta.$$

Now

$$\begin{aligned}&\int f(x, \theta)\left[\log \frac{f(x, \eta)}{f(x, \lambda)}\right]^2 1\left(\frac{f(x, \eta)}{f(x, \lambda)} \le 1\right) d\nu\\ &= \int \frac{f(x, \theta) f(x, \lambda)}{f(x, \eta)} \frac{f(x, \eta)}{f(x, \lambda)} \cdot \left[\log \frac{f(x, \eta)}{f(x, \lambda)}\right]^2 1\left(\frac{f(x, \eta)}{f(x, \lambda)} \le 1\right) d\nu. \qquad (4.4)\end{aligned}$$

Observe that

$$\text{a) } x[\log x]^2 \le \frac{4}{e^2} \text{ on } \{x \le 1\}; \qquad (4.5)$$

b) Since $\lambda > \eta > \theta$, there is $0 < \alpha < 1$ such that $\eta = \alpha\theta + (1 - \alpha)\lambda$. By (4.1), for each $x$, $f(x, \eta) \ge f(x, \theta)^\alpha f(x, \lambda)^{(1-\alpha)}$. Hence,

$$\frac{f(x, \theta) f(x, \lambda)}{f(x, \eta)} \le f(x, \theta)^{(1-\alpha)} f(x, \lambda)^\alpha. \qquad (4.6)$$

Let $\eta^o = \alpha\lambda + (1 - \alpha)\theta$. By (4.1) again,

$$f(x, \theta)^{(1-\alpha)} f(x, \lambda)^\alpha \le f(x, \eta^o). \qquad (4.7)$$

Putting (4.4)–(4.7) together gives $E_\theta[X_1 1(X_1 \le 0)]^2 \le 4/e^2$.

We may now use Lemma 4.2 to conclude $E_\theta N_K < \infty$ and

$$\limsup_{K\to\infty} \frac{E_\theta N_K}{\log K} \le \frac{1}{I(\theta, \lambda) - \epsilon}.$$

Letting $\epsilon \to 0$ gives

$$\limsup_{K\to\infty} \frac{E_\theta N_K}{\log K} \le \frac{1}{I(\theta, \lambda)}. \qquad (4.8)$$

We now bound $E_\theta N_K$ from below. Define the stopping time

$$T_K = \inf\{a \ge 1 \mid W_a(\lambda) \ge K\}.$$

Observe that $N_K \ge T_K - 1$. Thus, $E_\theta T_K < \infty$. Since

$$\begin{aligned}W_a(\lambda) &= \prod_{b=1}^{a} \frac{f(Y_b, \theta)}{f(Y_b, \lambda)} \int_{-\infty}^{0} \prod_{b=1}^{a} \frac{f(Y_b, \lambda + t)}{f(Y_b, \theta)} h(t)\, dt\\ &= L_a M_a\end{aligned}$$

where $M_a$ is a martingale under $\theta$ of mean 1, we see by Doob's optional sampling theorem [6] and Wald's lemma [1] that

$$\begin{aligned}\log K &\le E_\theta \log W_{T_K}(\lambda) = E_\theta \log L_{T_K} + E_\theta \log M_{T_K} \\ &\le E_\theta \log L_{T_K} + \log E_\theta M_{T_K} \\ &= I(\theta, \lambda) E_\theta T_K \le I(\theta, \lambda) E_\theta N_K\end{aligned}$$

which, together with (4.8), establishes (2). □

*Theorem 4.2:* Let $g_{ta}(Y_1, \cdots, Y_a) = \mu[U(a, Y_1, \cdots, Y_a, t (\log t)^p)]$ for some $p > 1$. Then for any $\lambda > \theta > \eta$

1) $$P_\theta\{g_{ta}(Y_1, \cdots, Y_a) > \mu(\eta) \text{ for all } a \le t\} = 1 - O(t^{-1}(\log t)^{-p}); \tag{4.9}$$

2) $$\limsup_{t\to\infty} \frac{\sum_{a=1}^{t} P_\theta\{g_{ta}(Y_1, \cdots, Y_a) \ge \mu(\lambda)\}}{\log t} \le \frac{1}{I(\theta, \lambda)}; \tag{4.10}$$

3) $g_{ta}$ is nondecreasing in $t$ for fixed $a$. (4.11)

*Proof:* 1) follows from 1) and 2) from 2) of Theorem 4.1. 3) follows from the form of $U(a, Y_1, \cdots, Y_a, K)$ and the assumption that $\mu(\theta)$ is monotonically increasing in $\theta$. □

As estimate for the mean reward of an arm we take the sample mean

$$h_a(Y_1, \cdots, Y_a) = \frac{Y_1 + \cdots + Y_a}{a}.$$

*Lemma 4.3:* For any $0 < \delta < 1$ and $\epsilon > 0$

$$P_\theta\{\max_{\delta t \le a \le t} |h_a(Y_1, \cdots, Y_a) - \mu(\theta)| > \epsilon\} = o(t^{-1}) \tag{4.12}$$

for every $\theta$.

*Proof:* Let $Z_a = Y_a - \mu(\theta) + \epsilon$ and $S_t = Z_1 + \cdots + Z_t$. By Lemma 4.1, using (4.2),

$$\sum_{t=1}^{\infty} P_\theta\{\inf_{a \ge t} S_a \le 0\} < \infty.$$

Hence, for $\rho > 0$, there is $T(\rho)$ such that

$$\sum_{t=T(\rho)}^{\infty} P_\theta\{\inf_{a \ge t} S_a \le 0\} < \rho.$$

For any $t \ge T(\rho)/\delta^2$,

$$P_\theta\{\min_{\delta t \le a \le t} h_a(Y_1, \cdots, Y_a) < \mu(\theta) - \epsilon\} = P_\theta\{\min_{\delta t \le a \le t} S_a \le 0\} \le P_\theta\{\inf_{a \ge b} S_a \le 0\}$$

for any $\delta^2 t \le b \le \delta t$. Hence,

$$\delta(1-\delta) t P_\theta\{\min_{\delta t \le a \le t} h_a(Y_1, \cdots, Y_a) < \mu(\theta) - \epsilon\} < \rho.$$

A similar argument applies to $P_\theta\{\max_{\delta t \le a \le t} h_a(Y_1, \cdots, Y_a) > \mu(\theta) + \epsilon\}$. Letting $\rho \to 0$ concludes the proof. □

## V. An Asymptotically Efficient Allocation Rule

Let the $N$ arms correspond to $C = (\theta_1, \cdots, \theta_N)$. Assume that the arms have been reindexed so that

$$\mu(\theta_1) \ge \cdots \ge \mu(\theta_N).$$

With $g_{ta}$ and $h_a$ as in Section IV, consider the following adaptive allocation rule.

1) In the first $N$ steps sample $m$ times from each of the arms in some order to establish an initial sample.

2) Choose $0 < \delta < 1/N^2$. Consider the situation when we are about to decide which $m$ arms to sample at time $t + 1$. Clearly, whatever the preceding decisions, at least $m$ among the arms have been sampled at least $\delta t$ times. Among these "well-sampled" arms choose the *m-leaders* at stage $t + 1$, namely the arms with the $m$ best values of the statistic $\mu_t(j)$, $j = 1, \cdots, N$, where

$$\mu_t(j) = h_{T_t(j)}(Y_{j1} \cdots, Y_{jT_t(j)}).$$

Let $j \in \{1, \cdots, N\}$ be the arm for which $t + 1 \equiv j \bmod N$. Calculate the statistic $U_t(j)$ where

$$U_t(j) = g_{tT_t(j)}(Y_{j1}, \cdots, Y_{jT_t(j)}).$$

a) If arm $j$ is already one of the $m$-leaders, then at stage $t + 1$ play the $m$-leaders.

b) If arm $j$ is not among the $m$-leaders, and $U_t(j)$ is less than $\mu_t(k)$ for every $m$-leader $k$, then again play the $m$-leaders.

c) If arm $j$ is not among the $m$-leaders, and $U_t(j)$ equals or exceeds the $\mu_t$ statistic of the least best of the $m$-leaders, then play the $m - 1$ best of the $m$-leaders and the arm $j$ at stage $t$.

Note that in any case the $m - 1$ best of the $m$-leaders always get played.

*Theorem 5.1:* The rule above is asymptotically efficient.

*Proof:* The proof consists of three main steps. We first summarize the steps and indicate how they combine to yield the result. First, define $0 \le l \le m - 1$ and $m \le n \le N$ by

$$\begin{aligned}\mu(\theta_1) \ge \cdots \ge \mu(\theta_l) > \mu(\theta_{l+1}) = \cdots = \mu(\theta_m) \\ = \cdots = \mu(\theta_n) > \mu(\theta_{n+1}) \ge \cdots \ge \mu(\theta_N).\end{aligned}$$

Notice that with reference to a) at the beginning of Section III, in case $\mu(\theta_{l+1}) = \cdots = \mu(\theta_m) > \mu(\theta_{m+1})$, we are setting $n = m$, so that the $m$-border arms are in this case also the arms $X_j$, $l + 1 \le j \le n$.

Throughout the proof fix $\epsilon > 0$, satisfying $\epsilon < \mu(\theta_l) - \mu(\theta_m)/2$ if $l > 0$ and $\epsilon < \mu(\theta_n) - \mu(\theta_{n+1})/2$ if $n < N$.

*Step A:* This step is required only if $l > 0$.

$$\text{If } \mu(\theta_j) \ge \mu(\theta_l) \text{ then } E(t - T_t(j)) = o(\log t).$$

*Step B:* This step is required only if $n < N$. Define the increasing sequence of integer-valued random variables $B_t$ by

$$B_t = \#\{N \le a \le t | \text{ for some } j \ge n+1, j \text{ is one of the } m\text{-leaders at stage } a+1\}$$

where $\#\{\ \}$ denotes the number of elements in $\{\ \}$.

$$\text{Then } EB_t = o(\log t).$$

*Step C:* This step is required only if $n < N$. For each $j \ge n + 1$ define the increasing sequence of integer-valued random variables $S_t(j)$ by

$$S_t(j) = \#\{N \le a \le t | \text{ All the } m\text{-leaders at stage } a+1 \text{ are among the arms } k \text{ with } \mu(\theta_k) \ge \mu(\theta_n) \text{ and for each } m\text{-leader at stage } a+1 \ |h_{T_a(k)}(Y_{k1}, \cdots, Y_{kT_a(k)}) - \mu(\theta_k)| < \epsilon, \text{ but still the rule plays arm } j \text{ at stage } a+1\}.$$

Then, for each $\rho > 0$ we can choose $\epsilon > 0$ so small that

$$ES_t(j) \le \frac{1 + \rho + o(1)}{I(\theta_j, \theta_m)} \log t.$$

We now indicate how these steps combine to yield the theorem. 1) $R_t(\theta_1, \cdots, \theta_N) = \Sigma_{j \geq n+1} [\mu(\theta_m) - \mu(\theta_j)] ET_t(j) + o(\log t)$. Indeed, from (2.5) and (2.6) we have

$$R_t(\theta_1, \cdots, \theta_N)$$

$$= \sum_{j=1}^{l} \mu(\theta_j)(t - ET_t(j)) + \sum_{j \geq n+1} [\mu(\theta_m) - \mu(\theta_j)] ET_t(j)$$

$$+ \mu(\theta_m) \left[ \sum_{j=l+1}^{m} (t - ET_t(j)) - \sum_{j \geq m+1} ET_t(j) \right]. \quad (5.1)$$

If we observe that

$$\sum_{j=1}^{N} ET_t(j) = mt$$

we get

$$\sum_{j=l+1}^{m} t - \sum_{j=l+1}^{N} ET_t(j) = \sum_{j=1}^{l} (ET_t(j) - t)$$

so the first and third terms on the right in (5.1) are $o(\log t)$, from Step A.

*Remark:* If $n = N$ this already yields the theorem:
2) Suppose $n < N$ and $j \geq n + 1$. Then

$$T_{t+1}(j) \leq S_t(j)$$

$$+ \#\{N \leq a \leq t \mid \text{All the } m\text{-leaders at stage } a+1 \text{ are among the arms with index } \leq n, \text{ but for at least one of the } m\text{-leaders at stage } a+1, \text{ say } k, \; |h_{T_a(k)}(Y_{k1}, \cdots, Y_{kT_a(k)}) - \mu(\theta_k)| > \epsilon\}$$

$$+ B_t + N. \quad (5.2)$$

Take expectations on both sides. By Step B, $EB_t = o(\log t)$. Noting that

$$P_C\{\text{The leaders at stage } a \text{ all have index} \leq n \text{ but at least one of them, say arm } k, \text{ has } |h_{T_a(k)}(Y_{k1}, \cdots, Y_{kT_a(k)}) - \mu(\theta_k)| > \epsilon\}$$

$$\leq P_C\left\{ \max_{1 \leq i \leq N} \max_{\delta a \leq b \leq a} |h_b(Y_{i1}, \cdots, Y_{ib}) - \mu(\theta_i)| > \epsilon \right\}$$

$$= o(a^{-1}) \text{ by (4.12)}$$

we see that the expected value of the middle term on the right-hand side of (5.2) is $o(\log t)$.

By Step C we have

$$\limsup_{t \to \infty} \frac{E_C S_t(j)}{\log t} \leq \frac{1}{I(\theta_j, \theta_m)}$$

from which the theorem follows.

We now prove the individual steps.

*Proof of Step A:* Recall that this step is required only if $l > 0$. Pick a positive integer $c$, satisfying $c > (1 - N^2\delta)^{-1}$. The idea behind the choice of $c$ is that

$$\frac{t - c^{r-1}}{N} > N\delta t \qquad \text{for } t > c^r.$$

*Lemma 5.1:* Let $r$ be a positive integer. Define the sets

$$A_r = \bigcap_{1 \leq j \leq N} \left\{ \max_{\delta c^{r-1} \leq t \leq c^{r+1}} |h_t(Y_{j1}, \cdots, Y_{jt}) - \mu(\theta_j)| \leq \epsilon \right\},$$

$$B_r = \bigcap_{k \leq l} \{ g_{ta}(Y_{k1}, \cdots, Y_{ka}) \geq \mu(\theta_l) - \epsilon \text{ for } 1 \leq a \leq \delta t \text{ and } c^{r-1} \leq t \leq c^{r+1} \}.$$

Then $P_C(A_r^c) = o(c^{-r})$ and $P_C(B_r^c) = o(c^{-r})$ where $A_r^c$ and $B_r^c$ denote the complements of $A_r$ and $B_r$, respectively.

*Proof:* From (4.12) we immediately get $P_C(A_r^c) = o(c^{-r})$. From (4.9) we see that $P_C(B_r^c) = O(c^{-r} r^{-p}) = o(c^{-r})$. □

*Lemma 5.2:* On the event $A_r \cap B_r$, if $t + 1 \equiv k \bmod N$ for some $k \leq l$ and $c^{r-1} \leq t \leq c^{r+1}$, the rule plays arm $X_k$.

*Proof:* On $A_r$ the $h_a$ statistics of the $m$-leaders are all within $\epsilon$ of their actual means. If arm $X_k$ is one of the $m$-leaders at stage $t + 1$, then according to the rule it is played. Suppose $X_k$ is not an $m$-leader at stage $t + 1$. On $A_r$ the least best of the $m$-leaders at stage $t + 1$, say $j_t$, has

$$\mu_t(j_t) < \mu(\theta_l) - \epsilon.$$

In case $T_t(k) \geq \delta t$, we have on $A_r$,

$$\mu(\theta_l) - \epsilon \leq h_{T_t(k)}(Y_{k1}, \cdots, Y_{kT_t(k)}),$$

hence, our rule will play $X_k$ since it will already be one of the $m$-leaders at stage $t + 1$.

In case $T_t(k) < \delta t$, we have on $B_r$,

$$\mu(\theta_l) - \epsilon \leq U_t(k)$$

so in any case, our rule plays $X_k$. □

By Lemma 5.2, on the event $A_r \cap B_r$, for $c^r \leq t \leq c^{r+1}$, the number of times we have played arm $X_k$, $k \leq l$, exceeds

$$N^{-1}(t - c^{r-1} - 2N)$$

which exceeds $N\delta t$ if $r \geq r_0$ for some $r_0$.

*Lemma 5.3:* If $r \geq r_0$, then on the event $A_r \cap B_r$, for every $c^r \leq t \leq c^{r+1}$, we play each arm $X_k$ with $k \leq l$.

*Proof:* By Lemma 5.2, on $A_r \cap B_r$, and $c^r \leq t \leq c^{r+1}$, $r \geq r_0$, all arms $X_k$, $k \leq l$, are well sampled. Since on $A_r$, every well-sampled arm has its $h_a$ statistic $\epsilon$ close to its actual mean, all arms $X_k$, $k \leq l$ must be among the $m$-leaders. Further, they cannot be replaced by a nonleading arm's $g_{ta}$ statistic indicating the need to learn from it, because none of them is the least best of the $m$-leaders. □

*Corollary:* For $r \geq r_0$, the expected number of times arm $X_k$, $k \leq l$, is not played during $c^r \leq t \leq c^{r+1}$ is less than

$$\sum_{c^r \leq t \leq c^{r+1}} P_C(A_r^c) + P_C(B_r^c) = o(1).$$

Hence, the expected number of times arm $X_k$, $k \leq l$, is not played in $t$ steps is $o(\log t)$. □

*Proof of Step B:* Recall that this step is required only if $n < N$. The proof is identical in form to that of Step A and proceeds as follows.

*Lemma 5.1 B:* Let $A_r$ be as in Lemma 5.1 and let

$$Z_r = \bigcap_{k \leq n} \{ g_{ta}(Y_{k1}, \cdots, Y_{ka}) \geq \mu(\theta_k) - \epsilon \text{ for } 1 \leq a \leq \delta t \text{ and } c^{r-1} \leq t \leq c^{r+1} \}.$$

Then $P_C(A_r^c) = o(c^{-r})$ and $P_C(Z_r^c) = o(c^{-r})$.

*Proof:* The proof is identical to the proof of Lemma 5.1.

*Lemma 5.2B:* On the $A_r \cap Z_r$, if $t + 1 \equiv k \bmod N$ for some $k \leq n$ and $c^{r-1} \leq t \leq c^{r+1}$, then at time $t + 1$ the rule only plays arms with index $\leq n$.

*Proof:* Suppose not. Then $k$ is not one of the $m$-leaders and the least best of the $m$-leaders has index $j_r > n$ on the event $A_r$ with $\mu_t(j_t) < \mu(\theta_n) - \epsilon$.

If $T_t(k) \geq \delta t$,

$$\mu(\theta_n) - \epsilon \leq h_{T_t(k)}(Y_{k1}, \cdots, Y_{kT_t(k)})$$

on $A_r$, hence, our rule will play $X_k$; in fact, $X_k$ will already be one of the $m$-leaders at stage $t+1$.

If $T_t(k) < \delta t$,

$$\mu(\theta_n) - \epsilon \leq U_t(k)$$

on $Z_r$, hence, our rule will play $X_k$. □

Let $r_0$ be defined as in the proof of Step A. We now show that on $A_r \cap Z_r$, for $r \geq r_0 + 1$ and $c^{r-1} \leq t \leq c^{r+1}$, $m - l$ of the $m$-border arms have been played $\delta t$ times.

1) First consider the case $n = m$. For each of the $m$-border arms $X_j$, $l+1 \leq j \leq n$, there are at least $t - c^{r-1} - 2N/N > N\delta t$ times prior to $t$ at which $t+1 \equiv j \bmod N$. Choose $\delta t$ of these times. By Lemma 5.2 B, on the event $A_r \cap Z_r$ each of the arms that is played at this time has index $\leq m$. But this means that the arm $X_j$ is played at this time. Thus, we see that at stage $t+1$, all $m$-border arms are well sampled, and there are $m - l$ of them.

2) Suppose $n > m$ and that fewer than $m - l$ of the $m$-border arms have been well sampled. Let $X_j$ be one of the arms that is not well sampled, $l+1 \leq j \leq n$. There are at least $t - c^{r-1} - 2N/N > N\delta t$ times prior to $t$ at which $t+1 \equiv j \bmod N$. Choose $N\delta t$ of these times. Since arm $j$ is not well sampled, we can choose $(N-1)\delta t$ of these times at which the rule plays only arms whose indexes are $\geq n$, by Lemma 5.2 B above. We know by Lemma 5.3 that at each of these times the rule plays all arms whose indexes are $\leq l$ on the event $A_r \cap B_r$ which contains the event $A_r \cap Z_r$. Thus, $(m-l)(N-1)\delta t$ plays of $m$-border arms with index $\neq j$ are made at these times. Note that there are $n - l - 1 \geq m - l$ such arms. Also note that at these $(N-1)\delta t$ times, not one of these arms can undergo more than $(N-1)\delta t$ plays. Suppose that only $p < m - l$ of these $n - l - 1$ arms undergo $\delta t$ plays or more at these times. Then the total number of plays of these arms at these times is strictly less than

$$\begin{aligned} &p(N-1)\delta t + (n-l-1-p)\delta t \\ &\quad \leq (m-l-1)(N-1)\delta t + (N-1)\delta t \\ &\quad = (m-l)(N-1)\delta t \end{aligned}$$

which gives a contradiction.

The analog of Lemma 5.3 is as follows.

*Lemma 5.3 B:* If $r \geq r_0 + 1$, then on the event $A_r \cap Z_r$, for every $c^r \leq t \leq c^{r+1}$, the $m$-leaders are among the arms $X_k$, $k \leq n$.

*Proof:* On $A_r$ a well-sampled arm has its $h_a$ statistic $\epsilon$ close to its mean. By the above reasoning, at least $m$ of the $X_k$, $k \leq n$, are well sampled at stage $t+1$, hence the $m$-leaders are constituted of such arms. (Note that, unlike in Lemma 5.3, we do not assert that the arms that are played at such times are among the $X_k$, $k \leq n$. This is in fact false.) □

Step B follows from Lemmas 5.1 B and 5.3 B.

*Proof of Step C:* Recall that this step is required only if $n < N$. Let $j \geq n+1$. Then observe that

$$\begin{aligned} S_t(j) &\leq \#\{N \leq a \leq t \mid g_{aT_a(j)}(Y_{j1}, \cdots, Y_{jT_a(j)}) \geq \mu(\theta_m) - \epsilon \\ &\leq \#\{N \leq a \leq t \mid g_{tT_a(j)}(Y_{j1}, \cdots, Y_{jT_a(j)}) \\ &\geq \mu(\theta_m) - \epsilon\}, \text{ by (4.11)} \\ &\leq \#\{N \leq b \leq t \mid g_{tb}(Y_{j1}, \cdots, Y_{jb}) \geq \mu(\theta_m) - \epsilon\}, \end{aligned}$$

where $Y_{j1}, Y_{j2}, \cdots$ denote the rewards on plays of arm $j$. Thus,

$$\begin{aligned} E_C S_t(j) &\leq E_{\theta_j} \#\{N \leq b \leq t \mid g_{tb}(Y_{j1}, \cdots, Y_{jb}) \geq \mu(\theta_m) - \epsilon\} \\ &\leq \sum_{b=1}^{t} P_\theta\{g_{tb}(Y_{j1}, \cdots, Y_{jb}) \geq \mu(\theta_m) - \epsilon\}. \end{aligned}$$

But by (4.10) we can, for each $\rho > 0$, choose $\epsilon > 0$ so small that

$$\sum_{b=1}^{t} P_\theta\{g_{tb}(Y_{j1}, \cdots, Y_{jb}) \geq \mu(\theta_m) - \epsilon\} \leq \frac{1 + \rho + o(1)}{I(\theta_j, \theta_m)} \log t$$

which establishes Step C and Theorem 5.1. □

*Remark:* We have not examined whether our $g_{ta}$ statistics can be recursively computed, or whether there are other recursively computable $g_{ta}$ statistics satisfying (4.9), (4.10), and (4.11). For exponential families this is possible, since $U(a, Y_1, \cdots, Y_a, K)$ depends only on the sample mean. Moreover, for Bernoulli, Poisson, normal, and double exponential families, explicit recursively computable $g_{ta}$ statistics are given by Lai and Robbins [4].

## VI. Isolated Parameter Values: Lower Bound

Following Lai and Robbins [5], we will now examine the situation for multiple plays when the denseness condition (2.4) is removed. Thus some of the allowed parameter values may be isolated. For a parameter configuration $C = (\theta_1, \cdots, \theta_N)$ let $\sigma$ be a permutation of $\{1, \cdots, N\}$ such that $\mu(\theta_{\sigma(1)}) \geq \cdots \geq \mu(\theta_{\sigma(N)})$. Throughout this section and Section VII $\lambda \in \Theta$ ($\lambda$ depends on $C$) is defined as

$$\lambda = \inf\{\theta \in \Theta \mid \theta > \theta_{\sigma(m)}\}. \tag{6.1}$$

In case $\theta_{\sigma(m)} = \sup_{\theta \in \Theta} \theta$, set $\lambda = \infty$.

*Theorem 6.1:* Let the family of reward distributions satisfy (2.2) and (2.3). Let $\Phi$ be a uniformly good rule. Let $C = (\theta_1, \cdots, \theta_N)$ be a parameter configuration and $\sigma$, $\lambda$ as above. If $\lambda$ is finite, then, for each of the $m$-worst arms $j$

$$\liminf_{t \to \infty} \frac{E_C T_t(j)}{\log t} \geq \frac{1}{I(\theta_j, \lambda)}. \tag{6.2}$$

Hence

$$\liminf_{t \to \infty} \frac{R_t(\theta_1, \cdots, \theta_N)}{\log t} \geq \sum_{j \text{ is } m\text{-worst}} \frac{[\mu(\theta_{\sigma(m)}) - \mu(\theta_j)]}{I(\theta_j, \lambda)}.$$

*Proof:* Let $j$ be an $m$-worst arm. Let $C^* = (\theta_1, \cdots, \theta_{j-1}, \lambda, \theta_{j+1}, \cdots, \theta_N)$ denote the parameter configuration when the arm $j$ has parameter $\lambda$ instead of $\theta_j$. Repeating the analysis of Theorem 3.1 we see that

$$\lim_{t \to \infty} P_C\left\{T_t(j) < \frac{\log t}{(1+2\rho) I(\theta_j, \lambda)}\right\} = 0$$

for every $\rho > 0$ [see (3.5)], which proves (6.2). □

## VII. Isolated Parameter Values: An Asymptotically Efficient Rule

We call an allocation rule *asymptotically efficient* if

$$\limsup_{t \to \infty} \frac{R_t(\theta_1, \cdots, \theta_N)}{\log t} \leq \sum_{j \text{ is } m\text{-worst}} \frac{[\mu(\theta_{\sigma(m)}) - \mu(\theta_j)]}{I(\theta_j, \lambda)}$$

when $\lambda$ is finite for the parameter configuration $C = (\theta_1, \cdots, \theta_N)$, and

$$\limsup_{t \to \infty} R_t(\theta_1, \cdots, \theta_N) < \infty$$

when $\lambda = \infty$.

The allocation rule of Section V is not asymptotically efficient in case $\lambda = \infty$. Note that this means $l = 0$, i.e., there are no distinctly $m$-best arms. For a rule to be asymptotically efficient in this case means that the expected number of plays of each of the distinctly $m$-worst arms is finite. However, with the rule of

Section V the least best $\mu_t$ statistic among the $m$-leaders will fall infinitely often below $\mu(\theta_{\sigma(m)})$, while the $g_{ta}$ statistics grow in such a way that we are forced to play the $m$-worst arms infinitely often.

To get around this problem, following Lai and Robbins [5], we make a simple modification of the rule of Section V sampling from the poorer looking arms only if their $g_{ta}$ statistic exceeds the least best $\mu_t$ statistic of the $m$-leaders by a margin, with the margin decreasing to zero suitably. Let $\gamma(t)$, $t \geq 1$ decrease monotonically to zero such that, for some $q > 1$, we have, for each $\theta \in \Theta$,

$$P_\theta \{ \max_{\delta t \leq a \leq t} |h_a(Y_1, \cdots, Y_a) - \mu(\theta)| > \gamma(t)\} = O(t^{-1} (\log t)^{-q}) \quad (7.1)$$

where $h_a(Y_1, \cdots, Y_a) = Y_1 + \cdots + Y_a/a$. Such functions can be found if, for example,

$$\int |x|^4 f(x, \theta)\, d\nu(x) < \infty \qquad \text{for all } \theta \in \Theta \quad (7.2)$$

which we assume henceforth.

*Lemma:* Let the family of reward distributions satisfy (7.2). Then (7.1) holds for $\gamma(t) = Kt^{-\alpha}$ for any $K > 0$ and $0 < \alpha < 1/4$.

*Proof:* Let $S_t = Z_1 + \cdots + Z_t$, where $Z_a = Y_a - EY_a$. Then $\{S_t^4\}$ is a positive integrable submartingale. By the maximal inequality [6]

$$P_\theta \{\sup_{a \leq t} S_a^4 > K^4 t^{4(1-\alpha)}\} \leq \frac{ES_t^4}{K^4 t^{4-4\alpha}}.$$

A simple calculation gives $ES_t^4 \leq 9t^2 EZ_1^4$, from which

$$P_\theta \{ \max_{\delta t \leq a \leq t} |h_a(Y_1, \cdots, Y_a) - \mu(\theta)| > Kt^{-\alpha}\} \leq \frac{E|Y_1|^4}{K^4 t^{2-4\alpha}}$$

which is $O(t^{-1}(\log t)^{-q})$ for any $q > 1$, when $0 < \alpha < 1/4$. □

Condition (7.2) can obviously be considerably relaxed. We have not examined this issue.

We now describe the modified rule.

1) In the first $N$ steps sample $m$ times from each of the arms in some order to establish an initial sample.

2) Choose $0 < \delta < 1/N^2$. Consider the situation at stage $t + 1$, when we are about to decide which $m$ arms to play at time $t + 1$. Let $\mu_t^*$ denote the $h_a$ statistic of the least best of the $m$-leaders at stage $t + 1$. Then calculate

$$\mu_t^+ = \inf_{\theta \in \Theta} \{\mu(\theta) | \mu(\theta) > \mu_t^* + \gamma(t)\}.$$

$\mu_t^+$ could be $\infty$.

3) Let $k$ be the arm for which $t + 1 \equiv k \bmod N$. Calculate the statistic $U_t(k)$,

$$U_t(k) = g_{tT_t(k)}(Y_1, \cdots, Y_{kT_t(k)}).$$

Decide which of the arms to play at time $t + 1$ based on $\mu_t^*$ and $U_t(k)$ as follows.

a) If arm $k$ is already one of the $m$-leaders, then at time $t + 1$ play the $m$-leaders.

b) If arm $k$ is not among the $m$-leaders and $U_t(k) < \mu_t^+$, then at time $t + 1$ play the $m$-leaders.

c) If arm $k$ is not among the $m$-leaders, and $U_t(k) \geq \mu_t^+$, then play the $m - 1$ best of the $m$-leaders and the arm $k$ at time $t + 1$.

*Theorem 7.1:* The rule above is asymptotically efficient.

*Proof:* The proof consists of three steps, parallel to the proof of Theorem 5.1.

*Step A:* This step is required only if $l > 0$.

$$\text{If } \mu(\theta_j) \geq \mu(\theta_l) \text{ then } E[t - T_t(j)] < \infty.$$

*Step B:* This step is required only if $n < N$. Define the increasing sequence of integer valued random variables $B_t$ by

$$B_t = \#\{N \leq a \leq t | \text{ for some } j \geq n+1,\ j \text{ is one of the } m\text{-leaders at stage } a+1\}.$$

Then $EB_t < \infty$.

*Step C:* This step is required only if $n < N$. For each $j \geq n + 1$ define the increasing sequence of integer valued random variables $S_t(j)$ by

$$S_t(j) = \#\{N \leq a \leq t | \text{ All the } m\text{-leaders at stage } a+1 \text{ are among the arms } k \text{ with } \mu(\theta_k) \geq \mu(\theta_n), \text{ and for each } m\text{-leader at stage } a+1, |h_{T_a(k)}(Y_{k1}, \cdots, Y_{kT_a(k)}) - \mu(\theta_k)| < \gamma(t), \text{ but still the rule plays arm } j \text{ at stage } a+1\}.$$

Then, if $\lambda < \infty$, for each $\rho > 0$ we have

$$ES_t(j) \leq \frac{1 + \rho + o(1)}{I(\theta_j, \theta_m)} \log t$$

while if $\lambda = \infty$, $S_t(j) = 0$.

The argument that shows how these steps combine to prove asymptotic efficiency is identical to that of Theorem 5.1. We proceed to the individual steps.

*Proof of Step A:* This step is required only if $l > 0$. Let $c > (1 - N^2\delta)^{-1}$ be an integer, and let $r_0$ be such that

$$N^{-1}(t - c^{r+1} - 2N) \geq N\delta t,$$

$$\gamma(c^{r-1}) < \frac{\mu(\theta_l) - \mu(\theta_m)}{2}$$

(if $l > 0$), and

$$\gamma(c^{r-1}) < \frac{\mu(\theta_n) - \mu(\theta_{n+1})}{2}$$

(if $n < N$), for all $r \geq r_0$.

*Lemma 7.1:* For $r = 1, 2, \cdots$, define the sets

$$A_r = \bigcap_{1 \leq j \leq N} \{ \max_{\delta c^{r-1} \leq t \leq c^{r+1}} |h_t(Y_{j1}, \cdots, Y_{jt}) - \mu(\theta_j)| \leq \gamma(c^{r+1})\},$$

$$B_r = \bigcap_{k \leq l} \{g_{ta}(Y_{k1}, \cdots, Y_{ka}) \geq \mu(\theta_l) \text{ for } 1 \leq a \leq \delta t \text{ and } c^{r-1} \leq t \leq c^{r+1}\}.$$

Then $P_C(A_r^c) = O(c^{-r} r^{-q})$ and $P_C(B_r^c) = O(c^{-r} r^{-p})$ where $A_r^c$ and $B_r^c$ denote the complements of $A_r$ and $B_r$, respectively.

*Proof:* From (7.1) we get $P_C(A_r^c) = O(c^{-r} r^{-q})$. From (4.9) we get $P_C(B_r^c) = O(c^{-r} r^{-p})$. □

*Lemma 7.2:* For $r \geq r_0$, on the event $A_r \cap B_r$ if $t + 1 \equiv k \bmod N$ with $k \leq l$ and $c^{r-1} \leq t \leq c^{r+1}$, the rule plays arm $k$.

*Proof:* As in Lemma 5.2, we can suppose arm $k$ is not an $m$-leader at stage $t + 1$. On $A_r$ the least best of the $m$-leaders at stage $t + 1$, say $j_t$, has

$$\mu_t(j_t) \leq \mu(\theta_m) + \gamma(c^{r+1}) < \mu(\theta_l) - \gamma(c^{r+1}).$$

If $T_t(k) \geq \delta t$ we have on $A_r$

$$\mu(\theta_l) - \gamma(c^{r+1}) \leq h_{T_t(k)}(Y_{k1}, \cdots, Y_{kT_t(k)}),$$

hence, our rule will play arm $k$. In fact, arm $k$ will already be one of the $m$-leaders at stage $t + 1$.

If $T_t(k) < \delta t$, we have on $B_r$

$$\mu_t(j_t) + \gamma(t) \leq \mu_t(j_t) + \gamma(c^{r-1}) < \mu(\theta_l),$$

so that

$$\mu_t^+ < \mu(\theta_l) \le U_t(k),$$

so in any case, our rule plays arm $k$. □

The next result follows from Lemma 7.2 exactly as Lemma 5.3 followed from Lemma 5.2.

*Lemma 7.3:* If $r \ge r_0$, then on the event $A_r \cap B_r$, for every $c^r \le t \le c^{r+1}$, we play each arm $k$ with $k \le l$.

*Corollary:* For $r \ge r_0$ and $c^r \le t \le c^{r+1}$ the number of times an arm $k$, $k \le l$, is not played is less than

$$\sum_{c^r \le t \le c^{r+1}} P_C(A_r^c) + P_C(B_r^c) = O(r^{-q}) + O(r^{-p})$$

so that the number of times an arm $k$, $k \le l$, is not played is finite.

*Proof of Step B:* This step is required only if $n < N$.

*Lemma 7.1 B:* Let $A_r$ be as in Lemma 5.1 and let

$$Z_r = \bigcap_{k \le n} \{g_{ta}(Y_{k1}, \cdots, Y_{ka}) \ge \mu(\theta_k) \quad \text{for all } 1 \le a \le \delta t \text{ and } c^{r-1} \le t \le c^{r+1}\}.$$

Then $P_C(A_r^c) = O(c^{-r} r^{-q})$ and $P_C(Z_r^c) = O(c^{-r} r^{-p})$.

*Proof:* The proof is identical to the proof of Lemma 7.1. □

*Lemma 7.2 B:* For $r \ge r_0$ on the event $A_r \cap Z_r$, if $t + 1 \equiv k$ mod $N$ for some $k \le n$ and $c^{r-1} \le t \le c^{r+1}$, then either the rule plays arm $k$ at time $t + 1$ or the rule plays only arms with index $\le n$ at time $t + 1$.

*Proof:* Suppose not. Then the least best of the $m$-leaders has index $j_t > n$. If $k$ is one of the $m$-leaders, it cannot be the least of the $m$-leaders and is therefore played. If $k$ is not one of the $m$-leaders, we can consider the cases $T_t(k) \ge \delta t$ and $T_t(k) \le \delta t$ separately, as in the proof of Lemma 5.2. □

*Lemma 7.3 B:* If $r \ge r_0 + 1$, then on the event $A_r \cap Z_r$, for every $c^r \le t \le c^{r+1}$, the $m$-leaders are among the arms $X_k$, $k \le n$.

*Proof:* On $A_r$ a well-sampled arm has its $h_a$ statistic $\gamma(c^{r+1})$ close to its mean. Reasoning exactly as in Theorem 5.1, we see that the $X_k$, $k \le n$, are well-sampled at stage $t + 1$ on $A_r \cap Z_r$, hence the $m$-leaders are constituted of such arms. □

*Proof of Step C:* This step is required only if $n < N$. Let $j \ge n + 1$. From the definition of $S_t(j)$, we see

$$S_t(j) \le \#\{N \le a \le t \mid g_{aT_a(j)}(Y_{j1}, \cdots, Y_{jT_a(j)}) \ge \mu(\theta_m)\}.$$

Thus $S_t(j) = 0$ when $\lambda = \infty$. If $\lambda < \infty$, since $\gamma(t) < \epsilon$ for any $\epsilon > 0$ for all large $t$, we can argue as in the proof of Theorem 5.1 to see that for each $\rho > 0$ we can choose $\epsilon > 0$ so small that

$$\sum_{b=1}^{t} P_\theta\{g_{tb}(Y_{j1}, \cdots, Y_{jb}) \ge \lambda\} \le \frac{1 + \rho + o(1)}{I(\theta_j, \lambda)} \log(t)$$

and conclude the proof. □

## References


[1] Y. S. Chow, H. Robbins, and D. Siegmund, *Great Expectations: The Theory of Optimal Stopping.* Boston, MA: Houghton Mifflin, 1971.

[2] M. Hogan, "Moments of the minimum of a random walk and complete convergence," Dep. Statistics, Stanford Univ., Stanford, CA, Tech. Rep. 21, Jan. 1983.

[3] T. L. Lai, "Some thoughts on stochastic adaptive control," in *Proc. 23rd IEEE Conf. Decision Contr.,* Las Vegas, NV, Dec. 1984, pp. 51-56.

[4] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.,* vol. 6, pp. 4-22, 1985.

[5] ——, "Asymptotically efficient allocation of treatments in sequential experiments," in *Design of Experiments,* T. J. Santner and A. C. Tamhane, Eds. New York: Marcel Dekker, pp. 127-142.

[6] J. Neveu, *Discrete Parameter Martingales.* Amsterdam, The Netherlands: North-Holland, 1975.

[7] M. Pollack and D. Siegmund, "Approximations to the expected sample size of certain sequential tests," *Ann. Stat.,* vol. 3, pp. 1267-1282, 1975.



**Venkatachalam Anantharam** (M'86) was born on August 4, 1960 in Ernakulam, India. He received the B.Tech degree in electrical engineering from the Indian Institute of Technology, Madras, in May 1980, the M.S. and Ph.D. degrees in electrical engineering in 1982 and 1986, and the M.A. and C.Phil degrees in mathematics in 1983 and 1984 from the University of California, Berkeley.

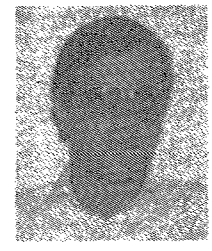
He was a Member of the Technical Staff at Bell Communications Research in Summer 1984. Since July 1986 he has been an Assistant Professor of Electrical Engineering at Cornell University, Ithaca, NY. He is the author of several technical publications.

Dr. Anantharam is a member of the American Mathematical Society and the London Mathematical Society. He was awarded the President of India Gold Medal and the Phillips India Medal in 1980 and has held several fellowships as a graduate student at Berkeley.



**Pravin Varaiya** (M'68-SM'78-F'80) received the B.S. degree from V.J.T. Institute, Bombay, India, and the M.S. and Ph.D. degrees from the University of California, Berkeley, all in electrical engineering.

He is currently Professor of Electrical Engineering and Computer Sciences and Economics at the University of California, Berkeley. He is the author, with P. R. Kumar, of *Stochastic Systems: Estimation, Identification, and Adaptive Control* (Englewood Cliffs, NJ: Prentice-Hall, 1986). His areas of research and teaching are in stochastic systems, communication networks, power systems and urban economics. He is also a Coordinator of FACHRES-CA, Faculty for Human Rights in El Salvador and Central America.


**Jean Walrand** (S'71-M'74-M'80) received the ingénieur civil degree in electrical engineering from the Université de Liège, Liège, Belgium, in 1974 and the Ph.D. degree in electrical engineering from the University of California, Berkeley, in 1979.

From 1979 to 1981 he taught at Cornell University, Ithaca, NY. Since 1981 he has been with the Department of Electrical Engineering and Computer Science at the University of California, Berkeley. His research interests are in queueing networks, communication networks, stochastic control, and stochastic processes.

Dr. Walrand is the author of *An Introduction to Queueing Networks* (Englewood Cliffs, NJ: Prentice-Hall, 1987). He has served as an Associate Editor for the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, and is on the Editorial Board of *Systems & Control Letters, Queueing Systems,* and *Probability in the Engineering and Informational Sciences.*