Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem with Multiple Plays—Part II: Markovian 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 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 $C = (\theta_1, \cdots, \theta_N)$ 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 explicitly in terms of the Kullback-Liebler number between pairs of transition probabilities.**

I. 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 $P(\theta) = [P(x, y, \theta)]$, $\theta \in \mathbf{R}$, $x, y \in X$, where $X \subset \mathbf{R}$ is a finite set of rewards. There are N arms X_j, $j = 1, \cdots, N$ with parameter configuration $C = (\theta_1, \cdots, \theta_N)$. Successive plays of arm j result in X-valued random variables $Y_{j1}, Y_{j2}, \cdots$ whose statistics are given by $P(\theta)$. The first play of an arm with parameter θ has reward distribution $p(\theta)$ 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.

We assume that

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

$P(\theta)$ is irreducible and aperiodic for all $\theta \in \mathbf{R}$, and

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

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

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

is assumed to be strictly monotone increasing in θ.

Manuscript received September 8, 1986; revised June 1, 1987. Paper recommended by 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 8716711.

The values that can actually arise as parameters of the arms belong to a subset $\Theta \subset \mathbf{R}$. In Sections II–V Θ is assumed to satisfy the denseness condition (2.12). This restriction is removed in Sections VI and VII.

II. 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 $F_\infty = \vee_t F_t$. 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

$$|EN(x, \tau) - \pi(x)E\tau| \leq K, \tag{2.1}$$

and

$$|EN(x, y, \tau) - \pi(x)P(x, y)E\tau| \leq K \tag{2.2}$$

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

Proof: Let $X^* = \cup_{t \geq 1} X^t$, with the Borel σ-algebra of the discrete topology, i.e., all subsets are measurable. The process $\{Y_t, t \geq 1\}$ allows us 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 $\tau_k < \infty$ 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 $x, y \in X$, $y = (y_1, y_2, \cdots, y_t) \in X^*$, let $l(y)$ = length of y, $N(x, y)$ = number of times x appears in y, and $N(x, y, y)$ = number of transitions from x to y in y where $y_t \to y_1$ is also considered a transition.

It is well-known (see, e.g., [4, ch. 1, Theorem (31)]) that $\{B_k\}$

is i.i.d. and for any $x, y \in X$

$$EN(x, B_1) = \pi(x)El(B_1),$$

$$EN(x, y, B_1) = \pi(x)P(x, y)El(B_1).$$

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

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

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

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

Observe that for a fixed constant K independent of p and τ, $E(T - \tau) \le K$. 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),

$$\pi(x)E(T-1) - K \le EN(x, \tau) < \pi(x)E(T-1) + 1,$$

so that

$$\pi(x)E\tau - K \le EN(x, \tau) \le \pi(x)E\tau + K. \quad (2.7)$$

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),

$$\pi(x)P(x, y)E(T-1) - K \le EN(x, y, \tau) \le \pi(x)P(x, y)E(T-1),$$

so that

$$\pi(x)P(x, y)E\tau - K \le EN(x, y, \tau) \le \pi(x)P(x, y)E\tau + K. \quad (2.8)$$

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

Let $Y_{j1}, Y_{j2}, \cdots$ denote the successive rewards from arm j. Let $F_t(j)$ denote the σ-algebra generated by $Y_{j1}, \cdots, Y_{jt}$, $F_\infty(j) = \vee_{t \ge 1} F_t(j)$, and $G(j) = \vee_{i \ne j} F_\infty(i)$. As in [1, sect. II], 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 Φ the number of plays we have made of arm j at time t, $T_t(j)$, is a stopping time of $\{F_s(j) \vee G(j), s \ge 1\}$. The total reward is

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

By Lemma 2.1,

$$\left|ES_t - \sum_{j=1}^{N} \mu(\theta_j)ET_t(j)\right| \le \text{const.} \quad (2.9)$$

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 $C = (\theta_1, \cdots, \theta_N)$ 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

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

Then the regret is

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

By (2.9),

$$\left|R_t(\theta_1, \cdots, \theta_N) - \left[t\sum_{i=1}^{m} \mu(\theta_{\sigma(i)}) - \sum_{j=1}^{N} \mu(\theta_{\sigma(j)})ET_t(j)\right]\right| \le \text{const.} \quad (2.10)$$

where the constant can depend on C.

An allocation rule is called *uniformly good* if for every configuration $R_t(\theta_1, \cdots, \theta_N) = o(t^\alpha)$ 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 $I(\theta, \lambda)$ denote $I(P(\theta), P(\lambda))$. Under (1.1) and (1.2), $0 < I(\theta, \lambda) < \infty$ for $\theta \ne \lambda$. We assume that

$$I(\theta, \lambda) \text{ is continuous in } \lambda > \theta \qquad \text{for fixed } \theta. \quad (2.11)$$

In Sections II–V we also assume the following denseness condition on Θ:

for all $\lambda \in \Theta$ and $\delta > 0$, there is $\lambda' \in \Theta$ s.t.

$$\mu(\lambda) < \mu(\lambda') < \mu(\lambda) + \delta. \quad (2.12)$$

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

For a parameter configuration $C = (\theta_1, \cdots, \theta_N)$, define the notions of m-best, m-worst, and m-border arms exactly as in [1, sect. III]. By (2.10), an adaptive allocation rule Φ 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 $\alpha > 0$.

Thereom 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 $C = (\theta_1, \cdots, \theta_N)$, then for each distinctly m-worst arm j and each $\epsilon > 0$,

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

so that

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

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

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

Consequently,

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

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

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

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

Consider the new configuration of parameters $C^* = (\theta_1, \cdots, \theta_{j-1}, \lambda, \theta_{j+1}, \cdots, \theta_N)$. Let $Y_1, Y_2, \cdots$ denote the sequence of rewards from plays of arm j under the uniformly good rule Φ. Define

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

By (1.1) and the ergodic theorem $L_t/t \to I(\theta_j, \lambda)$ 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

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

After this point the proof proceeds exactly as in [1, Theorem 3.1]. □

IV. Construction of Statistics

An allocation rule is *asymptotically efficient* if for each $C = (\theta_1, \cdots, \theta_N)$

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

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

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

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 $f: X \to \mathbf{R}$ be such that $\Sigma_{x\in X} \pi(x)f(x) > 0$ and let $S_t = \Sigma_{a=1}^{t} f(Y_a)$. Let $L = \Sigma_{t=1}^{\infty} 1(\inf_{a\geq t} S_a \leq 0)$. Then $EL < \infty$.

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

The process $\{Y_t\}$ defines for each $t \geq 1$ a probability measure Q_t on 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 M. From the large deviations theory for this weak convergence, [3, Theorem II.1], there are constants $A > 0$, $\alpha > 0$ such that

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

Now

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

so that

$$Q_t(K) = E1(S_t \leq 0)$$

and the result follows. □

Lemma 4.2: Let $\{Y_t, t \geq 1\}$, P, π, p be as in Lemma 4.1 and $f: X^2 \to \mathbf{R}$ be such that $\Sigma_{x,y\in X} \pi(x)P(x, y)f(x, y) > 0$. For $t \geq 2$ let $S_t = \Sigma_{a=1}^{t-1} f(Y_a, Y_{a+1})$. Let $N = \Sigma_{t=2}^{\infty} 1(S_t \leq 0)$. Then $EN < \infty$.

Proof: We appeal to the large deviations theory for the empirical transition count matrix of a finite state Markov chain (see [2]). Let $M^{(2)}$ be the unit simplex in $\mathbf{R}^{|X|^2}$ identified with the space of probability measures on X^2, and define $F: M^{(2)} \to \mathbf{R}$ by $F(v) = \Sigma_{x,y\in X} f(x, y)v(x, y)$. Let $K = \{v \in M^{(2)} | F(v) \leq 0\}$. Let $\pi P \in M^{(2)}$ 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 $t \geq 2$ a probability measure $Q_t^{(2)}$ on $M^{(2)}$ which is the distribution of the $M^{(2)}$ valued random variable whose component in the (x, y) direction is $N(x, y, t)/t - 1$. Then $Q_t^{(2)} \to \delta_{\pi P}$ weakly as probability measures on $M^{(2)}$. From the large deviations theory, [2, Problem IX.6.12], there are constants $A > 0$, $\alpha > 0$ such that

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

Now

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

so that

$$Q_t^{(2)}(K) = E1(S_t \leq 0)$$

from which the result follows. □

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

$$\limsup_{A\to\infty} \frac{EN_A}{A} \leq \frac{1}{\mu}.$$

Proof: For any $\epsilon > 0$

$$N_A \leq \frac{A(1+\epsilon)}{\mu} + 1 + \sum_{t=2}^{\infty} 1\left[S_t \leq (t-1)\frac{\mu}{1+\epsilon}\right].$$

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

$$EN_A \leq \frac{A(1+\epsilon)}{\mu} + \text{const.}$$

for some constant depending on ϵ. Thus,

$$\limsup_{A\to\infty} \frac{EN_A}{A} \leq \frac{1+\epsilon}{\mu}.$$

Letting $\epsilon \to 0$ 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

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

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

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

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

1) $P_\theta\{\eta < U(a, Y_1, \cdots, Y_a, K)$ for all $a \geq 2\} \geq 1 - 1/K$,

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

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

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

$$U(a, Y_1, \cdots, Y_a, K)<\theta \Leftrightarrow W_a(\theta)\geq K.$$

Now

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

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

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

establishing (1).

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

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

where

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

By Jensen's inequality

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

Now

$$\begin{aligned}&\sum_{x,y\in X} \pi(x, \theta)P(x, y, \theta) \int_{\substack{|\eta-\theta|<\delta\\ \eta>\theta}} \log \frac{P(x, y, \eta)}{P(x, y, \lambda)} h^o(\eta)\, d\eta\\ &= \sum_{x,y\in X} \pi(x, \theta)P(x, y, \theta) \left[\log \frac{P(x, y, \theta)}{P(x, y, \lambda)} - \int_{\substack{|\eta-\theta|<\delta\\ \eta>\theta}} \log \frac{P(x, y, \theta)}{P(x, y, \eta)} h^o(\eta)\, d\eta\right]\\ &=I(\theta, \lambda)- \int_{\substack{|\eta-\theta|<\delta\\ \eta>\theta}} I(\theta, \eta)h^o(\eta)\, d\eta\\ &\geq I(\theta, \lambda)-\epsilon>0\end{aligned}$$

for ϵ sufficiently small. By Lemma 4.3, $EN_K < \infty$ and

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

Letting $\epsilon \to 0$ gives

$$\limsup_{K\to\infty} \frac{E_\theta N_K}{\log K}\leq \frac{1}{I(\theta, \lambda)}. \tag{4.3}$$

To bound $E_\theta N_K$ 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 $E_\theta T_K < \infty$. Since

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

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

$$\begin{aligned}\log K\leq E_\theta \log W_{T_K}(\lambda)&=\log E_\theta L_{T_K}+E_\theta \log M_{T_K}\\ &\leq E_\theta \log L_{T_K}+\log E_\theta M_{T_K}\\ &=E_\theta \log L_{T_K}.\end{aligned} \tag{4.4}$$

Now

$$E_\theta \log L_{T_K}= \sum_{x,y\in X} E_\theta N(x, y, T_K) \log \frac{P(x, y, \theta)}{P(x, y, \lambda)}$$

and by Lemma 2.1

$$|E_\theta N(x, y, T_K)-\pi(x, \theta)P(x, y, \theta)E_\theta T_K|\leq \text{const.}$$

Hence

$$|E_\theta \log L_{T_K}-I(\theta, \lambda)E_\theta T_K|\leq \text{const.} \tag{4.5}$$

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

$$\liminf_{K\to\infty} \frac{E_\theta N_K}{\log K}\geq \frac{1}{I(\theta, \lambda)}$$

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

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

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

$$2)\ \limsup_{t\to\infty} \sum_{a=2}^{t} \frac{P_\theta\{g_{ta}(Y_1, \cdots, Y_a)\geq \mu(\lambda)\}}{\log t}\leq \frac{1}{I(\theta, \lambda)}, \tag{4.7}$$

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 $\mu(\theta)$ is monotonically increasing in θ. □

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

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

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

$$P_\theta\{\max_{\delta t \le a \le t} |h_a(Y_a, \cdots, Y_a) - \mu(\theta)| > \epsilon\} = o(t^{-1}) \tag{4.9}$$

for every θ.

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

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

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

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

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

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 Rule

Consider the allocation rule of [1, sect. V] using the g_{ta} and h_a statistics constructed in Section IV 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 $\mu(\theta_1) \ge \cdots \ge \mu(\theta_N)$. Let $0 \le l \le m - 1$ and $m \le n \le N$ 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 $E(t - T_t(j)) = o(\log t)$ for every $j \le l$.

B: If $n < N$, let

$B_t = \#\{N \le a \le t \mid$. There exists $j \ge n+1$ s.t. j is one of the m-leaders at stage $a+1\}$.

Then $EB_t = o(\log t)$.

C: If $n < N$ and $0 < \epsilon < \mu(\theta_n) - \mu(\theta_{n+1})$, then for $j \ge n + 1$ let

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

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

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

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

$$\limsup_{t\to\infty} \frac{t \sum_{i=1}^{m} \mu(\theta_i) - \sum_{j=1}^{N} \mu(\theta_j) ET_t(j)}{\log t} \le \sum_{j \text{ is } m\text{-worst}} \frac{\mu(\theta_m) - \mu(\theta_j)}{I(\theta_j, \theta_m)}$$

from which the proof follows using (2.10). □

VI. Isolated Parameter Values: Lower Bound

We proceed to examine the situation in the absence of the denseness condition (2.12). For a configuration $C = (\theta_1, \cdots, \theta_N)$, let σ be a permutation of $\{1, \cdots, N\}$ such that $\mu(\theta_{\sigma(1)}) \ge \cdots \ge \mu(\theta_{\sigma(N)})$. Throughout this section and Section VII, $\lambda \in \Theta$ (λ depending on C) is defined as

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

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

Theorem 6.1: Let the family of reward distributions satisfy (2.11). Let Φ be a uniformly good rule. Let $C = (\theta_1, \cdots, \theta_N)$ be a configuration and σ, λ as above. If $\lambda < \infty$, then, for each distinctly m-worst arm j,

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

Consequently, by (2.10),

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

for each C.

Proof: Let j be an m-worst arm. Consider the parameter configuration $C^* = (\theta_1, \cdots, \theta_{j-1}, \lambda, \theta_{j+1}, \cdots, \theta_N)$ when the arm j has parameter λ instead of θ_j and proceed as in Theorem 3.1. □

VII. Isolated Parameter Values: An Asymptotically Efficient Rule

As in [1, sect. VII], an allocation rule is called *asymptotically efficient* if

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

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

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

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 $0 < \alpha < 1/4$, with $\gamma(t) = Kt^{-\alpha}$ we have

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

for all $0 < \delta < 1$, $q > 1$ and $\theta \in \Theta$, where $h_a(Y_1, \cdots, Y_a) = (Y_1 + \cdots + Y_a)/a$.

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

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

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

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

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

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

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

$$P_\theta\{\max_{1 \le a \le t} |S_a| \ge Kt^{1-\alpha}\} = O(t^{4\alpha - 2}) \tag{7.2}$$

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

$$\{\max_{\delta t \le a \le t} |h_a(Y_1, \cdots, Y_a) - \mu(\theta)| > Kt^{-\alpha}\} \subset \bigcup_{x \in X} \left\{\max_{\delta t \le a \le t} |N(x, a) - a\pi(x, \theta)| > \frac{\delta K t^{1-\alpha}}{|X|}\right\}. \tag{7.3}$$

Further

$$\left\{N(x, a) > a\pi(x, \theta) + \frac{\delta K t^{1-\alpha}}{|X|}\right\} \subset \{T_{\lceil a\pi(x,\theta) + (\delta K t^{1-\alpha})/|X| - 1\rceil} \le a\} \subset \left\{\max_{1 \le a \le t} |S_b| \ge \frac{\delta K t^{1-\alpha}}{2|X|}\right\},$$

and

$$\left\{N(x, a) > a\pi(x, \theta) - \frac{\delta K t^{1-\alpha}}{|X|}\right\} \subset \{T_{\lfloor a\pi(x,\theta) - (\delta K t^{1-\alpha})/|X| - 1\rfloor} > a\} \subset \left\{\max_{1 \le a \le t} |S_b| \ge \frac{\delta K t^{1-\alpha}}{2|X|}\right\}$$

for t sufficiently large. The result follows from (7.2) and (7.3). □

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

Proof: Reindex the arms so that $\mu(\theta_1) \ge \cdots \ge \mu(\theta_N)$. Using (7.1) and the properties (4.6)–(4.8) of the g_{ta} statistic, we can argue exactly as in the proof of Theorem 7.2 of [1] to get

$$\limsup_{t \to \infty} \frac{t \sum_{i=1}^m \mu(\theta_i) - \sum_{j=1}^N \mu(\theta_j) ET_t(j)}{\log t} \le \sum_{j \text{ is } m\text{-worst}} \frac{\mu(\theta_m) - \mu(\theta_j)}{I(\theta_j, \lambda)}$$

if $\lambda < \infty$, and

$$\limsup_{t \to \infty} \frac{t \sum_{i=1}^m \mu(\theta_i) - \sum_{j=1}^N \mu(\theta_j) ET_t(j)}{\log t} < \infty$$

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," *IEEE Trans. Automat. Contr.*, this issue, pp. 968-976.

[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," *Ann. Probability*, vol. 12, pp. 1-12, 1984.

[4] D. Freedman, *Markov Chains.* New York: Springer-Verlag, 1983.

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

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

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

[8] 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. New York: Marcel Dekker, pp. 127-142.

Venkatachalam Anantharam (M'86), for a photograph and biography, see this issue, p. 976.

Pravin Varaiya (M'68-SM'78-F'80), for a photograph and biography, see this issue, p. 976.

Jean Walrand (S'71-M'74-M'80), for a photograph and biography, see this issue, p. 976.