ASYMPTOTICALLY EFFICIENT ALLOCATION RULES FOR THE MULTIARMED BANDIT PROBLEM WITH MULTIPLE PLAYS

PART I: I.I.D. REWARDS

PART II: MARKOVIAN REWARDS

by

V. Anantharam, P. Varaiya and J. Walrand

Memorandum No. UCB/ERL M86/62

7 August 1986

ASYMPTOTICALLY EFFICIENT ALLOCATION RULES FOR THE MULTIARMED BANDIT PROBLEM WITH MULTIPLE PLAYS

PART I: I.I.D. REWARDS

PART II: MARKOVIAN REWARDS

by

V. Anantharam, P. Varaiya and J. Walrand

Memorandum No. UCB/ERL M86/62

7 August 1986

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

TITLE PAGE

ASYMPTOTICALLY EFFICIENT ALLOCATION RULES FOR THE MULTIARMED BANDIT PROBLEM WITH MULTIPLE PLAYS

PART I: I.I.D. REWARDS

PART II: MARKOVIAN REWARDS

by

V. Anantharam, P. Varaiya and J. Walrand

Memorandum No. UCB/ERL M86/62

7 August 1986

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays Part I: I.I.D. rewards[1]

V. Anantharam[2], *P. Varaiya and J. Walrand*

Department of Electrical Engineering and Computer Science
and Electronics Research Laboratory,
University of California, Berkeley, CA 94720.

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 parametrized by a real number ϑ. 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 $C = (\vartheta_1, \cdots, \vartheta_N)$ 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 the paper considers the same problem when the reward processes are Markovian.

August 6, 1986

[1] Research supported in part by JSEP Contract F49620-84-C-0057.
[2] Present address: School of Electrical Engineering, Cornell Univ., Ithaca, NY 14853.

Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays Part I: I.I.D. rewards[1]

V. Anantharam[2], *P. Varaiya and J. Walrand*

Department of Electrical Engineering and Computer Science
and Electronics Research Laboratory,
University of California, Berkeley, CA 94720.

1. 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 $f(x,\vartheta)$ with respect to some measure ν on $\mathbb{R}$. ϑ is a real valued parameter. There are N arms X_j, $j = 1, \cdots, N$ with parameter configuration $C = (\vartheta_1, \ldots, \vartheta_N)$. When arm j is played, it gives a reward with distribution $f(x,\vartheta_j)d\nu(x)$. 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 ϑ_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.

The actual values ϑ that can arise as parameters of the arms are known *a priori* to belong to a subset $\Theta \subseteq \mathbb{R}$. In §2-5, Θ is assumed to satisfy the denseness condition (2.4). This restriction is removed in §6-7.

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

[1] Research supported in part by JSEP Contract F49620-84-C-0057.
[2] Present address: School of Electrical Engineering, Cornell Univ., Ithaca, NY 14853.

2. Setup

We assume that the rewards are integrable

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

and the mean reward

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

is a strictly monotone increasing function of the parameter ϑ.

The Kullback-Liebler number,

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

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

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

and

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

In §2-5, the following denseness condition on Θ is imposed:

$$\text{for all } \lambda \in \Theta \text{ and } \delta > 0, \text{ there is } \lambda' \in \Theta \text{ s.t. } \mu(\lambda) < \mu(\lambda') < \mu(\lambda)+\delta\,. \tag{2.4}$$

Let $Y_{j1}, Y_{j2}, \cdots$ denote successive rewards from arm j. Let $F_t(j)$ denote the σ-algebra generated by $Y_{j1}, \cdots, Y_{jt}$, let $F_\infty(j) = \bigvee_t F_t(j)$, and $G(j) = \bigvee_{i \neq j} F_\infty(i)$. 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 $Y_{j1}, \ldots, Y_{jT_t(j)}$, $j = 1, \cdots, N$ and the past decisions. For an adaptive allocation rule Φ, the number of plays we have made of arm j by time t, $T_t(j)$, is a stopping time of $\{F_s(j) \vee G(j)\,,\ s \ge 1\}$. By Wald's Lemma, (see, e.g., [1], Lemma 3.1), if S_t denotes the total reward received upto time t,

$$\mathbf{E}S_t = \sum_{j=1}^{N} \mu(\vartheta_j)\mathbf{E}T_t(j)\,. \tag{2.5}$$

For a configuration $C = (\vartheta_1, \ldots, \vartheta_N)$, 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 σ be a permutation of $\{1, \cdots, N\}$ such that

$$\mu(\vartheta_{\sigma(1)}) \ge \mu(\vartheta_{\sigma(2)}) \ge \cdots \ge \mu(\vartheta_{\sigma(N)}).$$

Then the regret is

$$R_t(\vartheta_1, \cdots, \vartheta_N) = t\sum_{i=1}^{m}\mu(\vartheta_{\sigma(i)}) - \mathbf{E}S_t. \tag{2.6}$$

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 $\mu(\vartheta_i) > \mu(\vartheta_j)$ for all $1 \le i \le m$ and $m+1 \le j \le N$. However, when a parameter configuration has $\mu(\vartheta_i) < \mu(\vartheta_j)$ for some $1 \le i \le m$ and $m+1 \le j \le N$, this rule will have regret proportional to t.

We call a rule *uniformly good* if for *every* parameter configuration $R_t(\vartheta_1, \cdots, \vartheta_N) = o(t^\alpha)$ for every real $\alpha > 0$. We consider as uninteresting any rule that is not uniformly good.

3. A lower bound for the regret of a uniformly good rule

Let the arms have parameter configuration $C = (\vartheta_1, \ldots, \vartheta_N)$ and let σ be a permutation of $\{1, \cdots, N\}$ such that

$$\mu(\vartheta_{\sigma(1)}) \ge \cdots \ge \mu(\vartheta_{\sigma(N)}).$$

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

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

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

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

(c) The arms with mean equal to $\mu(\vartheta_{\sigma(m)})$ 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-border arms are the arms j, $l+1 \le j \le n$.

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

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

$$\mathbf{E}(t - T_t(j)) = o(t^\alpha),$$

and for every distinctly m-worst arm j

$$\mathbf{E}(T_t(j)) = o(t^\alpha),$$

for every real $\alpha > 0$.

Theorem 3.1 : Let the family of reward distributions satisfy conditions (2.2), (2.3) and (2.4). Let Φ be a uniformly good rule. If the arms have parameter configuration $C = (\vartheta_1, \ldots, \vartheta_N)$, then for each distinctly m-worst arm j and

each $\varepsilon > 0$

$$\lim_{t\to\infty} P_C\{T_t(j) \geq \frac{(1-\varepsilon)\log t}{I(\vartheta_j,\vartheta_{\sigma(m)})}\} = 1\,,$$

so that

$$\liminf_{t\to\infty} \frac{\mathbf{E}_C T_t(j)}{\log t} \geq \frac{1}{I(\vartheta_j,\vartheta_{\sigma(m)})}\,,$$

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

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

Consequently,

$$\liminf_{t\to\infty} \frac{R_t(\vartheta_1, \cdots, \vartheta_N)}{\log t} \geq \sum_{j \text{ is } m-\text{worst}} \frac{[\mu(\vartheta_{\sigma(m)})-\mu(\vartheta_j)]}{I(\vartheta_j,\vartheta_{\sigma(m)})}$$

for every configuration $C = (\vartheta_1, \ldots, \vartheta_N)$.

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

$$\mu(\lambda) > \mu(\vartheta_{\sigma(m)}) > \mu(\vartheta_j)$$

and

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

Consider the new configuration of parameters $C^* = (\vartheta_1,..,\vartheta_{j-1},\lambda,\vartheta_{j+1},..,\vartheta_N)$, i.e., replace ϑ_j by λ. Then arm j is one of the distinctly m-best for the parameter constellation C^*. Let $Y_1, Y_2, \cdots$ denote the sequence of rewards from plays of arm j under the uniformly good rule Φ. Define

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

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

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

We write

$$\{T_t(j) < K\log t,\ L_{T_t}(j) \leq K(1+\rho)I(\vartheta_j,\lambda)\log t\}$$

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

and

$$P_{C^*}\{T_t(j) = a,\ L_a \leq K(1+\rho)I(\vartheta_j,\lambda)\log t\}$$

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

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

Thus

$$P_{C^*}\{T_t(j) < K\log t,\ L_{T_t}(j) \leq K(1+\rho)I(\vartheta_j,\lambda)\log t\}$$
$$\geq t^{-K(1+\rho)I(\vartheta_j,\lambda)} P_C\{T_t(j) < K\log t,\ L_{T_t}(j) \leq K(1+\rho)I(\vartheta_j,\lambda)\log t\}. \qquad (3.3)$$

Since Φ is uniformly good and arm j is distinctly m-best under $C^* = (\vartheta_1,\ldots,\vartheta_{j-1},\lambda,\vartheta_{j+1},\ldots,\vartheta_N)$

$$\mathbf{E}_{C^*}(t - T_t(j)) = o(t^a)\,,$$

so that

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

hence

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

for every real $\alpha > 0$.

Choosing $K = \dfrac{1}{(1+2\rho)I(\vartheta_j,\lambda)}$, we have, from (3.2), (3.3) and (3.4),

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

Since $(1+\rho)I(\vartheta_j,\vartheta_{\sigma(m)}) \geq I(\vartheta_j,\lambda)$ by (3.1), we have

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

for every $\rho > 0$. Writing $\dfrac{1}{(1+2\rho)(1+\rho)}$ as $1-\varepsilon$ proves the first claim. Letting $\varepsilon\to 0$ proves the second claim. ∎

4. Construction of statistics

Motivated by Theorem 3.1, we call an adaptive allocation rule *asymptotically efficient* if for each configuration $C = (\vartheta_1, \ldots, \vartheta_N)$,

$$\limsup_{t\to\infty} \frac{R_t(\vartheta_1, \cdots, \vartheta_N)}{\log t} \leq \sum_{j \text{ is } m-\text{worst}} \frac{[\mu(\vartheta_{\sigma(m)}) - \mu(\vartheta_j)]}{I(\vartheta_j,\vartheta_{\sigma(m)})}\,.$$

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 ϑ_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 $g_{ta}(Y_1, \ldots, Y_a)$, $1 \le a \le t$, $t = 1,2, \cdots$, 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 $g_{ta}(Y_1, \ldots, Y_a)$ under the following assumptions

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

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

The reader may wish to glance at the beginning of §5 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 $\varepsilon > 0$

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

Let $Z_t = X_t - \frac{EX_1}{(1+\varepsilon)}$. Then

$$E\{|Z_1|^2 1(Z_1 \le 0)\} \le 2\, E\{[|X_1|^2 + (\frac{EX_1}{1+\varepsilon})^2]\, 1(X_1 \le \frac{EX_1}{1+\varepsilon})\}$$

$$\le 2\, E|X_1|^2 1(X_1 \le 0) + 2E|X_1|^2 1(0 < X_1 \le \frac{EX_1}{1+\varepsilon}) + 2(\frac{EX_1}{1+\varepsilon})^2$$

$$< \infty.$$

By Lemma 4.1, for some constant K depending on ε,

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

so that

$$\limsup_{A\to\infty} \frac{\mathbf{E}N_A}{A} \le \frac{1+\varepsilon}{\mathbf{E}X_1}.$$

Letting $\varepsilon \to 0$ concludes the proof. ∎

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

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

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

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

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

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

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

Heuristics : Having observed samples $Y_1, \cdots, Y_a$, for any $\vartheta \in \mathbb{R}$, $W_a(\vartheta)$ is a natural statistic to test the hypothesis that the samples have been generated by a parameter value less than ϑ against the hypothesis that they have been generated by ϑ. By the log concavity assumption (4.1), $W_a(\vartheta)$ is increasing in ϑ. Therefore, for fixed K, for any $\vartheta > U(a, Y_1, \ldots, Y_a, K)$, it is more likely that the samples have been generated by parameter values below ϑ than by ϑ, whereas, for any $\vartheta < U(a, Y_1, \ldots, Y_a, K)$, it is more likely that the samples have been generated by ϑ than by parameter values below ϑ. When we use $U(a, Y_1, \ldots, Y_a, K)$ 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 ϑ before we reject the possibility that they may have been generated by ϑ.

Proof : By (4.1), $W_a(\vartheta)$ is increasing in ϑ, so

$$U(a, Y_1, \ldots, Y_a, K) \le \vartheta \iff W_a(\vartheta) \ge K.$$

Now

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

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

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

establishing (1).

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

$$|I(\vartheta,\eta)| < \varepsilon \text{ if } |\eta-\vartheta| < \delta .$$

Now

$$\{W_a(\lambda) < K\} \subset \{\log \int_{\substack{|\eta-\vartheta|<\delta \\ \eta>\vartheta}} \prod_{b=1}^{a} \frac{f(Y_b,\eta)}{f(Y_b,\lambda)} h(\eta-\lambda)d\eta < \log K\}$$

$$= \{\log \int_{\substack{|\eta-\vartheta|<\delta \\ \eta>\vartheta}} \prod_{b=1}^{a} \frac{f(Y_b,\eta)}{f(Y_b,\lambda)} h^o(\eta)d\eta < \log K - \log A\},$$

where

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

By Jensen's inequality

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

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

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

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

$$EX_1 = E_\vartheta[\log \frac{f(X,\vartheta)}{f(X,\lambda)} + \int_{\substack{|\eta-\vartheta|<\delta \\ \eta>\vartheta}} \log \frac{f(X,\eta)}{f(X,\vartheta)} h^o(\eta)d\eta] \geq I(\vartheta,\lambda) - \varepsilon > 0,$$

for ε sufficiently small.

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

$$0 \geq X_1 1(X_1 \leq 0) \geq \int_{\substack{|\eta-\vartheta|<\delta \\ \eta>\vartheta}} \log \frac{f(Y_1,\eta)}{f(Y_1,\lambda)} 1(\frac{f(Y_1,\eta)}{f(Y_1,\lambda)} \leq 1) h^o(\eta)d\eta,$$

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

Now

$$\int f(x,\vartheta)[\log \frac{f(x,\eta)}{f(x,\lambda)}]^2 1(\frac{f(x,\eta)}{f(x,\lambda)} \leq 1)d\nu$$

$$= \int \frac{f(x,\vartheta)f(x,\lambda)}{f(x,\eta)} \frac{f(x,\eta)}{f(x,\lambda)} [\log \frac{f(x,\eta)}{f(x,\lambda)}]^2 1(\frac{f(x,\eta)}{f(x,\lambda)} \leq 1)d\nu. \qquad (4.4)$$

Observe that

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

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

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

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

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

Putting (4.4), (4.5), (4.6) and (4.7) together gives $\mathbf{E}_{\vartheta}[X_1 1(X_1 \le 0)]^2 \le \frac{4}{e^2}$.

We may now use Lemma 4.2 to conclude $\mathbf{E}_{\vartheta} N_K < \infty$ and

$$\limsup_{K\to\infty} \frac{\mathbf{E}_{\vartheta} N_K}{\log K} \le \frac{1}{I(\vartheta,\lambda)-\varepsilon}.$$

Letting $\varepsilon \to 0$ gives

$$\limsup_{K\to\infty} \frac{\mathbf{E}_{\vartheta} N_K}{\log K} \le \frac{1}{I(\vartheta,\lambda)}. \tag{4.8}$$

We now bound $\mathbf{E}_{\vartheta} N_K$ 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 $\mathbf{E}_{\vartheta} T_K < \infty$. Since

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

$$= L_a M_a ,$$

where M_a is a martingale under ϑ of mean 1, we see that

$$\log K \le \mathbf{E}_{\vartheta} \log W_{T_K}(\lambda) = \mathbf{E}_{\vartheta} \log L_{T_K} + \mathbf{E}_{\vartheta} \log M_{T_K}$$

$$\le \mathbf{E}_{\vartheta} \log L_{T_K} + \log \mathbf{E}_{\vartheta} M_{T_K}$$

$$= I(\vartheta,\lambda)\mathbf{E}_{\vartheta} T_K \le I(\vartheta,\lambda)\mathbf{E}_{\vartheta} N_K ,$$

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

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

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

(2) $$\limsup_{t\to\infty} \frac{\sum_{a=1}^{t} P_{\vartheta}\{g_{ta}(Y_1, \ldots, Y_a) \ge \mu(\lambda)\}}{\log t} \le \frac{1}{I(\vartheta,\lambda)};$$ (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 $\mu(\vartheta)$ is monotonically

increasing in ϑ. ∎

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

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

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

$$P_\vartheta\{\max_{\delta t \le a \le t} | h_a(Y_1, \ldots, Y_a) - \mu(\vartheta) | > \varepsilon\} = o(t^{-1}) \qquad (4.12)$$

for every ϑ.

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

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

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

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

For any $t \ge \frac{T(\rho)}{\delta^2}$,

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

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

$$\delta(1-\delta) t P_\vartheta\{\min_{\delta t \le a \le t} h_a(Y_1, \ldots, Y_a) < \mu(\vartheta) - \varepsilon\} < \rho .$$

A similar argument applies to $P_\vartheta\{\max_{\delta t \le a \le t} h_a(Y_1, \ldots, Y_a) > \mu(\vartheta) + \varepsilon\}$. Letting $\rho \to 0$ concludes the proof. ∎

5. An asymptotically efficient allocation rule

Let the N arms correspond to $C = (\vartheta_1, \ldots, \vartheta_N)$. Assume that the arms have been reindexed so that

$$\mu(\vartheta_1) \ge \cdots \ge \mu(\vartheta_N)$$

With g_{ta} and h_a as in §4, 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 $0 < \delta < \frac{1}{N^2}$. 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 δt 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 μ_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 of 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

$$\mu(\vartheta_1) \ge .. \ge \mu(\vartheta_l) > \mu(\vartheta_{l+1}) = .. = \mu(\vartheta_m) = .. = \mu(\vartheta_n) > \mu(\vartheta_{n+1}) \ge .. \ge \mu(\vartheta_N).$$

Notice that with reference to (a) at the beginning of §3, in case $\mu(\vartheta_{l+1}) = \cdots = \mu(\vartheta_m) > \mu(\vartheta_{m+1})$, 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 $\varepsilon > 0$, satisfying $\varepsilon < \frac{\mu(\vartheta_l)-\mu(\vartheta_m)}{2}$ if $l > 0$ and $\varepsilon < \frac{\mu(\vartheta_n)-\mu(\vartheta_{n+1})}{2}$ if $n < N$.

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

If $\mu(\vartheta_j) \ge \mu(\vartheta_l)$ then $\mathbf{E}(t - T_t(j)) = o(\log t)$.

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

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

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

Then $\mathbf{E}B_t = o(\log t)$.

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

$$S_t(j) = \#\{N \le \alpha \le t \mid \text{All the } m\text{-leaders at stage } \alpha + 1 \text{ are among the arms } k \text{ with } \mu(\vartheta_k) \ge \mu(\vartheta_n) \text{ and for each } m\text{-leader at stage } \alpha + 1 \mid h_{T_\alpha(k)}(Y_{k1}, \ldots, Y_{kT_\alpha(k)}) - \mu(\vartheta_k) \mid < \varepsilon,$$

but still the rule plays arm j at stage $\alpha+1$ }.

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

$$\mathbf{E}S_t(j) \le \frac{1+\rho+o(1)}{I(\vartheta_j, \vartheta_m)} \log t$$

We now indicate how these steps combine to yield the theorem.

1. $R_t(\vartheta_1, \cdots, \vartheta_N) = \sum_{j \ge n+1} [\mu(\vartheta_m) - \mu(\vartheta_j)] \mathbf{E}T_t(j) + o(\log t)$.

Indeed, from (2.5) and (2.6) we have

$$R_t(\vartheta_1, \cdots, \vartheta_N) = \sum_{j=1}^{l} \mu(\vartheta_j)(t - \mathbf{E}T_t(j)) + \sum_{j \ge n+1} [\mu(\vartheta_m) - \mu(\vartheta_j)] \mathbf{E}T_t(j)$$

$$+ \mu(\vartheta_m)[\sum_{j=l+1}^{m} (t - \mathbf{E}T_t(j)) - \sum_{j \ge m+1} \mathbf{E}T_t(j)]. \qquad (5.1)$$

If we observe that

$$\sum_{j=1}^{N} \mathbf{E}T_t(j) = mt$$

we get

$$\sum_{j=l+1}^{m} t - \sum_{j=l+1}^{N} \mathbf{E}T_t(j) = \sum_{j=1}^{l} (\mathbf{E}T_t(j) - t),$$

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 $j \ge n+1$. Then

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

$$+ \#\{N \le \alpha \le t \mid \text{All the } m\text{-leaders at stage } \alpha+1 \text{ are among the arms with index } \le n, \text{ but for at least one of the } m\text{-leaders at stage } \alpha+1, \text{ say } k,$$

$$\mid h_{T_\alpha(k)}(Y_{k1}, \ldots, Y_{kT_\alpha(k)}) - \mu(\vartheta_k) \mid > \varepsilon \}$$

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

Take expectations on both sides. By Step B, $\mathbf{E}B_t = o(\log t)$. Noting that

$$P_C\{ \text{The leaders at stage } \alpha \text{ all have index} \le n \text{ but at least one of them, say arm } k, \text{ has } \mid h_{T_\alpha(k)}(Y_{k1}, \ldots, Y_{kT_\alpha(k)}) - \mu(\vartheta_k) \mid > \varepsilon\}$$

$$\le P_C\{ \max_{1 \le i \le N} \max_{\delta\alpha \le b \le \alpha} \mid h_b(Y_{i1}, \ldots, Y_{ib}) - \mu(\vartheta_i) \mid > \varepsilon\}$$

$$= o(\alpha^{-1}) \quad \text{by (4.12)},$$

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

$$\limsup_{t\to\infty} \frac{E_C S_t(j)}{\log t} \le \frac{1}{I(\vartheta_j,\vartheta_m)}$$

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

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

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

$$A_r = \bigcap_{1\le j\le N} \{ \max_{\delta c^{r-1}\le t\le c^{r+1}} | h_t(Y_{j1},\ldots,Y_{jt}) - \mu(\vartheta_j) | \le \varepsilon \},$$

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

Then $P_C(A_r^c) = o(c^{-r})$ 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 $t+1 \equiv k \mod N$ for some $k \le l$ and $c^{r-1} \le t \le c^{r+1}$, the rule plays arm X_k.

Proof : On A_r the h_a statistics of the m-leaders are all within ε 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

$$\mu_t(j_t) < \mu(\vartheta_l) - \varepsilon .$$

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

$$\mu(\vartheta_l) - \varepsilon \le h_{T_t(k)}(Y_{k1},\ldots,Y_{kT_t(k)})$$

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,

$$\mu(\vartheta_l) - \varepsilon \le U_t(k),$$

so in any case, our rule plays X_k. ∎

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

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

which exceeds $N\delta t$ if $r \ge r_0$ 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 ε 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

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

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 5.2 B : On the event $A_r \cap Z_r$, 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 $j_t > n$ on the event A_r with $\mu_t(j_t) < \mu(\vartheta_n) - \varepsilon$.

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

$$\mu(\vartheta_n) - \varepsilon \leq h_{T_t(k)}(Y_{k1}, \ldots, Y_{kT_t(k)})$$

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 $T_t(k) < \delta t$,

$$\mu(\vartheta_n) - \varepsilon \leq U_t(k)$$

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 δt 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 $\frac{t - c^{r-1} - 2N}{N} > N\delta t$ times prior to t at which $t+1 \equiv j \bmod N$. Choose δt 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 $\frac{t-c^{r-1}-2N}{N} > N\delta t$ 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 indices are $\leq n$, by Lemma 5.2 B above. We know by Lemma 5.3 that at each of these times the rule plays all arms whose indices 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, no 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 δt plays or more at these times. Then the total number of plays of these arms at these times is strictly less than

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

which gives a contradiction.

The analog of Lemma 5.3 is

Lemma 5.3 B : If $r \geq r_0+1$, than 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 ε 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

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

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

$$\begin{aligned} \mathbf{E}_C S_t(j) &\leq \mathbf{E}_{\vartheta_j} \#\{N \leq b \leq t \mid g_{tb}(Y_{j1}, \ldots, Y_{jb}) \geq \mu(\vartheta_m)-\varepsilon\} \\ &\leq \sum_{b=1}^{t} P_\vartheta\{g_{tb}(Y_{j1}, \ldots, Y_{jb}) \geq \mu(\vartheta_m)-\varepsilon\}. \end{aligned}$$

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

$$\sum_{b=1}^{t} P_\vartheta\{g_{tb}(Y_{j1}, \ldots, Y_{jb}) \geq \mu(\vartheta_m) - \varepsilon\} \leq \frac{1+\rho+o(1)}{I(\vartheta_j, \vartheta_m)} \log t$$

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 $U(a, Y_1, \ldots, Y_a, K)$ 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].

6. 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 $C = (\vartheta_1, \ldots, \vartheta_N)$ let σ be a permutation of $\{1, \cdots, N\}$ such that $\mu(\vartheta_{\sigma(1)}) \geq \cdots \geq \mu(\vartheta_{\sigma(N)})$. Throughout this section and §7, $\lambda \in \Theta$ (λ depends on C) is defined as

$$\lambda = \inf\{\vartheta \in \Theta \mid \vartheta > \vartheta_{\sigma(m)}\} \tag{6.1}$$

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

Theorem 6.1 : Let the family of reward distributions satisfy (2.2) and (2.3). Let Φ be a uniformly good rule. Let $C = (\vartheta_1, \ldots, \vartheta_N)$ be a parameter constellation and σ, λ as above. If λ is finite, then, for each of the m-worst arms j

$$\liminf_{t \to \infty} \frac{\mathbf{E}_C T_t(j)}{\log t} \geq \frac{1}{I(\vartheta_j, \lambda)}. \tag{6.2}$$

Hence

$$\liminf_{t \to \infty} \frac{R_t(\vartheta_1, \cdots, \vartheta_N)}{\log t} \geq \sum_{j \text{ is } m-\text{worst}} \frac{[\mu(\vartheta_{\sigma(m)}) - \mu(\vartheta_j)]}{I(\vartheta_j, \lambda)}.$$

Proof : Let j be an m-worst arm. Let $C^* = (\vartheta_1, \ldots, \vartheta_{j-1}, \lambda, \vartheta_{j+1}, \ldots, \vartheta_N)$ denote the parameter configuration when the arm j has parameter λ instead of ϑ_j. Repeating the analysis of Theorem 3.1 we see that

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

for every $\rho > 0$, (see Eqn (3.5)), which proves (6.2). ∎

7. Isolated parameter values : An asymptotically efficient rule

We call an allocation rule *asymptotically effficient* if

$$\limsup_{t \to \infty} \frac{R_t(\vartheta_1, \cdots, \vartheta_N)}{\log t} \leq \sum_{j \text{ is } m-\text{worst}} \frac{[\mu(\vartheta_{\sigma(m)}) - \mu(\vartheta_j)]}{I(\vartheta_j, \lambda)}$$

when λ is finite for the parameter constellation $C = (\vartheta_1, \ldots, \vartheta_N)$, and

$$\limsup_{t\to\infty} R_t(\vartheta_1, \cdots, \vartheta_N) < \infty$$

when $\lambda = \infty$.

The allocation rule of §5 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 §5, the least best μ_t statistic among the m-leaders will fall infinitely often below $\mu(\vartheta_{\sigma(m)})$, 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 §5, sampling from the poorer looking arms only if their g_{ta} statistic exceeds the least best μ_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 $\vartheta \in \Theta$,

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

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

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

which we assume henceforth.

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

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],

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

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

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

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

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

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

μ_t^+ could be ∞.

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

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

Decide which of the arms to play at time $t+1$ based on μ_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 $l > 0$.

If $\mu(\vartheta_j) \geq \mu(\vartheta_l)$ then $\mathbf{E}[t - T_t(j)] < \infty$.

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

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

Then $\mathbf{E}B_t < \infty$.

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

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

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

$$\mathbf{E}S_t(j) \leq \frac{1+\rho+o(1)}{I(\vartheta_j, \vartheta_m)} \log t$$

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

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

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

(if $l > 0$), and

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

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

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

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

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

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

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

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

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

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

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

so that

$$\mu_t^+ < \mu(\vartheta_l) \leq U_t(k) ,$$

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

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

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

so that the number of times an arm k ,$k \leq l$, 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

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

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

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

Lemma 7.2 B : For $r \geq r_0$ on the event $A_r \cap Z_r$, if $t+1 \equiv k \bmod N$ for some $k \leq n$ and $c^{r-1} \leq t \leq c^{r+1}$, then either the rule plays arm k at time $t+1$ or the rule plays only arms with index $\leq n$ 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 $T_t(k) \geq \delta t$ and $T_t(k) \leq \delta t$ seperately, as in the proof of Lemma 5.2. ∎

Lemma 7.3 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 $\gamma(c^{r+1})$ close to its mean. Reasoning exactly as in Theorem 5.1 , we see that the X_k, $k \leq n$, 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 $j \geq n+1$. From the definition of $S_t(j)$, we see

$$S_t(j) \leq \#\{N \leq a \leq t \mid g_{aT_a(j)}(Y_{j1}, \ldots, Y_{jT_a(j)}) \geq \mu(\vartheta_m)\}.$$

Thus $S_t(j) = 0$ when $\lambda = \infty$. If $\lambda < \infty$, since $\gamma(t) < \varepsilon$ for any $\varepsilon > 0$ 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 $\varepsilon > 0$ so small that

$$\sum_{b=1}^{t} P_\vartheta\{g_{tb}(Y_{j1}, \ldots, Y_{jb}) \geq \lambda\} \leq \frac{1+\rho+o(1)}{I(\vartheta_j, \lambda)} \log(t)$$

and conclude the proof. ∎

References

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

[2] M. Hogan, *Moments of the minimum of a random walk and complete convergence*, Technical report No. 21, Department of Statistics, Stanford University, Jan 1983.

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

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

[5] T. L. Lai and H. Robbins, "Asymptotically efficient allocation of treatments in sequential experiments", in *Design of Experiments*, T. J. Santner and A. C. Tamhane (eds.), Marcel Dekker, New York, 127-142.

[6] J. Neveu, *Discrete Parameter Martingales*, North-Holland, 1975.

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

Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays Part II: Markovian rewards[1]

V. Anantharam[2], *P. Varaiya and J. Walrand*

Department of Electrical Engineering and Computer Science
and Electronics Research Laboratory,
University of California, Berkeley, CA 94720.

ABSTRACT

At each instant of time we are required to sample a fixed number $m \geq 1$ out of N Markov chains whose stationary transition probability matrices belong to a family suitably parameterized by a real number ϑ. The objective is to maximize the long run expected value of the samples. The learning loss of a sampling scheme corresponding to a parameters configuration $C = (\vartheta_1, \cdots, \vartheta_N)$ is quantified by the *regret* $R_n(C)$. This is the difference between the maximum expected reward that could be achieved if C were known and the expected reward actually achieved. We provide a lower bound for the regret associated with any uniformly good scheme, and construct a sampling scheme which attains the lower bound for every C. The lower bound is given explicity in terms of the Kullback-Liebler number between pairs of transition probabilities.

1. Introduction

We study the problem of Part I of this paper [1] when the reward statistics are Markovian and given by a one-parameter family of stochastic transition matrices $P(\vartheta) = [P(x, y, \vartheta)]$, $\vartheta \in \mathbb{R}$, $x, y \in X$, where $X \subset \mathbb{R}$ is a finite set of rewards. There are N arms X_j, $j = 1, \cdots, N$ with parameter configuration $C = (\vartheta_1, \ldots, \vartheta_N)$. Successive plays of arm j result in X-valued random variables $Y_{j1}, Y_{j2}, \cdots$ whose statistics are given by $P(\vartheta)$. The first play of an arm with parameter ϑ has reward distribution $p(\vartheta)$ which need not be the invariant distribution. We are required at each stage to play m arms. The aim is to maximize in some sense the total expected reward for every parameter configuration.

[1] Research supported in part by JSEP Contract F49620-84-C-0057
[2] Present address: School of Electrical Engineering, Cornell Univ., Ithaca, NY 14853.

We assume that

$$\text{for } x,y \in X,\ \vartheta,\vartheta' \in \mathbb{R},\ \ P(x,y,\vartheta) > 0 \Rightarrow P(x,y,\vartheta') > 0,$$

$$P(\vartheta) \text{ is irreducible and aperiodic for all } \vartheta \in \mathbb{R}, \text{ and}$$

$$p(x,\vartheta) > 0 \text{ for all } x \in X \text{ and } \vartheta \in \mathbb{R}. \tag{1.1}$$

For $\vartheta \in \mathbb{R}$, $\pi(\vartheta) = [\pi(x,\vartheta)]$, $x \in X$, denotes the invariant probability distribution on X and the mean reward

$$\mu(\vartheta) = \sum_{x \in X} x\pi(x,\vartheta) \tag{1.2}$$

is assumed to be strictly monotone increasing in ϑ.

The values that can actually arise as parameters of the arms belong to a subset $\Theta \subset \mathbb{R}$. In §2-5, Θ is assumed to satisfy the denseness condition (2.12). This restriction is removed in §6-7.

2. Setup

Let $Y_1, Y_2, \cdots$ be Markovian with state space X, initial distribution p, stationary distribution π and transition matrix P, satisfying (1.1).

Lemma 2.1 : Let F_t denote the σ-algebra generated by $Y_1, Y_2, \cdots, Y_t$ and G a σ-algebra independent of $F_\infty = \bigvee_t F_t$. Let τ be a stopping time of $\{F_t \vee G\}$. Let

$$N(x,\tau) = \sum_{a=1}^{\tau} 1(Y_a = x)$$

and

$$N(x,y,\tau) = \sum_{a=1}^{\tau-1} 1(Y_a = x, Y_{a+1} = y).$$

Then for some fixed constant K

$$|\mathbf{E}N(x,\tau) - \pi(x)\mathbf{E}\tau| \le K, \tag{2.1}$$

and

$$|\mathbf{E}N(x,y,\tau) - \pi(x)\,P(x,y)\mathbf{E}\tau| \le K, \tag{2.2}$$

for all p and all τ with $\mathbf{E}\tau < \infty$.

Proof : Let $X^* = \bigcup_{t \ge 1} X^t$, with the Borel σ-algebra of the discrete topology, i.e., all subsets are measurable. The process $\{Y_t\ ,\ t \ge 1\}$ allows to define random variables $B_1, B_2, \cdots$ called *blocks* with values in X^*. First define the $\{F_t\}$ stopping times $\tau_1, \tau_2, \cdots$ by

$$\tau_k = \inf\{t > \tau_{k-1} \mid Y_t = Y_1\}$$

with $\tau_0 = 1$. Then $\tau_k \le \infty$ a.s., and for a sample path $\omega = (y_1, y_2, \cdots)$ the kth block is the sequence $(y_{\tau_{k-1}(\omega)}, y_{\tau_{k-1}(\omega)+1}, \cdots, y_{\tau_k(\omega)-1})$. Observe that the range

of B_k is restricted to sequences whose first letter appears only once. It is simple to check that

$$F_{\tau_k} = \sigma(B_1, B_2, \cdots, B_k). \tag{2.3}$$

For $x, y \in X$, $\mathbf{y} = (y_1, y_2, \ldots, y_t) \in X^*$, let $l(\mathbf{y})$ = length of $\mathbf{y}$, $N(x, \mathbf{y})$ = number of times x appears in $\mathbf{y}$, and $N(x, y, \mathbf{y})$ = number of transitions from x to y in $\mathbf{y}$ where $y_t \to y_1$ is also considered a transition.

It is well-known, see e.g., [4] Chapter 1, Theorem (31), that $\{B_k\}$ is i.i.d. and for any $x, y \in X$

$$\mathbf{E}N(x, B_1) = \pi(x)\, \mathbf{E}\, l(B_1),$$

$$\mathbf{E}N(x, y, B_1) = \pi(x)\, P(x, y)\, \mathbf{E}\, l(B_1).$$

Let $T = \inf\{t > \tau \mid Y_t = Y_1\}$. Then $T = \tau_\kappa$, where κ is a stopping time of F_{τ_k}. Indeed $\{\tau_{k-1} \le \tau\} \in F_{\tau_{k-1}}$, see [5], Prop. II-1-5. By Wald's lemma

$$\mathbf{E}\sum_{a=1}^{T-1} 1(Y_a = x) = \mathbf{E}\sum_{k=1}^{\kappa} N(x, B_k) = \pi(x)\, \mathbf{E}l(B_1)\, \mathbf{E}\kappa, \tag{2.4}$$

$$\mathbf{E}\sum_{a=1}^{T-1} 1(Y_a = x,\ Y_{a+1} = y) = \mathbf{E}\sum_{k=1}^{\kappa} N(x, y, B_k) = \pi(x)\, P(x, y)\, \mathbf{E}l(B_1)\, \mathbf{E}\kappa, \tag{2.5}$$

$$\mathbf{E}(T-1) = \mathbf{E}\sum_{k=1}^{\kappa} l(B_k) = \mathbf{E}l(B_1)\, \mathbf{E}\kappa. \tag{2.6}$$

Observe that for a fixed constant K independent of p and τ, $\mathbf{E}(T-\tau) \le K$. In fact the mean time to visit any state starting at Y_τ is finite.

For $x \in X$,

$$N(x,T) - (T-\tau) \le N(x,\tau) < N(x,T)\,.$$

Using (2.4), (2.5) and (2.6),

$$\pi(x)\, \mathbf{E}(T-1) - K \le \mathbf{E}N(x,\tau) < \pi(x)\, \mathbf{E}(T-1) + 1,$$

so that

$$\pi(x)\, \mathbf{E}\tau - K \le \mathbf{E}N(x,\tau) \le \pi(x)\, \mathbf{E}\tau + K. \tag{2.7}$$

For $x, y \in X$,

$$N(x,y,T) - (T-\tau) \le N(x,y,\tau) \le N(x,y,T).$$

Using (2.4), (2.5) and (2.6),

$$\pi(x)\, P(x,y)\mathbf{E}(T-1) - K \le \mathbf{E}N(x,y,\tau) \le \pi(x)\, P(x,y)\mathbf{E}(T-1),$$

so that

$$\pi(x)\, P(x,y)\, \mathbf{E}\tau - K \le \mathbf{E}N(x,y,\tau) \le \pi(x)\, P(x,y)\, \mathbf{E}\tau + K. \tag{2.8}$$

The result follows from (2.7) and (2.8). ∎

Let $Y_{j1}, Y_{j2}, \cdots$ denote the succesive rewards from arm j. Let $F_t(j)$ denote the σ-algebra generated by $Y_{j1}, \cdots, Y_{jt}$, $F_\infty(j) = \bigvee_{t \geq 1} F_t(j)$, and $G(j) = \bigvee_{i \neq j} F_\infty(i)$. As in §2 of [1], 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 $Y_{j1}, \ldots, Y_{jT_t(j)}$, $j = 1, \cdots, N$ and the past decisions. For an adaptive allocation rule Φ the number of plays we have made of arm j at time t, $T_t(j)$, is a stopping time of $\{F_s(j) \vee G(j),\ s \geq 1\}$. The total reward is

$$S_t = \sum_{j=1}^{N} \sum_{a=1}^{T_t(j)} Y_{ja} = \sum_{j=1}^{N} \sum_{x \in X} xN(x, T_t(j)) .$$

By Lemma 2.1,

$$|\mathbf{E}S_t - \sum_{j=1}^{N} \mu(\vartheta_j)\mathbf{E}T_t(j)| \leq \text{const} , \tag{2.9}$$

where the constant may depend on the parameter configuration, but not on t.

As in the i.i.d. case, the loss associated to an adaptive allocation rule Φ and a configuration $C = (\vartheta_1, \ldots, \vartheta_N)$ is a function of the number of plays t, called the ***regret***. It is the difference between the maximum expected reward that could have been achieved with prior knowledge of C and the actual expected reward. Let σ be a permutation of $\{1, \cdots, N\}$ such that

$$\mu(\vartheta_{\sigma(1)}) \geq \mu(\vartheta_{\sigma(2)}) \geq \cdots \geq \mu(\vartheta_{\sigma(N)}) .$$

Then the regret is

$$R_t(\vartheta_1, \cdots, \vartheta_N) = t \sum_{i=1}^{m} \mu(\vartheta_{\sigma(i)}) - \mathbf{E}S_t .$$

By (2.9),

$$|R_t(\vartheta_1, \cdots, \vartheta_N) - [t \sum_{i=1}^{m} \mu(\vartheta_{\sigma(i)}) - \sum_{j=1}^{N} \mu(\vartheta_{\sigma(j)})\mathbf{E}T_t(j)]| \leq \text{const}, \tag{2.10}$$

where the constant can depend on the C.

An allocation rule is called ***uniformly good*** if for every configuration $R_t(\vartheta_1, \cdots, \vartheta_N) = o(t^\alpha)$ for every $\alpha > 0$.

Let P and Q be irreducible and aperiodic stochastic matrices with P having invariant distribution π, which satisfy $P(x, y) > 0 \Leftrightarrow Q(x, y) > 0$. The Kullback-Liebler number

$$I(P, Q) = \sum_{x \in X} \pi(x) \sum_{y \in X} P(x, y) \log \frac{P(x, y)}{Q(x, y)},$$

is a well-known measure of dissimilarity between P and Q. Note that $I(P, Q)$ is just the expectation with respect to the invariant measure of P of the Kullback-Liebler numbers between the individual rows of P and Q thought of as probability distributions on X. Let $I(\vartheta, \lambda)$ denote $I(P(\vartheta), P(\lambda))$. Under (1.1) and (1.2), $0 < I(\vartheta, \lambda) < \infty$ for $\vartheta \neq \lambda$. We assume that

$$I(\vartheta,\lambda) \text{ is continuous in } \lambda > \vartheta \text{ for fixed } \vartheta. \tag{2.11}$$

In §2-5 we also assume the following denseness condition on Θ:

$$\text{for all } \lambda \in \Theta \text{ and } \delta > 0, \text{ there is } \lambda' \in \Theta \text{ s.t. } \mu(\lambda) < \mu(\lambda') < \mu(\lambda) + \delta. \tag{2.12}$$

3. A lower bound for the regret of a uniformly good rule

For a parameter configuration $C = (\vartheta_1, \ldots, \vartheta_N)$, define the notions of m-best, m-worst and m-border arms exactly as in § 3 of [1]. By (2.10), an adaptive allocation rule Φ is uniformly good iff for every distinctly m-best arm j

$$\mathbf{E}(t - T_t(j)) = o(t^\alpha),$$

and for every distinctly m-worst arm j

$$\mathbf{E}(T_t(j)) = o(t^\alpha),$$

for every real $\alpha > 0$.

Theorem 3.1 : Let the family of reward distributions satisfy conditions (2.11) and (2.12). Let Φ be a uniformly good rule. If the arms have parameter configuration $C = (\vartheta_1, \ldots, \vartheta_N)$, then for each distinctly m-worst arm j and each $\varepsilon > 0$,

$$\lim_{t\to\infty} P_C\{T_t(j) \geq \frac{(1-\varepsilon)\log t}{I(\vartheta_j, \vartheta_{\sigma(m)})}\} = 1,$$

so that

$$\liminf_{t\to\infty} \frac{\mathbf{E}_C T_t(j)}{\log t} \geq \frac{1}{I(\vartheta_j, \vartheta_{\sigma(m)})},$$

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

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

Consequently,

$$\liminf_{t\to\infty} \frac{R_t(\vartheta_1, \cdots, \vartheta_N)}{\log t} \geq \sum_{j \text{ is } m-\text{worst}} \frac{[\mu(\vartheta_{\sigma(m)}) - \mu(\vartheta_j)]}{I(\vartheta_j, \vartheta_{\sigma(m)})}.$$

Proof: As in the proof of Theorem 3.1 of [1], let j be an m-worst arm and, for any $\rho > 0$, choose λ satisfying

$$\mu(\lambda) > \mu(\vartheta_{\sigma(m)}) > \mu(\vartheta_j), \text{ and } |I(\vartheta_j,\lambda) - I(\vartheta_j, \vartheta_{\sigma(m)})| \leq \rho I(\vartheta_j, \vartheta_{\sigma(m)})$$

which is possible by (2.11) and (2.12).

Consider the new configuration of parameters $C^* = (\vartheta_1, \ldots, \vartheta_{j-1}, \lambda, \vartheta_{j+1}, \ldots, \vartheta_N)$. Let $Y_1, Y_2, \cdots$ denote the sequence of rewards from plays of arm j under the uniformly good rule Φ. Define

$$L_t = \log \frac{p(Y_1, \vartheta_j)}{p(Y_1, \lambda)} + \sum_{a=1}^{t-1} \log \frac{P(Y_a, Y_{a+1}, \vartheta_j)}{P(Y_a, Y_{a+1}, \lambda)}.$$

By (1.1) and the ergodic theorem $\frac{L_t}{t} \to I(\vartheta_j, \lambda)$ a.s. $[P_C]$. Hence $\frac{1}{t}\max_{a \leq t} L_a \to$

$I(\vartheta_j,\lambda)$ a.s. $[P_C]$. For any $K > 0$ we have

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

After this point the proof proceeds exactly as in Theorem 3.1 of [1]. ∎

4. Construction of statistics

An allocation rule is *asymptotically efficient* if for each $C = (\vartheta_1, \ldots, \vartheta_N)$

$$\limsup_{t\to\infty} \frac{R_t(\vartheta_1, \cdots, \vartheta_N)}{\log t} \le \sum_{j \text{ is } m-worst} \frac{[\mu(\vartheta_{\sigma(m)})-\mu(\vartheta_j)]}{I(\vartheta_j,\vartheta_{\sigma(m)})}.$$

We will construct an asymptotically efficient rule using a family of statistics $g_{ta}(Y_1, \ldots, Y_a)$, $2 \le a \le t$, $t = 2, 3, \cdots$ as in §4 of [1], under the following assumption:

$$\text{for } x, y \in X, \log P(x,y,\vartheta) \text{ is a concave function of } \vartheta. \tag{4.1}$$

The following lemmas are needed later.

Lemma 4.1 : Let $Y_1, Y_2, \cdots$ be Markovian with finite state space X, transition matrix P, invariant distribution π and initial distribution p. Let $f : X \to \mathbb{R}$ be such that $\sum_{x\in X} \pi(x)f(x) > 0$ and let $S_t = \sum_{a=1}^{t} f(Y_a)$. Let $L = \sum_{t=1}^{\infty} 1(\inf_{a\ge t} S_a \le 0)$. Then $\mathbf{E}L < \infty$.

Proof : We appeal to the large deviations theory for the empirical distribution of a finite state Markov chain, see especially [2] and [3]. Let $\mathbf{M}$ be the unit simplex in $\mathbb{R}^{|X|}$ identified with the space of probability measures on X. Define $F : \mathbf{M} \to \mathbb{R}$ by $F(\nu) = \sum_{x\in X} f(x)\nu(x)$ and let $K = \{\nu \in \mathbf{M} \mid F(\nu) \le 0\}$. K is closed and $\pi \notin K$.

The process $\{Y_t\}$ defines for each $t \ge 1$ a probability measure Q_t on $\mathbf{M}$ which is the distribution of the t-sample empirical distribution of $\{Y_t\}$. By the ergodic theorem $Q_t \to \delta_\pi$ weakly as probability measures on $\mathbf{M}$. From the large deviations theory for this weak convergence, [3] Theorem II.1, there are constants $A > 0$, $\alpha > 0$ such that

$$Q_t(K) < Ae^{-\alpha t} \text{ for all } t \ge 1.$$

Now

$$S_t = \sum_{x\in X} N(x,t)f(x),$$

so that

$$Q_t(K) = \mathbf{E}1(S_t \le 0),$$

and the result follows. ∎

Lemma 4.2 : Let $\{Y_t, t \ge 1\}$, P, π, p be as in Lemma 4.1 and $f : X^2 \to \mathbb{R}$ be such that $\sum_{x,y\in X} \pi(x)P(x,y)f(x,y) > 0$. For $t \ge 2$ let $S_t = \sum_{a=1}^{t-1} f(Y_a, Y_{a+1})$. Let

$N = \sum_{t=2}^{\infty} 1(S_t \le 0)$. Then $\mathbf{E}N < \infty$.

Proof : We appeal to the large deviations theory for the empirical transition count matrix of a finite state Markov chain, see [2]. Let $\mathbf{M}^{(2)}$ be the unit simplex in $\mathbb{R}^{|X|^2}$ identified with the space of probability measures on X^2, and define $F : \mathbf{M}^{(2)} \to \mathbb{R}$ by $F(\nu) = \sum_{x,y \in X} f(x,y)\nu(x,y)$. Let $K = \{\nu \in \mathbf{M}^{(2)} \mid F(\nu) \le 0\}$. Let $\pi P \in \mathbf{M}^{(2)}$ be given by $\pi P(x,y) = \pi(x)P(x,y)$. Then K is closed and $\pi P \notin K$.

$\{Y_t\}$ defines for each $t \ge 2$ a probability measure $Q_t^{(2)}$ on $\mathbf{M}^{(2)}$ which is the distribution of the $\mathbf{M}^{(2)}$ valued random variable whose component in the (x,y) direction is $\frac{N(x,y,t)}{t-1}$. Then $Q_t^{(2)} \to \delta_{\pi P}$ weakly as probability measures on $\mathbf{M}^{(2)}$. From the large deviations theory, [2] Problem IX.6.12, there are constants $A > 0$, $\alpha > 0$ such that

$$Q_t^{(2)}(K) < Ae^{-\alpha t} \text{ for all } t \ge 2.$$

Now

$$S_t = \sum_{x,y \in X} N(x,y,t) f(x,y),$$

so that

$$Q_t^{(2)}(K) = \mathbf{E}1(S_t \le 0),$$

from which the result follows. ∎

Lemma 4.3 : With the same conditions as in Lemma 4.2, write μ for $\sum_{x,y \in X} \pi(x)P(x,y)f(x,y)$. Given $A > 0$, let $N_A = \sum_{t=2}^{\infty} 1(S_t \le A)$. Then

$$\limsup_{A \to \infty} \frac{\mathbf{E}N_A}{A} \le \frac{1}{\mu}.$$

Proof : For any $\varepsilon > 0$,

$$N_A \le \frac{A(1+\varepsilon)}{\mu} + 1 + \sum_{t=2}^{\infty} 1[S_t \le (t-1)\frac{\mu}{1+\varepsilon}].$$

Let $g(x,y) = f(x,y) - \frac{\mu}{1+\varepsilon}$. Then $\sum_{x,y \in X} \pi(x)P(x,y)g(x,y) > 0$ and $\{S_t \le (t-1)\frac{\mu}{1+\varepsilon}\} = \{\sum_{a=1}^{t-1} g(Y_a, Y_{a+1}) \le 0\}$, so by Lemma 4.2,

$$\mathbf{E}N_A \le \frac{A(1+\varepsilon)}{\mu} + \text{const},$$

for some constant depending on ε. Thus

$$\limsup_{A \to \infty} \frac{\mathbf{E}N_A}{A} \le \frac{1+\varepsilon}{\mu}.$$

Letting $\varepsilon \to 0$ yields the result.

Theorem 4.1 : Let $Y_1, Y_2, \cdots$ be the sequence of rewards from an arm. For $a \geq 2$ write $P^a(Y^a)$ for $P(Y_1, Y_2) \cdots P(Y_{a-1}, Y_a)$. For $a \geq 2$, let

$$W_a(\vartheta) = \int_{-\infty}^{0} \frac{P^a(Y^a, \vartheta+t)}{P^a(Y^a, \vartheta)} h(t)dt \ ,$$

where h: $(-\infty, 0) \to \mathbb{R}_+$ is a positive continuous function satisfying $\int_{-\infty}^{0} h(t)dt = 1$.

For any $K > 0$, let

$$U(a, Y_1, \cdots, Y_a, K) = \inf \{\vartheta \mid W_a(\vartheta) \geq K\}. \tag{4.2}$$

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

(1) $P_\vartheta\{\eta < U(a, Y_1, \cdots, Y_a, k) \text{ for all } a \geq 2\} \geq 1 - \frac{1}{K}$,

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

Heuristics : The reason for introducing U is similar to that in Theorem 4.1 of [1].

Proof : By (4.1), W_a is increasing in ϑ, so

$$U(a, Y_1, \ldots, Y_a, K) < \vartheta \iff W_a(\vartheta) \geq K.$$

Now

$$\begin{aligned} \{U(a, Y_1, \ldots, Y_a, K) \leq \eta \text{ for some } a \geq 2\} & \\ \subset \{U(a, Y_1, \ldots, Y_a, K) < \vartheta \text{ for some } a \geq 2\} & \\ = \{W_a(\vartheta) \geq K \text{ for some } a \geq 2\} \ . & \end{aligned}$$

$W_a(\vartheta)$ is a nonnegative martingale under ϑ with mean 1. By the maximal inequality,

$$P_\vartheta\{W_a(\vartheta) \geq K \text{ for some } a \geq 2\} \leq \frac{1}{K},$$

establishing (1).

Let $N_K = \sum_{a=2}^{\infty} 1(W_a(\lambda) < K)$. Given $\varepsilon > 0$, choose $\delta > 0$ so that $|I(\vartheta, \eta)| < \varepsilon$ if $|\eta - \vartheta| < \delta$. Now

$$\begin{aligned} \{W_a(\lambda) < K\} & \subset \{\log \int_{\substack{|\eta - \vartheta| < \delta \\ \eta > \vartheta}} \frac{P^a(Y^a, \eta)}{P^a(Y^a, \lambda)} h(\eta - \lambda)d\eta < \log K\} \\ & = \{\log \int_{\substack{|\eta - \vartheta| < \delta \\ \eta > \vartheta}} \frac{P^a(Y^a, \eta)}{P^a(Y^a, \lambda)} h^\circ(\eta)d\eta < \log K - \log A\} \ , \end{aligned}$$

where

$$A = \int_{\substack{|\eta - \vartheta| < \delta \\ \eta > \vartheta}} h(\eta - \lambda)d\eta \quad \text{and} \quad h^\circ(\eta) = \frac{h(\eta - \lambda)}{A}.$$

By Jensen's inequality

$$\{W_a(\lambda) < K\} \subset \{\int_{\substack{|\eta-\vartheta|<\delta \\ \eta>\vartheta}} \log \frac{P^a(Y^a,\eta)}{P^a(Y^a,\lambda)} h^o(\eta)d\eta < \log K - \log A\}.$$

Now

$$\sum_{x,y\in X} \pi(x,\vartheta)P(x,y,\vartheta) \int_{\substack{|\eta-\vartheta|<\delta \\ \eta>\vartheta}} \log \frac{P(x,y,\eta)}{P(x,y,\lambda)} h^o(\eta)d\eta$$

$$= \sum_{x,y\in X} \pi(x,\vartheta)P(x,y,\vartheta)[\log \frac{P(x,y,\vartheta)}{P(x,y,\lambda)} - \int_{\substack{|\eta-\vartheta|<\delta \\ \eta>\vartheta}} \log \frac{P(x,y,\vartheta)}{P(x,y,\eta)} h^o(\eta)d\eta]$$

$$= I(\vartheta,\lambda) - \int_{\substack{|\eta-\vartheta|<\delta \\ \eta>\vartheta}} I(\vartheta,\eta)\, h^o(\eta)d\eta$$

$$\geq I(\vartheta,\lambda) - \varepsilon > 0$$

for ε sufficiently small. By Lemma 4.3, $\mathbf{E}N_K < \infty$ and

$$\limsup_{K\to\infty} \frac{\mathbf{E}_\vartheta N_K}{\log K} \leq \frac{1}{I(\vartheta,\lambda)-\varepsilon}.$$

Letting $\varepsilon \to 0$ gives

$$\limsup_{K\to\infty} \frac{\mathbf{E}_\vartheta N_K}{\log K} \leq \frac{1}{I(\vartheta,\lambda)}. \tag{4.3}$$

To bound $\mathbf{E}_\vartheta N_K$ from below, define the stopping time

$$T_K = \inf\{a \geq 2 \mid W_a(\lambda) \geq K\}.$$

Observe that $N_K \geq T_K - 1$. Thus $\mathbf{E}_\vartheta T_K < \infty$. Since

$$W_a(\lambda) = \frac{P^a(Y^a,\vartheta)}{P^a(Y^a,\lambda)} \int_{-\infty}^{0} \frac{P^a(Y^a,\lambda+t)}{P^a(Y^a,\vartheta)} h(t)dt = L_a M_a,$$

where M_a is a martingale under ϑ with mean 1, we obtain

$$\log K \leq \mathbf{E}_\vartheta \log W_{T_K}(\lambda) = \log \mathbf{E}_\vartheta L_{T_K} + \mathbf{E}_\vartheta \log M_{T_K}$$

$$\leq \mathbf{E}_\vartheta \log L_{T_K} + \log \mathbf{E}_\vartheta M_{T_K}$$

$$= \mathbf{E}_\vartheta \log L_{T_K}. \tag{4.4}$$

Now

$$\mathbf{E}_\vartheta \log L_{T_K} = \sum_{x,y\in X} \mathbf{E}_\vartheta N(x,y,T_K) \log \frac{P(x,y,\vartheta)}{P(x,y,\lambda)},$$

and by Lemma 2.1

$$|\mathbf{E}_\vartheta N(x,y,T_K) - \pi(x,\vartheta)\, P(x,y,\vartheta)\mathbf{E}_\vartheta T_K| \leq \text{const.}$$

Hence

$$|\mathbf{E}_\vartheta \log L_{T_K} - I(\vartheta,\lambda)\mathbf{E}_\vartheta T_K| \le \text{const.} \tag{4.5}$$

From (4.4) and (4.5), and using $N_K \ge T_K - 1$, we have

$$\liminf_{K\to\infty} \frac{\mathbf{E}_\vartheta N_K}{\log K} \ge \frac{1}{I(\vartheta,\lambda)}$$

which, together with (4.3), establishes (2). ∎

Theorem 4.2 : Fix $p > 1$. For $t = 2, 3, \cdots$ and $2 \le a \le t$, let $g_{ta}(Y_1, \ldots, Y_a) = \mu[U(a, Y_1, \cdots, Y_a, t(\log t)^p\ ]$. Then for all $\lambda > \vartheta > \eta$,

(1) $P_\vartheta\{g_{ta}(Y_1, \ldots, Y_a) > \mu[\eta] \text{ for all } 2 \le a \le t\} = 1 - O(t^{-1}(\log t)^{-p})$, (4.6)

(2) $\limsup_{t\to\infty} \sum_{a=2}^{t} \frac{P_\vartheta\{g_{ta}(Y_1, \ldots, Y_a) \ge \mu(\lambda)\}}{\log t} \le \frac{1}{I(\vartheta,\lambda)}$, (4.7)

(3) g_{ta} is nondecreasing in t for fixed a. (4.8)

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

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

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

Lemma 4.4 : For any $0 < \delta < 1$ and $\varepsilon > 0$

$$P_\vartheta\{\max_{\delta t \le a \le t} |h_a(Y_1, \ldots, Y_a) - \mu(\vartheta)| > \varepsilon\} = o(t^{-1}) \tag{4.9}$$

for every ϑ.

Proof : Consider $f(x) = x - \mu(\vartheta) + \varepsilon$. Then $\sum_{x\in X} \pi(x,\vartheta) f(x) > 0$. By Lemma 4.1, for any $\rho > 0$, there is $T(\rho)$ such that

$$\sum_{t=T(\rho)}^{\infty} P_\vartheta\{\inf_{a \ge t} S_a\} < \rho\ ,$$

where $S_t = \sum_{a=1}^{t} f(Y_a)$. For any $t \ge \frac{T(\rho)}{\delta^2}$

$$P_\vartheta\{\min_{\delta t \le a \le t} h_a(Y_1, \ldots, Y_a) < \mu(\vartheta) - \varepsilon\} = P_\vartheta\{\min_{\delta t \le a \le t} S_a \le 0\}$$

$$\le P_\vartheta\{\inf_{a \ge b} S_b \le 0\}$$

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

$$\delta(1-\delta)t\, P_\vartheta\{\min_{\delta t \le a \le t} h_a(Y_1, \ldots, Y_a) < \mu(\vartheta) - \varepsilon\} < \rho.$$

A similar argument applies to $P_\vartheta\{\max_{\delta t \le a \le t} h_a(Y_1, \ldots, Y_a) > \mu(\vartheta) + \varepsilon\}$. Letting $\rho \to 0$ concludes the proof. ∎

5. An asymptotically efficient rule

Consider the allocation rule of §5 of [1] using the g_{ta} and h_a statistics constructed in §4 above, and an initial sample of size $2N$ to initiate the g_{ta} statistics.

Theorem 5.1 : The rule above is asymptotically efficient.

Proof : Reindex the arms so that $\mu(\vartheta_1) \geq \cdots \geq \mu(\vartheta_N)$. Let $0 \leq l \leq m-1$ and $m \leq n \leq N$ be defined as in the proof of Theorem 5.1 of [1]. Given the properties (4.6), (4.7), (4.8) and (4.9) of the g_{ta} and h_a statistics which we have already established, the proof of Theorem 5.1 of [1] carries over word for word to establish the following assertions A, B, and C.

A: If $l > 0$, then $\mathbf{E}(t - T_t(j)) = o(\log t)$ for every $j \leq l$.

B: If $n < N$, let

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

Then $\mathbf{E}B_t = o(\log t)$.

C: If $n < N$ and $0 < \varepsilon < \mu(\vartheta_n) - \mu(\vartheta_{n+1})$, then for $j \geq n+1$ let

$$\begin{aligned} S_t(j) = \#\{N \leq a \leq t \mid & \text{All the } m\text{-leaders at stage } a+1 \text{are among the} \\ & \text{arms } k \text{ with } \mu(\vartheta_k) \geq \mu(\vartheta_n), \text{ and} \\ & \text{for each } m\text{-leader at stage } a+1 \\ & \mid h_{T_a(k)}(Y_{k1}, \ldots, Y_{kT_a(k)}) - \mu(\vartheta_k) \mid < \varepsilon, \text{but} \\ & \text{still the rule samples from arm } j \text{ at stage } a+1\}. \end{aligned}$$

For each $\rho > 0$ we can then choose $\varepsilon > 0$ so small that

$$\mathbf{E}S_t(j) \leq \frac{1+\rho+o(1)}{I(\vartheta_j, \vartheta_m)} \log t \ .$$

As indicated in Theorem 5.1 of [1], these steps can be combined to obtain

$$\limsup_{t \to \infty} \frac{t \sum_{i=1}^{m} \mu(\vartheta_i) - \sum_{j=1}^{N} \mu(\vartheta_j) \mathbf{E}T_t(j)}{\log t} \leq \sum_{j \text{ is } m\text{-worst}} \frac{\mu(\vartheta_m) - \mu(\vartheta_j)}{I(\vartheta_j, \vartheta_m)},$$

from which the proof follows using (2.10). ∎

6. Isolated parameter values : Lower bound

We proceed to examine the situation in the absence of the denseness condition (2.12). For a configuration $C = (\vartheta_1, \ldots, \vartheta_N)$, let σ be a permutation of $\{1, \cdots, N\}$ such that $\mu(\vartheta_{\sigma(1)}) \geq \cdots \geq \mu(\vartheta_{\sigma(N)})$. Throughout this section and §7, $\lambda \in \Theta$ (λ depending on C) is defined as

$$\lambda = \inf\{\vartheta \in \Theta \mid \vartheta > \vartheta_{\sigma(m)}\}.$$

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

Theorem 6.1 : Let the family of reward distributions satisfy (2.11). Let Φ be a uniformly good rule. Let $C = (\vartheta_1, \ldots, \vartheta_N)$ be a configuration and σ, λ as above. If $\lambda < \infty$, then, for each distinctly m-worst arm j,

$$\liminf_{t\to\infty} \frac{\mathbf{E}_C T_t(j)}{\log t} \geq \frac{1}{I(\vartheta_j,\lambda)}.$$

Consequently, by (2.10),

$$\liminf_{t\to\infty} \frac{R_t(\vartheta_1, \cdots, \vartheta_N)}{\log t} \geq \sum_{j \text{ is } m-\text{worst}} \frac{(\mu(\vartheta_{\sigma(m)})-\mu(\vartheta_j))}{I(\vartheta_j,\lambda)}$$

for each C.

Proof: Let j be an m-worst arm. Consider the parameter configuration $C^* = (\vartheta_1,..,\vartheta_{j-1},\lambda,\vartheta_{j+1},..,\vartheta_N)$ when the arm j has parameter λ instead of ϑ_j and proceed as in Theorem 3.1 . ∎

7. Isolated parameter values : An asymptotically efficient rule

As in §7 of [1], an allocation rule is called *asymptotically efficient* if

$$\limsup_{t\to\infty} \frac{R_t(\vartheta_1, \cdots, \vartheta_N)}{\log t} \leq \sum_{j \text{ is } m-\text{worst}} \frac{(\mu(\vartheta_{\sigma(m)})-\mu(\vartheta_j))}{I(\vartheta_j,\lambda)}$$

when λ is finite for the configuration $C = (\vartheta_1, \ldots, \vartheta_N)$, and

$$\limsup_{t\to\infty} R_t(\vartheta_1, \cdots, \vartheta_N) < \infty$$

when $\lambda = \infty$.

The following lemma allows the construction of asymptotically efficient rules.

Lemma 7.1: Let $Y_1, Y_2, \cdots$ be samples coming under parameter ϑ. For any $K > 0$ and $0 < \alpha < \frac{1}{4}$, with $\gamma(t) = K t^{-\alpha}$ we have

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

for all $0 < \delta < 1$, $q > 1$ and $\vartheta \in \Theta$, where $h_a(Y_1, \ldots, Y_a) = \frac{Y_1 + \cdots + Y_a}{a}$.

Proof : Fix $x \in X$. Let $\tau_0 = \inf\{t \geq 1 \mid Y_t = x\}$ and define $\tau_1, \tau_2, \cdots$ and T_n by

$$\tau_n = \inf\{t \geq 1 \mid Y_{T_{n-1}+t} = x\},$$

$$T_n = \tau_0 + \tau_1 + \cdots + \tau_n .$$

The random variables τ_n, $n \geq 1$, are i.i.d. Further, τ_0 and $\{\tau_n, n \geq 1\}$ have geometrically bounded tails, see e.g. [4], Chapter 1, Prop. (79), and hence have moments of all orders. Moreover, $\mathbf{E}\tau_1 = \frac{1}{\pi(x,\vartheta)}$. Note that T_n is the time of the $(n+1)$st visit to x.

Let $S_n = T_n - \frac{n}{\pi(x,\vartheta)} - E\tau_0$, so that $\{S_n,\ n \geq 1\}$ is a martingale. A simple calculation gives

$$ES_t^4 \leq E(\tau_0 - E\tau_0)^4 + 6tE(\tau_0 - E\tau_0)^2 E(\tau_1 - \frac{1}{\pi(x,\vartheta)})^2 + 3t^2 E(\tau_1 - \frac{1}{\pi(x,\vartheta)})^4 .$$

The maximal inequality applied to the positive submartingale $\{S_t^4\}$ gives, for any $K > 0$,

$$P_\vartheta\{\max_{1 \leq a \leq t} |S_a| \geq K t^{1-\alpha}\} = O(t^{4\alpha-2}) \tag{7.2}$$

which is $O(t^{-1}(\log t)^{-q})$ for any $q > 1$ if $0 < \alpha < \frac{1}{4}$. We have

$$\{\max_{\delta t \leq a \leq t} |h_a(Y_1, \ldots, Y_a) - \mu(\vartheta)| > K t^{-\alpha}\}$$

$$\subset \bigcup_{x \in X} \{\max_{\delta t \leq a \leq t} |N(x,a) - a\pi(x,\vartheta)| > \frac{\delta K t^{1-\alpha}}{|X|}\} . \tag{7.3}$$

Further,

$$\{N(x,a) > a\pi(x,\vartheta) + \frac{\delta K t^{1-\alpha}}{|X|}\} \subset \{T_{\lfloor a\pi(x,\vartheta) + \frac{\delta K t^{1-\alpha}}{|X|} - 1 \rfloor} \leq a\}$$

$$\subset \{\max_{1 \leq a \leq t} |S_b| \geq \frac{\delta K t^{1-\alpha}}{2|X|}\},$$

and

$$\{N(x,a) < a\pi(x,\vartheta) - \frac{\delta K t^{1-\alpha}}{|X|}\} \subset \{T_{\lfloor a\pi(x,\vartheta) - \frac{\delta K t^{1-\alpha}}{|X|} - 1 \rfloor} > a\}$$

$$\subset \{\max_{1 \leq a \leq t} |S_b| \geq \frac{\delta K t^{1-\alpha}}{2|X|}\},$$

for t sufficienty large. The result follows from (7.2) and (7.3). ∎

Theorem 7.1 : The allocation rule of §7 of [1], with an initial sample of size $2N$ to initiate the g_{ta} statistics, is asymptotically efficient.

Proof : Reindex the arms so that $\mu(\vartheta_1) \geq \cdots \geq \mu(\vartheta_N)$. Using (7.1) and the properties (4.6), (4.7) and (4.8) of the g_{ta} statistic, we can argue exactly as in the proof of Theorem 7.2 of [1] to get

$$\limsup_{t \to \infty} \frac{t \sum_{i=1}^{m} \mu(\vartheta_i) - \sum_{j=1}^{N} \mu(\vartheta_j) E T_t(j)}{\log t} \leq \sum_{j \text{ is } m\text{-worst}} \frac{\mu(\vartheta_m) - \mu(\vartheta_j)}{I(\vartheta_j, \lambda)}$$

if $\lambda < \infty$, and

$$\limsup_{t \to \infty} \frac{t \sum_{i=1}^{m} \mu(\vartheta_i) - \sum_{j=1}^{N} \mu(\vartheta_j) E T_t(j)}{\log t} < \infty$$

if $\lambda = \infty$. The proof is concluded using (2.10).

References

[1] V. Anantharam, P. Varaiya and J. Walrand, *Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays : Part I : I.I.D. rewards.*

[2] R.S. Ellis, *Entropy, Large Deviations and Statistical Mechanics*, Grund. der Math. Wiss., Vol 271, Springer Verlag, 1985.

[3] R.S. Ellis, "Large deviations for a general class of random vectors", *Annals of Probability*, 12, 1984, pp 1 -12.

[4] D. Freedman, *Markov Chains*, Springer Verlag, 1983.

[5] J. Neveu, *Discrete Parameter Martingales*, North Holland, 1975.