APPLICATIONES MATHEMATICAE
23,4 (1996), pp. 449–473

E. DRABIK (Białystok)

ON NEARLY SELFOPTIMIZING STRATEGIES FOR MULTIARMED BANDIT PROBLEMS WITH CONTROLLED ARMS

Abstract. Two kinds of strategies for a multiarmed Markov bandit problem with controlled arms are considered: a strategy with forcing and a strategy with randomization. The choice of arm and control function in both cases is based on the current value of the average cost per unit time functional. Some simulation results are also presented.

1. Introduction. This paper presents allocation rules for the multiarmed bandit problem with $N > 1$ arms, the dynamics of which is characterized by controlled Markov chains $X^j = (X_i^j),\ i = 1, 2, \ldots;\ j = 1, \ldots, N$ (on a state space E), whose transition probability operators are parametrized by an unknown parameter $\theta^{0,j} \in \Theta$, where Θ is a given compact set.

We assume that at each time t always one of the N arms is played. The arm that we play is also controlled. In general a control strategy is a sequence $(v_0, v_1, \ldots)$ of U-valued (U is a given compact set of control parameters) random variables that are adapted to the σ-field generated by the observations of the arms.

When at time t the jth ($j = 1, \ldots, N$) arm is played and the control v_t is used the cost $c(x_t^j, v_t)$ is incurred, with x_t^j denoting the position of the jth arm at time t. The problem is to find a strategy that minimizes the average cost per unit time. In what follows we shall restrict the class of admissible controls to the so-called Markov controls, i.e. controls of the form $v_t = u(x_t^j)$, where $u : E \to U$ is a measurable function (we write $u \in B(E, U)$), assuming that at time t the jth arm is played. By the general theory of controlled Markov processes with average cost per unit time (see [8]) it is known that

---

1991 *Mathematics Subject Classification*: 93E20, 60J20.
Key words and phrases: stochastic control, multiarmed bandit, invariant measure, adaptative control, selfoptimizing strategies.

optimal controls are usually Markov, in particular, when we assume an ergodic condition (1.1) that we formulate below. Given a control $v_t = u(x_t^j)$ at time t, the transition operator that describes the evolution of the jth arm until time $t+1$ is of the form $P_{v_t}^{\theta^{0,j}}(x_t^j, A)$, where $\theta^{0,j}$ is the unknown value of the parameter corresponding to the arm j.

To indicate the dependence of $P_{v_t}^{\theta^{0,j}}(x_t^j, A)$, on the Markov control function u we shall simply write $P_u^{\theta^{0,j}}(x_t^j, A)$.

We assume that for $j = 1, \ldots, N$ and $u \in B(E,U)$ the operator $P_u^\theta(x, A)$ is uniformly ergodic, that is, there exists $0 < \gamma < 1$ and a unique invariant measure π_u^θ satisfying

$$\sup_{\theta\in\Theta} \sup_{u\in B(E,U)} \sup_{x\in E} \sup_{A\in B(E)} |(P_u^\theta)^n(x, A) - \pi_u^\theta(A)| \le \gamma^n. \tag{1.1}$$

Our purpose is to minimize

$$J := \limsup_{t\to\infty} t^{-1} \sum_{j=1}^{N} \sum_{i=0}^{t-1} c(x_i^j, v_i) S_j(i), \tag{1.2}$$

where $c : E \times U \to \mathbb{R}^+$ is a bounded measurable function and

$$S_j(i) = \begin{cases} 1 & \text{when the } j\text{th arm is played at time } i, \\ 0 & \text{otherwise.} \end{cases}$$

At each time t we choose one of the N arms to be played and then the control is applied to this arm. Since the transition operators of the arms depend on the unknown parameter θ^0 we cannot determine immediately the arm and control that guarantee the minimal value of the cost functional (1.2). Although the dynamics of the arms depends on the unknown parameters $\theta^{0,j}$, $j = 1, \ldots, N$, in this paper we do not estimate them directly. Instead we compare the average per unit time costs for different arms and controls. To make this approach feasible, we have to adopt from [9] the assumption that for $\varepsilon > 0$ there exists a finite set $\vartheta = \{u_1, \ldots, u_{r(\varepsilon)}\}$ of ε-optimal control functions, i.e. a family ϑ such that for all $\theta \in \Theta$ there exists $u \in \vartheta$ satisfying

$$J^\theta(u) := \limsup_{t\to\infty} t^{-1} \sum_{i=0}^{t-1} E^\theta c(x_i, u(x_i)) \le \lambda(\theta) + \varepsilon \tag{1.3}$$

with

$$\lambda(\theta) = \inf_{u\in B(E,U)} J^\theta(u).$$

Notice that by (1.1), we clearly have

$$J^\theta(u) = \int_E c(x, u(x))\, \pi_u^\theta(dx).$$

Sufficient conditions under which there exists a finite set of ε-optimal controls can be found in [9].

The multiarmed bandit processes with controlled arms are called sometimes superprocesses and were studied so far with discounted cost criterion only (see [5], [7] and the references therein). In this paper the superprocesses are considered with long run average cost (1.2). The approach based on the existence of ε-optimal functions introduced above seems to be new. The multiarmed bandit problems with noncontrolled arms and long run average cost were thoroughly investigated in the series of papers [1]–[4].

The present paper consists of 5 sections. In Section 2 a nearly optimal strategy with constant decision horizon is considered. The next Section 3 is devoted to the construction of an optimal strategy with increasing decision horizon. In Section 4 a nearly optimal strategy with randomization is studied. Finally, in Section 5 some simulation results are presented.

For the construction of our strategy, it is important to find, for a given $\varepsilon > 0$, a decision time horizon $\kappa > 0$ which satisfies the inequality

$$\sup_{\theta\in\Theta}\sup_{u\in B(E,U)}\sup_{x\in E}\Big|\kappa^{-1}E_x^\theta\Big\{\sum_{i=0}^{\kappa-1}c(x_i^{\theta^j},u(x_i^{\theta^j}))\Big\} - \int_E c(x,u(x))\,\pi_u^\theta(dx)\Big| \le \varepsilon. \tag{1.4}$$

We have

LEMMA 1.1. *Assume that* (1.1) *holds. Then the inequality* (1.4) *is satisfied for*

$$\kappa > \frac{2\|c\|}{1-\gamma}\cdot\frac{1}{\varepsilon}. \tag{1.5}$$

Proof. From (1.1) we have

$$\sup_{\theta\in\Theta}\sup_{u\in B(E,U)}\sup_{x\in E}\Big|E_x^\theta\{c(x_i,u(x_i))\} - \int_E c(x,u(x))\,\pi_u^\theta(dx)\Big| \le 2\|c\|\gamma^i.$$

Then

$$\begin{aligned}
&\sup_{\theta\in\Theta}\sup_{u\in B(E,U)}\sup_{x\in E}\Big|\kappa^{-1}\Big\{\sum_{i=0}^{\kappa-1}E_x^\theta c(x_i^{\theta^j},u(x_i^{\theta^j}))\Big\} - \int_E c(x,u(x))\,\pi_u^\theta(dx)\Big| \\
&\quad\le \sup_{\theta\in\Theta}\sup_{u\in B(E,U)}\sup_{x\in E}\Big\{\kappa^{-1}\sum_{i=0}^{\kappa-1}\Big|E_x^\theta c(x_i^{\theta^j},u(x_i^{\theta^j})) - \int_E c(x,u(x))\,\pi_u^\theta(dx)\Big|\Big\} \\
&\quad\le \sup_{\theta\in\Theta}\sup_{u\in B(E,U)}\sup_{x\in E}\Big\{\kappa^{-1}\sum_{i=0}^{\kappa-1}2\|c\|\gamma^i\Big\} \le \frac{2\|c\|}{\kappa}\sum_{i=0}^{\kappa-1}\gamma^i \le \frac{2\|c\|}{\kappa}\cdot\frac{1}{1-\gamma}.
\end{aligned}$$

Therefore for κ satisfying (1.5) the inequality (1.4) holds. ■

In order to illustrate the problem we consider the following

EXAMPLE 1. Assume (x_i^j) satisfies the equation

$$x_{i+1}^j = f(x_i^j, v_i, \theta^j) + g(x_i^j)w_i, \quad x_0^j = x,$$

where f is a bounded continuous vector function, g is a square matrix which has a bounded inverse and w_i is a sequence of i.i.d. Gaussian vectors with expected value 0 and covariance matrix I. Then

$$\begin{aligned} P_u^{\theta^j}(x_i^j, A) &:= P\{f(x_i^j, u(x_i^j), \theta^j) + g(x_i^j)w_i \in A\} \\ &= N(f(x_i^j, u(x_i^j), \theta^j), g(x_i^j)g^*(x_i^j)). \end{aligned}$$

In particular, in the one-dimensional case the transition probability function has the form

$$P_u^{\theta^j}(x, A) := \frac{1}{\sqrt{2\pi g^2(x_i^j)}} \int_A e^{-(y - f(x_i^j, u(x_i^j), \theta^j))^2/(2g^2(x_i^j))} dy.$$

It can be shown (see [9]) that the transition operators $P_u^{\theta^j}$ defined above satisfy (1.1), and γ can be calculated explicitly. Moreover, for every $\varepsilon > 0$ there exists a finite set of ε-optimal control functions (Lemma 2 of [9]).

Acknowledgments. The author would like to thank Prof. Ł. Stettner for helpful comments and encouragement. The paper is a part of the author's Ph.D. thesis written under the supervision of Prof. Ł. Stettner at the Technical University of Warsaw.

2. Construction of an ε-optimal strategy with forcing and constant time decision horizon. In this section we shall consider a strategy under which at certain times, called forcing times, successively each arm is played and each control of the class $\vartheta(\varepsilon)$ with fixed $\varepsilon > 0$ is applied.

Denote by F the set of all forcing times to be defined. It is characterized by a sequence a_i, $i = 0, 1, \ldots$, such that $a_{i+1} + a_i \geq Nr(\varepsilon)\kappa$, with $a_0 = 0$.

At time a_i we choose the first arm and apply the control function u_1 for κ consecutive moments of time. Then, at time $a_i + \kappa$ we play again the first arm but apply the control function u_2 for the next κ moments of time. We continue to play the first arm applying successively the controls $(u_3, \ldots, u_{r(\varepsilon)})$ for consecutive κ moments of time. At time $a_i + r(\varepsilon)\kappa$, we start to play a second arm and test successively for κ moments of time each of the control functions of the class $\vartheta(\varepsilon)$. Then we test in a similar way all the remaining arms. At time $a_i + Nr(\varepsilon)\kappa - 1$ we finish the forcing.

Therefore

$$F = \{0, 1, \ldots, Nr(\varepsilon)\kappa - 1, a_1, a_1 + 1, \ldots, a_1 + Nr(\varepsilon)\kappa - 1, \ldots$$
$$\ldots, a_i, a_i + 1, \ldots, a_i + Nr(\varepsilon)\kappa - 1, \ldots \ (i = 1, 2, \ldots)\}.$$

We choose a_i in such a way that for F we have

$$\limsup_{t\to\infty} t^{-1} \sum_{i=0}^{t-1} \chi_F(i) = 0.$$

Let

$F_j^k =$ the set of forcing moments when we play the jth arm and the control function u_k,

$F_j =$ the set of forcing moments when we play the jth arm.

It is clear that $F_j \cap F_i = \emptyset$ for $i \neq j$, $F = \bigcup_{j=1}^{N} F_j$ and $F_j = \bigcup_{k=1}^{r(\varepsilon)} F_j^k$.

Let $\Delta = r(\varepsilon)\kappa$. We construct our nearly optimal strategy in the following way.

A. Strategy in the forcing intervals. For the jth arm, we use the control function u_{i+1} in the time interval $[(j-1)\Delta + i\kappa, (j-1)\Delta + (i+1)\kappa - 1]$ $(j = 1, \ldots, N,\ i = 1, \ldots, r(\varepsilon))$.

The forcing is finished at time $N\Delta - 1$. At time a_1 we start again the forcing and in the intervals $[a_1, a_1 + \kappa - 1], \ldots, [a_1 + i\kappa, a_1 + (i+1)\kappa - 1]$ we play the first arm and use the control functions $u_1, \ldots, u_{i+1}$, respectively.

At time $a_1 + \Delta$ we start to play the second arm and the procedure is continued until time $a_1 + N\Delta - 1$. We proceed in the same way for other times a_i.

B. Strategy outside of the forcing intervals. Let $T_j(t)$ be the number of times arm j was used up to stage t, and $T_j^k(t)$ be the number of times arm j and the control function u_k were used up to stage t. Clearly

$$t = T_1(t) + \ldots + T_N(t), \qquad T_j(t) = T_j^1(t) + \ldots + T_j^{r(\varepsilon)}(t).$$

Let

$$J_j^k(t) := (T_j^k(t))^{-1} \sum_{i=0}^{t-1} c(x_i^j, u_k(x_i^j)) S_j^k(i) \tag{2.1}$$

be the average cost at time t for the jth arm when the control function u_k is used; here

$$S_j^k(i) = \begin{cases} 1 & \text{if the } j\text{th arm is played and } u_k \text{ is applied,} \\ 0 & \text{otherwise.} \end{cases}$$

Let

$$J_j(t) := (T_j(t))^{-1} \sum_{i=0}^{t-1} \sum_{k=1}^{r(\varepsilon)} c(x_i^j, u_k(x_i^j)) S_j^k(i) \tag{2.2}$$

be the average cost for the jth arm.

Outside the forcing set F we use the following decision rule.

Let t be a multiple of κ.

B1. We find j, $j = 1, \ldots, N$, and k, $k = 1, \ldots, r(\varepsilon)$, such that

$$J_j^k(t) = \min_{i=1,\ldots,N} \min_{l=1,\ldots,r(\varepsilon)} J_i^l(t).$$

B2. If $J_j^k(t) = J_i^l(t)$ and $j \neq i$ or $k \neq l$ then we choose the jth arm and the control function u_k when $j < i$; if $j = i$ we choose the jth arm and the control function u_k provided $k < l$. For the next κ moments of time we play the jth arm and use the control function u_k.

The next decision is made at time $t + \kappa$. If $t + \kappa \in F$ we apply step A; if $t + \kappa \notin F$ we repeat step B of our strategy. ∎

Notice that under the above notation the average cost at time t is of the form

$$J(t) := t^{-1} \sum_{i=0}^{t-1} \sum_{j=1}^{N} \sum_{k=1}^{r(\varepsilon)} c(x_i^j, u_k(x_i^j)) S_j^k(i). \tag{2.3}$$

We define

$$J := \limsup_{t\to\infty} t^{-1} J(t).$$

In what follows we shall need the following sequence of lemmas.

LEMMA 2.1. *Let c_i, $i = 0, 1, \ldots$, be a bounded sequence of numbers. Assume that the nonnegative integers $\mathbb{N}$ are partitioned into N disjoint infinite subsets $\Phi(j)$, $j = 1, \ldots, N$. If, for a given $\varepsilon > 0$, there exist numbers g_j^t, $j = 1, \ldots, N$, $t = 0, 1, 2,$ such that*

$$\limsup_{t\to\infty} \Big| \Big(\sum_{i=0}^{t-1} \chi_{\Phi(j)}(i) \Big)^{-1} \sum_{i=0}^{t-1} c_i \chi_{\Phi(j)}(i) - g_j^t \Big| \leq \varepsilon \tag{2.4}$$

for every $j \in \{1, \ldots, N\}$ then

$$\limsup_{t\to\infty} \Big| t^{-1} \sum_{i=0}^{t-1} c_i - \sum_{j=1}^{N} g_j^t t^{-1} \sum_{i=0}^{t-1} \chi_{\Phi(j)}(i) \Big| \leq \varepsilon. \tag{2.5}$$

Proof. Clearly

$$t^{-1} \sum_{i=0}^{t-1} c_i = \sum_{j=1}^{N} \Big(\sum_{i=0}^{t-1} \chi_{\Phi(j)}(i) \Big)^{-1} \Big(\sum_{i=0}^{t-1} c_i \chi_{\Phi(j)}(i) \Big) t^{-1} \sum_{i=0}^{t-1} \chi_{\Phi(j)}(i). \tag{2.6}$$

By (2.4) for every $\varepsilon_0 > 0$ there exists t_0 such that for $t \geq t_0$ and $j = 1, \ldots, N$ we have

$$\Big|\Big(\sum_{i=0}^{t-1} \chi_{\Phi(j)}(i)\Big)\Big(\sum_{i=0}^{t-1} c_i \chi_{\Phi(j)}(i)\Big) - g_j^t\Big| \leq \varepsilon + \varepsilon_0. \tag{2.7}$$

Then for $t \geq t_0$, from (2.6) and (2.7) we obtain

$$\begin{aligned}
&\limsup_{t\to\infty}\Big|t^{-1}\sum_{i=0}^{t-1} c_i - \sum_{j=1}^{N} g_j^t t^{-1} \sum_{i=0}^{t-1} \chi_{\Phi(j)}(i)\Big| \\
&\quad \leq \limsup_{t\to\infty} \sum_{j=1}^{N} \Big|\Big(\sum_{i=0}^{t-1} \chi_{\Phi(j)}(i)\Big)^{-1} \sum_{i=0}^{t-1} c_i \chi_{\Phi(j)}(i) - g_j^t\Big| t^{-1} \sum_{i=0}^{t-1} \chi_{\Phi(j)}(i) \\
&\quad \leq (\varepsilon + \varepsilon_0) \limsup_{t\to\infty} t^{-1} \sum_{j=1}^{N} \sum_{i=0}^{t-1} \chi_{\Phi(j)}(i) \leq \varepsilon + \varepsilon_0.
\end{aligned}$$

Since ε_0 can be chosen arbitrarily small, we obtain (2.5). ■

R e m a r k 2.1. From (2.5), under (2.4) in particular we have

$$\Big|\limsup_{t\to\infty} t^{-1} \sum_{i=0}^{t-1} c_i - \limsup_{t\to\infty} \sum_{j=1}^{N} g_j^t t^{-1} \sum_{i=0}^{t-1} \chi_{\Phi(j)}(i)\Big| \leq \varepsilon. \tag{2.8}$$

LEMMA 2.2. *Let c_i, $i = 0, 1, \ldots,$ be a bounded sequence of numbers. Then*

$$\limsup_{t\to\infty} t^{-1} \sum_{i=0}^{t-1} c_i = \limsup_{t\to\infty} (t\kappa)^{-1} \sum_{i=0}^{t-1} \sum_{k=i\kappa}^{(i+1)\kappa-1} c_k.$$

P r o o f. The right hand side of the above equation satisfies

$$\limsup_{t\to\infty} (t\kappa)^{-1} \sum_{i=0}^{t-1} \sum_{k=i\kappa}^{(i+1)\kappa-1} c_k = \limsup_{t\to\infty} t^{-1} \sum_{i=0}^{t\kappa-1} c_i.$$

Hence

$$\limsup_{t\to\infty} t^{-1} \sum_{i=0}^{t-1} c_i \geq \limsup_{t\to\infty} (t\kappa)^{-1} \sum_{i=0}^{t-1} \sum_{k=i\kappa}^{(i+1)\kappa-1} c_k.$$

We can select $t_k \to \infty$ such that

$$\limsup_{t\to\infty} t^{-1} \sum_{i=0}^{t-1} c_i = \lim_{k\to\infty} t_k^{-1} \sum_{i=0}^{t_k-1} c_i.$$

Let n_k be such that $t_k \in [n_k\kappa, (n_k+1)\kappa[$. Then we have

$$\frac{1}{n_k\kappa}\sum_{i=0}^{n_k\kappa-1} c_i = \frac{1}{n_k\kappa}\Big(\sum_{i=0}^{t_k-1} c_i - \sum_{i=n_k\kappa}^{t_k-1} c_i\Big)$$
$$= \frac{t_k}{n_k\kappa}\cdot\frac{1}{t_k}\Big(\sum_{i=0}^{t_k-1} c_i - \sum_{i=n_k\kappa}^{t_k-1} c_i\Big) \to \lim_{k\to\infty}\sum_{i=0}^{t_k-1} c_i.$$

The above convergence follows from the facts that

(a) $t_k/(n_k\kappa) \to 1$ as $t_k \to \infty$ and

(b) the second term of the sum has at most k terms and it does not affect the whole sum for sufficiently large t, because c_i's are bounded. Therefore

$$\lim_{k\to\infty}\frac{1}{n_k\kappa}\sum_{i=0}^{n_k\kappa-1} c_i = \limsup_{t\to\infty}\sum_{i=0}^{t-1} c_i = \limsup_{t\to\infty}\sum_{i=0}^{t\kappa-1} c_i,$$

which completes the proof. ∎

LEMMA 2.3. *Let (x_i) be a controlled Markov chain with controls v_l. Then*

$$Z_t := \sum_{i=0}^{t\kappa-1} c(x_i,v_i) - \sum_{i=0}^{t-1} E\Big\{\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l,v_l)\;\Big|\;\mathfrak{F}_{i\kappa}\Big\}$$

is a martingale with respect to the σ-field $\mathfrak{F}_{i\kappa} = \sigma\{x_0,\ldots,x_{t\kappa}\}$ and $(1/t)Z_t \to 0$ as $t\to\infty$ P-a.e.

Proof. In order to prove that $(1/t)Z_t \to 0$ we use the law of large numbers for martingales ([6], Vol. II, VII, Th. 2). We show first that Z_t is a martingale and that the assumptions of the law of large numbers for martingales are satisfied. Let $Z_t = \sum_{i=0}^{t-1} X_i$ with

$$X_i = \sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l,v_l) - E\Big\{\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l,v_l)\;\Big|\;\mathfrak{F}_{i\kappa}\Big\}.$$

We have

$$E\{X_i \mid \mathfrak{F}_{i\kappa}\} = E\Big\{\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l,v_l) - E\Big\{\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l,v_l)\;\Big|\;\mathfrak{F}_{i\kappa}\Big\}\;\Big|\;\mathfrak{F}_{i\kappa}\Big\}$$
$$= E\Big\{\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l,v_l)\;\Big|\;\mathfrak{F}_{i\kappa}\Big\}$$
$$-E\Big\{\Big\{\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l,v_l)\;\Big|\;\mathfrak{F}_{i\kappa}\Big\}\;\Big|\;\mathfrak{F}_{i\kappa}\Big\} = 0.$$

Therefore Z_t is a martingale. Since

$$\begin{aligned}|X_i| &= \Big|\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l,v_l) - E\Big\{\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l,v_l) \,\Big|\, \mathfrak{F}_{i\kappa}\Big\}\Big| \\ &\le \Big|\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l,v_l)\Big| + \Big|E\Big\{\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l,v_l) \,\Big|\, \mathfrak{F}_{i\kappa}\Big\}\Big| \le \kappa\|c\| + \kappa\|c\| \\ &= 2\kappa\|c\|,\end{aligned}$$

we have $\sup_i |X_i| \le 2\kappa\|c\|$ and $\sum_{i=0}^{\infty} E^2\{X_i\}/i^2 < \infty$. Consequently, the assumptions of the law for large numbers of martingales are satisfied and $(1/t)Z_t \to 0$ as $t \to \infty$ P-a.e. ∎

From Lemma 2.3 we immediately have

Corollary 2.1. *For $k \in \{1,\ldots,r(\varepsilon)\}$ and $j \in \{1,\ldots,N\}$ we have*

$$\limsup_{t\to\infty} (T_j^k(t\kappa))^{-1}\Big\{\sum_{i=0}^{t\kappa-1} c(x_i^j, u_k(x_i^j))S_j^k(i) - \sum_{i=0}^{t-1} E\Big\{\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l^j, u_k(x_l^j))S_j^k(i\kappa) \,\Big|\, \mathfrak{F}_{i\kappa}\Big\}\Big\} = 0 \qquad P\text{-a.e.}$$ ∎

By the choice of the decision horizon κ (see (1.4)) we get

Proposition 2.1. *There exists $C \subset \Omega$ such that $P(C) = 0$ and for $\omega \in \Omega \setminus C$, $k \in \{1,\ldots,r(\varepsilon)\}$ and $j \in \{1,\ldots,N\}$ we have*

$$\limsup_{t\to\infty} |J_j^k(t)(\omega) - J^{\theta^{0,j}}(u_k)| \le \varepsilon. \tag{2.9}$$

Proof. To simplify notations set $J_j^k(t)(\omega) =: J_j^k(t)$ and $\pi_{u_k}^{\theta^j} =: \pi_k^j$. Notice first that by Lemma 2.2,

$$\limsup_{t\to\infty} J_j^k(t) = \limsup_{t\to\infty} J_j^k(t\kappa).$$

By Corollary 2.1 and the definition of κ (see (1.4)) for $\omega \in \Omega \setminus C$, where $P(C) = 0$, we have

$$\limsup_{t\to\infty} \Big| J_j^k(t\kappa) - \int_E c(x,u(x))\,\pi_k^j(dx)\Big|$$

$$\le \limsup_{t\to\infty} \Big| J_j^k(t\kappa) - (T_j^k(t\kappa))^{-1} \sum_{i=0}^{t-1} E\Big\{ \sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l^j, u_k(x_l^j)) S_j^k(l) \;\Big|\; \mathfrak{F}_{i\kappa} \Big\} \Big|$$

$$+ \limsup_{t\to\infty} \Big| (T_j^k(t\kappa))^{-1} \sum_{i=0}^{t-1} E\Big\{ \sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l^j, u_k(x_l^j)) S_j^k(l) \;\Big|\; \mathfrak{F}_{i\kappa} \Big\}$$

$$- (T_j^k(t\kappa))^{-1} \sum_{i=0}^{t\kappa-1} S_j^k(i) \int_E c(x, u(x))\, \pi_k^j(dx) \Big|$$

$$\le \limsup_{t\to\infty} \Big| (T_j^k(t\kappa))^{-1} \sum_{i=0}^{t-1} S_j^k(i\kappa) E\Big\{ \sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l^j, u_k(x_l^j)) S_j^k(l) \;\Big|\; \mathfrak{F}_{i\kappa} \Big\}$$

$$-\kappa \int_E c(x, u(x))\, \pi_k^j(dx) \Big| \le \varepsilon.$$

Since $J^{\theta^{0,j}}(u_k) = \int_E c(x,u(x))\,\pi_k^j(dx)$ we obtain (2.9) and the proof of Proposition 2.1 is complete. ■

Remark 2.2. It immediately follows from (2.9) that $\limsup_{t\to\infty} J_j^k \le J^{\theta^{0,j}}(u_k) + \varepsilon$ P-a.e. ■

Combining Lemma 2.1 and Proposition 2.1 we obtain

Corollary 2.2. *For $\omega \in \Omega \setminus C$, with C as in Proposition* 2.1, *and every $k \in \{1, \ldots, r(\varepsilon)\}$ and $j \in \{1, \ldots, N\}$ we have*

$$\limsup_{t\to\infty} \Big| J_j(t) - \sum_{k=1}^{r(\varepsilon)} J^{\theta^{0,j}}(u_k)(T_j(t))^{-1}(T_j^k(t)) \Big| \le \varepsilon \tag{2.10}$$

and consequently

$$\Big| \limsup_{t\to\infty} J_j(t) - \limsup_{t\to\infty} \sum_{k=1}^{r(\varepsilon)} J^{\theta^{0,j}}(u_k)(T_j(t))^{-1}(T_j^k(t)) \Big| \le \varepsilon. \tag{2.11}$$

Proof. Observe that by Proposition 2.1 the assumptions of Lemma 2.1 are satisfied, that is,

$$\limsup_{t\to\infty} \Big| (T_j^k(t))^{-1} \sum_{i=0}^{t-1} c(x_i, v_i) S_j^k(i) - J^{\theta^{0,j}}(u_k) \Big| \le \varepsilon.$$

Therefore from (2.5) we have

$$\limsup_{t\to\infty} \Big| (T_j(t))^{-1} \sum_{i=0}^{t-1} c(x_i, v_i) S_j^k(i) - \sum_{k=1}^{r(\varepsilon)} J^{\theta^{0,j}}(u_k)(T_j(t))^{-1} \sum_{i=0}^{t-1} S_j^k(i) \Big| \le \varepsilon.$$

Since $\sum_{i=0}^{t-1} S_j^k(i) = T_j^k(t)$ we obtain (2.10). The inequality (2.11) follows immediately from (2.10). ∎

Furthermore, we have

COROLLARY 2.3. *For $\omega \in \Omega \setminus C$, with C as in Proposition 2.1, and every $k \in \{1, \ldots, r(\varepsilon)\}$ and $j \in \{1, \ldots, N\}$ we have*

$$\limsup_{t\to\infty} \Big| J(t) - \sum_{j=1}^{N} \sum_{k=1}^{r(\varepsilon)} J^{\theta^{0,j}}(u_k) t^{-1} (T_j^k(t)) \Big| \le \varepsilon \tag{2.12}$$

and consequently

$$\Big| \limsup_{t\to\infty} J(t) - \limsup_{t\to\infty} \sum_{j=1}^{N} \sum_{k=1}^{r(\varepsilon)} J^{\theta^{0,j}}(u_k) t^{-1} (T_j^k(t)) \Big| \le \varepsilon. \tag{2.13}$$

Proof. By (2.10) and Lemma 2.1 we obtain

$$\limsup_{t\to\infty} \Big| J(t) - \sum_{j=1}^{N} \sum_{k=1}^{r(\varepsilon)} J^{\theta^{0,j}}(u_k) (T_j(t))^{-1} (T_j(t)) (T_j^k(t)) t^{-1} \Big| \le \varepsilon.$$

Hence we have (2.12) and, as a consequence, (2.13). ∎

We can now formulate the main result of this section.

THEOREM 2.1. *There exists $C \subset \Omega$ such that $P(C) = 0$ and for $\omega \in \Omega \setminus C$, $k \in \{1, \ldots, r(\varepsilon)\}$ and $j \in \{1, \ldots, N\}$ we have*

$$\begin{aligned} \limsup_{t\to\infty} J(t) &\le \min_{j=1,\ldots,N} \min_{k=1,\ldots,r(\varepsilon)} J^{\theta^{0,j}}(u_k) + 2\varepsilon \\ &\le \min_{j=1,\ldots,N} \lambda(\theta^{0,j}) + 3\varepsilon. \end{aligned} \tag{2.14}$$

Proof. By Corollary 2.3 we have to estimate

$$\limsup_{t\to\infty} t^{-1} \sum_{j=1}^{N} \sum_{k=1}^{r(\varepsilon)} J^{\theta^{0,j}}(u_k) (T_j^k(t)).$$

For this purpose we define

$$\begin{aligned} Z = \{ &(j,k) \in \{1, \ldots, N\} \times \{1, \ldots, r(\varepsilon)\} : \\ &|J^{\theta^{0,j}}(u_k) - \min_{l=1,\ldots,N} \min_{i=1,\ldots,r(\varepsilon)} J^{\theta^{0,l}}(u_i)| \le 2\varepsilon \}. \end{aligned} \tag{2.15}$$

We shall need the following lemma.

LEMMA 2.4. *If $(j,k) \notin Z$, then with probability 1 there is no sequence t_n, $t_n \to \infty$, $t_n \notin F$, such that at time t_n we select the jth arm and the control function u_k.*

Proof. Assume $(j,k) \notin Z$ and at time t_n, $t_n \to \infty$, $t_n \notin F$ being a multiple of κ, we select the jth arm, $j \in \{1,\dots,N\}$, and the control function u_k. Then $J_j^k(t_n) \le J_l^i(t_n)$ for all $l \in \{1,\dots,N\}$ and $i \in \{1,\dots,r(\varepsilon)\}$. Letting $n \to \infty$ and by Proposition 2.1 with probability 1 we obtain

$$-\varepsilon + J^{\theta^{0,j}}(u_k) \le J^{\theta^{0,j}}(u_i) + \varepsilon$$

for all $l \in \{1,\dots,N\}$ and $i \in \{1,\dots,r(\varepsilon)\}$. Therefore $(j,k) \in Z$, and we have a contradiction. ■

We are now in a position to complete the proof of Theorem 2.1. Namely, from Lemma 2.4 it follows that for each pair $(j,k) \notin Z$ the jth arm and the control function u_k are played, with probability 1, at the forcing times only. On the other hand, we know that the forcing times are Cesàro rare. Denote by $\chi_Z(j,k)$ the characteristic function of the set Z. Then we have

$$\begin{aligned}
&\limsup_{t\to\infty} t^{-1} \sum_{j=1}^{N}\sum_{k=1}^{r(\varepsilon)} J^{\theta^{0,j}}(u_k)(T_j^k(t)) \\
&\quad = \limsup_{t\to\infty} t^{-1} \sum_{j=1}^{N}\sum_{k=1}^{r(\varepsilon)} J^{\theta^{0,j}}(u_k)\chi_Z(j,k)(T_j^k(t)) \\
&\quad \le \Big(\min_{l=1,\dots,N}\ \min_{i=1,\dots,r(\varepsilon)} J^{\theta^{0,l}}(u_i) + 2\varepsilon\Big) \limsup_{t\to\infty} \sum_{j=1}^{N}\sum_{k=1}^{r(\varepsilon)} \chi_Z(j,k)(T_j^k(t))t^{-1} \\
&\quad \le \min_{l=1,\dots,N}\ \min_{i=1,\dots,r(\varepsilon)} J^{\theta^{0,l}}(u_i) + 2\varepsilon \le \min_{l=1,\dots,N} \lambda(\theta^{0,j}) + 3\varepsilon,
\end{aligned}$$

which completes the proof. ■

3. Strategy with forcing and increasing decision horizon. We now present a strategy with forcing and increasing decision horizon which enables us to obtain a better accuracy of approximation.

The difference between the strategy considered in Section 2 and the one presented below consists in the consideration of an increasing decision horizon. The remaining elements of the strategy are similar.

We start with an auxiliary lemma.

LEMMA 3.1. *Let c_i, $i = 0,1,\dots$, be a bounded sequence. Assume that the set $\mathbb{N}$ of nonnegative integers is partitioned into disjoint infinite subsets $\Phi(i)$, $i = 1,\dots,N$. If for every $j \in \{1,\dots,N\}$ there exist g_j^t, $t = 0,1,\dots$, such that*

$$\limsup_{t\to\infty} \Big| \Big(\sum_{i=0}^{t-1} \chi_{\Phi(j)}(i)\Big)^{-1} \sum_{i=0}^{t-1} c_i \chi_{\Phi(j)}(i) - g_j^t \Big| = 0 \tag{3.1}$$

then

$$\limsup_{t\to\infty} t^{-1}\sum_{i=0}^{t-1} c_i = \limsup_{t\to\infty}\sum_{j=1}^{N} g_j^t t^{-1}\sum_{i=0}^{t-1}\chi_{\Phi(j)}(i). \tag{3.2}$$

P r o o f. We recall formula (2.6):

$$t^{-1}\sum_{i=0}^{t-1} c_i = \sum_{j=1}^{N}\Big(\sum_{i=0}^{t-1}\chi_{\Phi(j)}(i)\Big)^{-1}\sum_{i=0}^{t-1} c_i\chi_{\Phi(j)}(i)t^{-1}\sum_{i=0}^{t-1}\chi_{\Phi(j)}(i).$$

By (3.1) for every $\varepsilon_0 > 0$ there exists t_0 such that for $t \ge t_0$ and $j = 1,\ldots,N$ we have

$$\Big|\Big(\sum_{i=0}^{t-1}\chi_{\Phi(j)}(i)\Big)^{-1}\sum_{i=0}^{t-1} c_i\chi_{\Phi(j)}(i) - g_j^t\Big| \le \varepsilon_0.$$

Then for $t \ge t_0$,

$$\begin{aligned}
&\limsup_{t\to\infty}\Big|t^{-1}\sum_{i=0}^{t-1} c_i - \sum_{j=1}^{N} g_j^t t^{-1}\sum_{i=0}^{t-1}\chi_{\Phi(j)}(i)\Big| \\
&\quad\le \limsup_{t\to\infty}\sum_{j=1}^{N}\Big\{\Big|\Big(\sum_{i=0}^{t-1}\chi_{\Phi(j)}(i)\Big)^{-1}\sum_{i=0}^{t-1} c_i\chi_{\Phi(j)}(i) - g_j^t\Big| t^{-1}\sum_{i=0}^{t-1}\chi_{\Phi(j)}(i)\Big\} \\
&\quad\le \varepsilon_0 \limsup_{t\to\infty} t^{-1}\sum_{j=1}^{N}\sum_{i=0}^{t-1}\chi_{\Phi(j)}(i) = \varepsilon_0.
\end{aligned}$$

Since ε_0 can be chosen arbitrarily small, we obtain (3.2). ∎

By analogy to Section 2 we define a set F' of forcing times

$$\begin{aligned}
F' = \{&0,1,\ldots,Nr(\varepsilon)\kappa, a_1', a_1'+1,\ldots,a_1'+2Nr(\varepsilon)\kappa-1,\ldots \\
&\ldots, a_i', a_i'+1,\ldots,a_i'+2^iNr(\varepsilon)\kappa-1,\ldots\ (i=1,2,\ldots)\}.
\end{aligned}$$

We assume that the sequence a_i' is such that

1) $\limsup_{t\to\infty} t^{-1}\sum_{i=0}^{t-1}\chi_{F'}(i) = 0$,
2) $a_{i+1}' > a_i' + 2^iNr(\varepsilon)\kappa - 1$.

The modification of our control strategy consists now in the fact that we have an increasing decision horizon. First, until a_1' the changes of arms and control functions take place every κ units of time, from a_1' till a_2' every 2κ units of time; and inductively from a_i' till a_{i+1}' every $2^i\kappa$ units.

To construct the sequence a_i' let

$$S(t) = t^{-1}\sum_{i=0}^{t-1}\chi_{F'}(i)$$

and define a'_i such that

$$S(a'_1 + 2\kappa N r(\varepsilon)) = 1/2, \ldots, \qquad S(a'_i + 2^i \kappa N r(\varepsilon)) = 1/2^i.$$

Then $a'_i = \kappa N r(\varepsilon) 2^{i+1}(2^i - 1)$. ■

We divide the time axis into the $Nr(\varepsilon)$ disjoint subsets $\Phi(k,j)$, $k = 1,\ldots,r$, $j = 1,\ldots,N$, such that $\Phi(k,j) = \{\tau_1(k,j), \tau_2(k,j), \ldots\}$ with $\tau_1(k,j), \tau_2(k,j), \ldots$ indicating the successive times at which the control function u_k is used and the jth arm is played.

We have

PROPOSITION 3.1. *There exists C such that $P(C) = 0$ and for $\omega \in \Omega \setminus C$, $k \in \{1,\ldots,r(\varepsilon)\}$ and $j \in \{1,\ldots,N\}$ we have*

$$\lim_{t\to\infty} t^{-1} \sum_{i=0}^{t-1} c(x^j_{\tau_i(k,j)}, u_k(x^j_{\tau_i(k,j)})) = \int_E c(x, u_k(x))\, \pi^j_k(dx) = J^{\theta^{0,j}}(u_k). \tag{3.3}$$

Proof. For $n = 1,2,\ldots,$ $k = 1,\ldots,r(\varepsilon)$ and $j = 1,\ldots,N$ define

$$d(k,j,n) = \inf\{i = 1,2,\ldots : \tau_i(k,j) \ge a^i_n\}.$$

By the strong law of large numbers for martingales, for $n = 1,2,\ldots,$ $k = 1,2,\ldots,r(\varepsilon)$ and $j = 1,\ldots,N$ we have (Lemma 2.3)

$$t^{-1} \sum_{i=0}^{t-1} \Big\{ \sum_{l=d(k,j,n)+i2^n}^{d(k,j,n)+(i+1)2^n-1} c(x^j_{\tau_l(k,j)}, u_k(x^j_{\tau_l(k,j)})) - E\Big\{ \sum_{l=d(k,j,n)+i2^n}^{d(k,j,n)+(i+1)2^n-1} c(x^j_{\tau_l(k,j)}, u_k(x^j_{\tau_l(k,j)})) \,\Big|\, \mathfrak{F}_{\tau_{d(k,j,n)+i2^n}} \Big\}\Big\} \to 0 \quad P\text{-a.e.} \tag{3.4}$$

Using the uniform ergodicity (1.1) we obtain

$$\Big| E\Big\{ \sum_{l=d(k,j,n)+i2^n}^{d(k,j,n)+(i+1)2^n-1} c(x^j_{\tau_l(k,j)}, u_k(x^j_{\tau_l(k,j)})) \,\Big|\, \mathfrak{F}_{\tau_{d(k,j,n)+i2^n}} \Big\} - 2^n \int c(x, u_k(x))\, \pi^j_k(dx) \Big| \le 2\|c\|(1-\gamma)^{-1}. \tag{3.5}$$

Since c is a bounded function for $k = 1,2,\ldots$ we have (compare to Lemma 2.2 and its proof)

$$\limsup_{t\to\infty} t^{-1} \sum_{i=0}^{t-1} c(x^j_{\tau_i(k,j)}, u_k(x^j_{\tau_i(k,j)})) \tag{3.6}$$

$$= \limsup_{t\to\infty} t^{-1} \sum_{i=0}^{t-1} \frac{1}{2^n} \Big\{ \sum_{l=d(k,j,n)+i2^n}^{d(k,j,n)+(i+1)2^n-1} c(x^j_{\tau_l(k,j)}, u_k(x^j_{\tau_l(k,j)})) \Big\}$$

and also

$$\liminf_{t\to\infty} t^{-1} \sum_{i=0}^{t-1} c(x^j_{\tau_i(k,j)}, u_k(x^j_{\tau_i(k,j)})) \tag{3.7}$$
$$= \liminf_{t\to\infty} t^{-1} \sum_{i=0}^{t-1} \frac{1}{2^n} \Big\{ \sum_{l=d(k,j,n)+i2^n}^{d(k,j,n)+(i+1)2^n-1} c(x^j_{\tau_l(k,j)}, u_k(x^j_{\tau_l(k,j)})) \Big\}.$$

Therefore, in order to prove (3.3) it is sufficient to show that for every $\varepsilon_0 > 0$ there exists n_0 such that for $n \ge n_0$,

$$\limsup_{t\to\infty} t^{-1} \sum_{i=0}^{t-1} \frac{1}{2^n} \Big\{ \sum_{l=d(k,j,n)+i2^n}^{d(k,j,n)+(i+1)2^n-1} c(x^j_{\tau_l(k,j)}, u_k(x^j_{\tau_l(k,j)})) \Big\} \le \int_E c(x, u_k(x))\, \pi^j_k(dx) + \varepsilon_0 \tag{3.8}$$

and

$$\liminf_{t\to\infty} t^{-1} \sum_{i=0}^{t-1} \frac{1}{2^n} \Big\{ \sum_{l=d(k,j,n)+i2^n}^{d(k,j,n)+(i+1)2^n-1} c(x^j_{\tau_l(k,j)}, u_k(x^j_{\tau_l(k,j)})) \Big\} \ge \int_E c(x, u_k(x))\, \pi^j_k(dx) - \varepsilon_0. \tag{3.9}$$

Let n be such that $2^{-n}2\|c\|(1-\gamma)^{-1} \le \varepsilon_0$. Then from (3.4) and (3.5) we obtain (3.8) and (3.9), which completes the proof. ■

From Proposition 3.1 and Lemma 3.1 we almost immediately obtain the following corollary:

COROLLARY 3.1. *There exists C such that $P(C) = 0$ and for $\omega \in \Omega \setminus C$, $k \in \{1, \ldots, r(\varepsilon)\}$ and $j \in \{1, \ldots, N\}$ we have*

$$\limsup_{t\to\infty} t^{-1} \sum_{i=0}^{t-1} \sum_{j=1}^{N} c(x^j_i, v_i) S_j(i) = \limsup_{t\to\infty} \sum_{k=1}^{r(\varepsilon)} \sum_{j=1}^{N} \Big(\int_E c(x, u_k(x))\, \pi^j_k(dx) \Big) t^{-1} \sum_{i=0}^{t-1} S^k_j(i), \tag{3.10}$$

where $S^k_j(i)$ is as in (2.1). ■

Outside the forcing moments we use the arm and the control function for which the average cost per unit time over the trajectory is minimal.

Therefore by Proposition 3.1, for sufficiently large t we choose the jth arm and the control function u_k such that

$$\int\limits_E c(x,u_k(x))\,\pi_k^j(dx) = \min_{l=1,\dots,r(\varepsilon)} \int\limits_E c(x,u_l(x))\,\pi_l^j(dx).$$

From the construction of F' it follows that the forcing moments are Cesàro rare, so that from Corollary 3.1 we have

Corollary 3.2. *There exists C such that $P(C)=0$ and for $\omega\in\Omega\setminus C$, $k\in\{1,\dots,r(\varepsilon)\}$ and $j\in\{1,\dots,N\}$ we have*

$$\limsup_{t\to\infty} t^{-1}\sum_{k=1}^{r(\varepsilon)}\sum_{j=1}^{N}\sum_{i=0}^{t-1} c(x_i^j,u_k(x_i^j))S_j^k(i)$$
$$= \min_{j=1,\dots,N}\min_{k=1,\dots,r(\varepsilon)} \int\limits_E c(x,u_k(x))\,\pi_k^j(dx)$$
$$= \min_{j=1,\dots,N}\min_{k=1,\dots,r(\varepsilon)} J^{\theta^{0,j}}(u_k). \ \blacksquare$$

From the above corollary in view of the definition of the class ϑ we obtain

Theorem 3.1. *There exists C such that $P(C)=0$ and for $\omega\in\Omega\setminus C$, $k\in\{1,\dots,r(\varepsilon)\}$ and $j\in\{1,\dots,N\}$ we have*

$$\limsup_{t\to\infty} J(t) = \min_{j=1,\dots,N}\min_{k=1,\dots,r(\varepsilon)} J^{\theta^{0,j}}(u_k) \le \min_{j=1,\dots,N}\lambda(\theta^{0,j})+\varepsilon. \ \blacksquare \tag{3.11}$$

4. Strategy with randomization.

In this section we consider a strategy with randomization. It consists in a randomized choice of arms and control functions. The probabilities in the randomized choice depend on successive calculation of average costs.

The strategy is defined as follows.

1. First for κ (with κ as in (1.4)) moments of time we test every arm and every control function.

2. Let $J(t)$ denote the matrix $J_j^k(t)$, $k=1,\dots,r(\varepsilon)$, $j=1,\dots,N$, defined in (2.1). Define the function $\eta:\mathbb{R}^{Nr(\varepsilon)}\to\mathbb{N}^2$ by

$$\eta([J(t)]) = (\eta_1([J(t)]),\eta_2([J(t)])) = (j_t(\omega),k_t(\omega)) = (j,k),$$

where j,k are such that

$$J_j^k(t) = \min_{l=1,\dots,N}\min_{i=1,\dots,r(\varepsilon)} J_l^i(t)$$

and if $J_j^k(t)=J_l^i(t)$ then either $j<l$ or $j=l$ and $k\le i$.

2a. Let $t^* = Nr(\varepsilon)\kappa$. Define the random variable ξ_{t^*} by the conditional distribution

$$P\{\xi_{t^*}(\omega) = \eta([J(t^*)]) \mid \mathfrak{F}_{t^*}\} = 1 - \varepsilon,$$

where for $t \ge 0$, $\xi_t(\omega) \in \{1, \ldots, N\} \times \{1, \ldots, r(\varepsilon)\}$ and $\mathfrak{F}_t = \sigma(x_0, \ldots, x_t)$, and for $(j, k) \ne \eta([J(t^*)])$,

$$P\{\xi_{t^*}(\omega) = (j, k) \mid \mathfrak{F}_{t^*}\} = \frac{1}{Nr(\varepsilon) - 1}.$$

For the next κ moments of time we choose the pair: arm + number of a control function according to the value of the random variable $\xi_{t^*}(\omega)$.

2b. Let $t \ge (Nr(\varepsilon) + 1)\kappa$. Let $\xi_t(\omega) = \xi_{[t/\kappa]\kappa}(\omega)$, where $[\ ]$ denotes the integer part, and $\xi_0(\omega) = 0$ if $t < Nr(\varepsilon)\kappa$. Define the σ-field $\mathfrak{G}_t(\omega) = \sigma(\xi_0, \ldots, \xi_{t-1})$. For $t > Nr(\varepsilon)\kappa$ such that $t = [t/\kappa]\kappa$ define ξ_t by the conditional distribution

$$P\{\xi_{t^*}(\omega) = \eta([J(t)]) \mid \mathfrak{F}_t \vee \mathfrak{G}_t\} = 1 - \varepsilon, \tag{4.1}$$

where $\mathfrak{F}_t \vee \mathfrak{G}_t = \sigma(x_0, \ldots, x_t, \xi_0, \ldots, \xi_{t-1})$ and for $(j, k) \ne \eta([J(t)])$,

$$P\{\xi_t(\omega) = (j, k) \mid \mathfrak{F}_t \vee \mathfrak{G}_t\} = \frac{1}{Nr(\varepsilon) - 1}. \tag{4.2}$$

For the next κ units of time the arm and the control function are chosen according to the value of $\xi_t(\omega)$. ■

Let

$$\begin{aligned} Z_t = {} & \sum_{i=0}^{t\kappa-1} \sum_{j=1}^{N} c(x_i^j, v_i) S_j(i) \\ & - \sum_{i=0}^{t-1} \sum_{j=1}^{N} \sum_{k=1}^{r(\varepsilon)} \Big\{ \chi_{(j,k)=\eta([J(i\kappa)])} (1-\varepsilon) E_{x_{i\kappa}} \Big\{ \sum_{l=0}^{\kappa-1} c(x_l^j, u_k(x_l^j)) \Big\} \\ & + \chi_{(j,k)\ne\eta([J(i\kappa)])} \frac{\varepsilon}{Nr(\varepsilon) - 1} E_{x_{i\kappa}} \Big\{ \sum_{l=0}^{\kappa-1} c(x_l^j, u_k(x_l^j)) \Big\} \Big\}, \end{aligned}$$

where $\chi_{(j,k)=\eta([J(i\kappa)])} = 1$ if $(j, k) = \eta([J(t)])$ and 0 otherwise, and $\chi_{(j,k)\ne\eta([J(i\kappa)])} = 1$ if $(j, k) \ne \eta([J(t)])$ and 0 otherwise.

LEMMA 4.1. *Z_t is a square integrable martingale with respect to the σ-field $\mathfrak{F}_{t\kappa} \vee \mathfrak{G}_{t\kappa}$ and $(1/(t\kappa))Z_t \to 0$ P-a.e. as $t \to \infty$.*

P r o o f. Notice first that

$$\sum_{i=0}^{t-1} \sum_{j=1}^{N} E\Big\{ \sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l^j, v_l) S_j(l) \;\Big|\; \mathfrak{F}_{i\kappa} \vee \mathfrak{G}_{i\kappa} \Big\} \tag{4.3}$$

$$= \sum_{i=0}^{t-1}\sum_{j=1}^{N}\sum_{k=1}^{r(\varepsilon)}\Big\{\chi_{(j,k)=\eta([J(i\kappa)])}(1-\varepsilon)E_{x_{i\kappa}}\Big\{\sum_{l=0}^{\kappa-1} c(x_l^j, u_k(x_l^j)) \,\Big|\, \mathfrak{F}_{i\kappa} \vee \mathfrak{G}_{i\kappa}\Big\}$$

$$+ \chi_{(j,k)\neq\eta([J(i\kappa)])}\frac{\varepsilon}{Nr(\varepsilon)-1}E_{x_{i\kappa}}\Big\{\sum_{l=0}^{\kappa-1} c(x_l^j, u_k(x_l^j)) \,\Big|\, \mathfrak{F}_{i\kappa} \vee \mathfrak{G}_{i\kappa}\Big\}\Big\}.$$

In fact,

$$\sum_{i=0}^{t-1}\sum_{j=1}^{N} E\Big\{\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l^j, v_l)S_j(l) \,\Big|\, \mathfrak{F}_{i\kappa} \vee \mathfrak{G}_{i\kappa}\Big\}$$

$$= \sum_{i=0}^{t-1}\sum_{j=1}^{N}\sum_{l=i\kappa}^{(i+1)\kappa-1} E\{c(x_l^j, v_l)S_j(l) \mid \mathfrak{F}_{i\kappa} \vee \mathfrak{G}_{i\kappa}\}$$

$$= \sum_{i=0}^{t-1}\sum_{j=1}^{N}\sum_{l=i\kappa}^{(i+1)\kappa-1} E\{E\{c(x_l^j, v_l)S_j(i\kappa) \mid \mathfrak{F}_{i\kappa} \vee \mathfrak{G}_{i\kappa+1}\} \mid \mathfrak{F}_{i\kappa} \vee \mathfrak{G}_{i\kappa}\}$$

$$= \sum_{i=0}^{t-1}\sum_{j=1}^{N}\sum_{l=i\kappa}^{(i+1)\kappa-1} E\{S_j(i\kappa)E\{c(x_l^j, v_l) \mid \mathfrak{F}_{i\kappa} \vee \mathfrak{G}_{i\kappa+1}\} \mid \mathfrak{F}_{i\kappa} \vee \mathfrak{G}_{i\kappa}\}$$

since $S_j(i\kappa)$ is a measurable function with respect to the σ-field $\mathfrak{F}_{t\kappa} \vee \mathfrak{G}_{t\kappa}$. Moreover, for $i\kappa \le l \le (i+1)\kappa$,

$$E\{c(x_l^j, v_l) \mid \mathfrak{F}_{i\kappa} \vee \mathfrak{G}_{i\kappa+1}\} = E_{x_{i\kappa}}\{c(x_{l-i\kappa}^j, u_k(x_{l-i\kappa}^j))\}$$

provided $\xi_{i\kappa}(\omega) = (j,k)$, and

$$\sum_{i=0}^{t-1}\sum_{j=1}^{N}\sum_{k=1}^{r(\varepsilon)} E\Big\{S_j^k(i\kappa)E_{x_{i\kappa}}\Big\{\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_{l-i\kappa}^j, u_k(x_{l-i\kappa}^j))\Big\} \,\Big|\, \mathfrak{F}_{i\kappa} \vee \mathfrak{G}_{i\kappa}\Big\}$$

$$= \sum_{i=0}^{t-1}\sum_{j=1}^{N}\sum_{k=1}^{r(\varepsilon)} E\Big\{\chi_{\xi_{i\kappa}=(j,k)}E_{x_{i\kappa}}\Big\{\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_{l-i\kappa}^j, u_k(x_{l-i\kappa}^j))\Big\} \,\Big|\, \mathfrak{F}_{i\kappa} \vee \mathfrak{G}_{i\kappa}\Big\}$$

$$= \sum_{i=0}^{t-1}\sum_{j=1}^{N}\sum_{k=1}^{r(\varepsilon)} E_{x_{i\kappa}}\Big\{\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_{l-i\kappa}^j, u_k(x_{l-i\kappa}^j))\Big\}P\{\xi_{i\kappa} = (j,k) \mid \mathfrak{F}_{i\kappa} \vee \mathfrak{G}_{i\kappa}\}$$

$$= \sum_{i=0}^{t-1}\sum_{j=1}^{N}\sum_{k=1}^{r(\varepsilon)}\Big\{\chi_{(j,k)=\eta([J(i\kappa)])}(1-\varepsilon)E_{x_{i\kappa}}\Big\{\sum_{l=0}^{\kappa-1} c(x_l^j, u_k(x_l^j))\Big\}$$

$$+ \chi_{(j,k)\neq\eta([J(i\kappa)])}\frac{\varepsilon}{Nr(\varepsilon)-1}E_{x_{i\kappa}}\Big\{\sum_{l=0}^{\kappa-1} c(x_l^j, u_k(x_l^j))\Big\}\Big\}.$$

Therefore (4.3) holds.

Similarly to the proof of Lemma 2.3, we can now show that Z_t is a square integrable martingale and $(1/(t\kappa))Z_t \to 0$ P-a.e. as $t \to \infty$. ∎

From Lemma 4.1 we obtain

COROLLARY 4.1. *The total average cost*

$$J = \limsup_{t\to\infty}(t\kappa)^{-1}\sum_{i=0}^{t-1}\sum_{j=1}^{N}\sum_{k=1}^{r(\varepsilon)} c(x_i^j, u_k(x_i^j))S_j^k(i)$$

is equal to

$$\begin{aligned} J = \limsup_{t\to\infty} t^{-1}\sum_{i=0}^{t-1}\sum_{j=1}^{N}\sum_{k=1}^{r(\varepsilon)} \kappa^{-1}\Big\{&\chi_{(j,k)=\eta([J(i\kappa)])}(1-\varepsilon) \\ &\times E_{x_{i\kappa}}\Big\{\sum_{l=0}^{\kappa-1} c(x_l^j, u_k(x_l^j))\Big\} + \chi_{(j,k)\neq\eta([J(i\kappa)])} \\ &\times \frac{\varepsilon}{Nr(\varepsilon)-1}E_{x_{i\kappa}}\Big\{\sum_{l=0}^{\kappa-1} c(x_l^j, u_k(x_l^j))\Big\}\Big\} \qquad P\text{-a.e.} \ \blacksquare \end{aligned}$$

Moreover, by (1.4) we have

COROLLARY 4.2. *For* $J^{\theta^{0,j}}(u_k) = \int_E c(x, u_k(x))\,\pi_k^j(dx)$ *we have*

$$\limsup_{t\to\infty}\Big|J(t) - t^{-1}\sum_{i=0}^{t-1}\sum_{j=1}^{N}\sum_{k=1}^{r(\varepsilon)} J^{\theta^{0,j}}(u_k)\chi_{(j,k)=\eta([J(t)])}\Big| \le \varepsilon(2\|c\|+1) \qquad P\text{-a.e.}$$

Proof. Let

$$I_1(t) = t^{-1}\sum_{i=0}^{t-1}\sum_{j=1}^{N}\sum_{k=1}^{r(\varepsilon)} \kappa^{-1}\Big\{\chi_{(j,k)=\eta([J(t)])}E_{x_{i\kappa}}\Big\{\sum_{l=0}^{\kappa-1} c(x_l^j, u_k(x_l^j))\Big\}\Big\},$$
$$I_2(t) = t^{-1}\sum_{i=0}^{t-1}\sum_{j=1}^{N}\sum_{k=1}^{r(\varepsilon)} \kappa^{-1}\Big\{\chi_{(j,k)\neq\eta([J(t)])}E_{x_{i\kappa}}\Big\{\sum_{l=0}^{\kappa-1} c(x_l^j, u_k(x_l^j))\Big\}\Big\}.$$

We have

$$\limsup_{t\to\infty}\Big|(1-\varepsilon)I_1(t) + \frac{\varepsilon}{Nr(\varepsilon)-1}I_2(t) - t^{-1}\sum_{i=0}^{t-1}\sum_{j=1}^{N}\sum_{k=1}^{r(\varepsilon)} J^{\theta^{0,j}}(u_k)\chi_{(j,k)=\eta([J(t)])}\Big|$$

$$\leq \limsup_{t\to\infty} \Big| I_1(t) - t^{-1} \sum_{i=0}^{t-1} \sum_{j=1}^{N} \sum_{k=1}^{r(\varepsilon)} J^{\theta^{0,j}}(u_k) \chi_{(j,k)=\eta([J(t)])} \Big|$$

$$+ \varepsilon \limsup_{t\to\infty} |I_1(t)| + \frac{\varepsilon}{Nr(\varepsilon)-1} \limsup_{t\to\infty} |I_2(t)|$$

$$\leq \limsup_{t\to\infty} \Big| I_1(t) - t^{-1} \sum_{i=0}^{t-1} \sum_{j=1}^{N} \sum_{k=1}^{r(\varepsilon)} J^{\theta^{0,j}}(u_k) \chi_{(j,k)=\eta([J(t)])} \Big| + 2\varepsilon \|c\|.$$

Moreover, by uniform ergodicity and the definition of κ,

$$\limsup_{t\to\infty} \Big| I_1(t) - t^{-1} \sum_{i=0}^{t-1} \sum_{j=1}^{N} \sum_{k=1}^{r(\varepsilon)} J^{\theta^{0,j}}(u_k) \chi_{(j,k)=\eta([J(t)])} \Big|$$

$$\leq \limsup_{t\to\infty} \Big| t^{-1} \sum_{i=0}^{t-1} \sum_{j=1}^{N} \sum_{k=1}^{r(\varepsilon)} \kappa^{-1} \Big\{ \chi_{(j,k)=\eta([J(t)])} E_{x_{i\kappa}} \Big\{ \sum_{l=0}^{\kappa-1} c(x_l^j, u_k(x_l^j)) \Big\} \Big\}$$

$$- t^{-1} \sum_{i=0}^{t-1} \sum_{j=1}^{N} \sum_{k=1}^{r(\varepsilon)} \chi_{(j,k)=\eta([J(t)])} \int_E c(x, u_k(x))\, \pi_j^k(dx) \Big|$$

$$\leq \limsup_{t\to\infty} \Big| t^{-1} \sum_{i=0}^{t-1} \sum_{j=1}^{N} \sum_{k=1}^{r(\varepsilon)} \chi_{(j,k)=\eta([J(t)])} \kappa^{-1} \Big\{ E_{x_{i\kappa}} \Big\{ \sum_{l=0}^{\kappa-1} c(x_l^j, u_k(x_l^j)) \Big\} \Big\}$$

$$- \int_E c(x, u_k(x))\, \pi_j^k(dx) \Big| \leq \varepsilon$$

and the proof is complete. ∎

To show the near optimality of the randomized strategy defined above we prove the following auxiliary lemmas:

LEMMA 4.2. *For every $k \in \{1, \ldots, r(\varepsilon)\}$ and $j \in \{1, \ldots, N\}$, under the randomized strategy we have*

$$\limsup_{t\to\infty} t^{-1} T_j^k(t) \geq \frac{\varepsilon}{Nr(\varepsilon)-1} \qquad P\text{-a.e.}$$

Proof. By the definition of the strategy we play the pair (j,k) at each moment of time $t \geq Nr(\varepsilon)\kappa$ with probability greater than or equal to $\varepsilon/(Nr(\varepsilon)-1)$. Let

$$b_t = \sum_{i=0}^{t-1} (\chi_{(j,k)=\eta([J(t)])} - P\{\xi_{i\kappa}(\omega) = (j,k) \mid \mathfrak{F}_{i\kappa} \vee \mathfrak{G}_{i\kappa}\}).$$

Clearly b_t is a square integrable martingale and therefore $(1/t)b_t \to 0$ P-a.e., i.e.

$$\lim_{t\to\infty} t^{-1}\sum_{i=0}^{t-1}(\chi_{(j,k)=\eta([J(i\kappa)])} - P\{\xi_{i\kappa}(\omega)=(j,k)\mid \mathfrak{F}_{i\kappa}\vee\mathfrak{G}_{i\kappa}\}) = 0 \quad P\text{-a.e.}$$

Consequently,

$$\begin{aligned}\liminf_{t\to\infty} t^{-1}\sum_{i=0}^{t-1}\chi_{(j,k)=\eta([J(i\kappa)])}&\\ &= \liminf_{t\to\infty} t^{-1}\sum_{i=0}^{t-1} P\{\xi_{i\kappa}(\omega)=(j,k)\mid \mathfrak{F}_{i\kappa}\vee\mathfrak{G}_{i\kappa}\}\\ &\ge \varepsilon/(Nr(\varepsilon)-1) \quad P\text{-a.e.}\end{aligned}$$

and the conclusion of Lemma 4.2 holds. ∎

LEMMA 4.3. *For $k\in\{1,\ldots,r(\varepsilon)\}$ and $j\in\{1,\ldots,N\}$, there exists C such that $P(C)=0$ and for $\omega\in\Omega\setminus C$ we have*

$$\limsup_{t\to\infty}|J_j^k(t)-J^{\theta^{0,j}}(u_k)|\le\varepsilon. \tag{4.4}$$

P r o o f. The proof parallels that of Proposition 2.1. Observe first that

$$Z_t=\sum_{i=0}^{t\kappa-1} c(x_i^j,v_i)S_j^k(i)-\sum_{i=0}^{t-1}E\Big\{\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_i^j,v_l)S_j^k(i)\ \Big|\ \mathfrak{F}_{i\kappa}\vee\mathfrak{G}_{i\kappa}\Big\}$$

is a square integrable martingale with respect to the σ-field $\mathfrak{F}_{i\kappa}\vee\mathfrak{G}_{i\kappa}$. Hence $(1/(t\kappa))Z_t>0$ as $t\to\infty$ P-a.e. Therefore from Lemma 4.2, $(T_j^k(t\kappa))^{-1}Z_t\to 0$ P-a.e., i.e.

$$\begin{aligned}\limsup_{t\to\infty}(T_j^k(t\kappa))^{-1}\Big|&\sum_{i=0}^{t\kappa-1} c(x_i^j,v_i)S_j^k(i)\\ &-\sum_{i=0}^{t-1}E\Big\{\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_i^j,v_l)S_j^k(i)\ \Big|\ \mathfrak{F}_{i\kappa}\vee\mathfrak{G}_{i\kappa}\Big\}\Big|\to 0 \quad P\text{-a.e.}\end{aligned}$$

Since

$$\begin{aligned}\limsup_{t\to\infty}\Big|(T_j^k(t\kappa))^{-1}\sum_{i=0}^{t-1}E\Big\{\sum_{l=i\kappa}^{(i+1)\kappa-1} c(x_l^j,v_l)S_j^k(l)\ \Big|\ \mathfrak{F}_{i\kappa}\vee\mathfrak{G}_{i\kappa}\Big\}&\\ -\kappa S_j^k(t\kappa)\int_E c(x,u_k(x))\,\pi_k^j(dx)\Big|&\le\varepsilon\end{aligned}$$

and $J^{\theta^{0,j}}(u_k)=\int_E c(x,u_k(x))\,\pi_k^j(dx)$ we obtain (4.4). ∎

LEMMA 4.4. *Let Z be the set of pairs defined in* (2.15). *There exists C such that $P(C)=0$ and for $\omega\in\Omega\setminus C$ if $\xi_{t_n}(\omega)=(j,k)=\eta([J(t_n)])$ for some $(j,k)\in\{1,\dots,N\}\times\{1,\dots,r(\varepsilon)\}$ and $t_n\to\infty$ then $(j,k)\in Z$.*

Proof. Let $\xi_{t_n}(\omega)=(j,k)=\eta([J(t_n)])$ for $t_n\to\infty$ and $\omega\in\Omega\setminus C$ with C as in Lemma 4.3. Then $J^k_j(t_n)\le J^i_l(t_n)$ for $l\in\{1,\dots,N\}$ and $i\in\{1,\dots,r(\varepsilon)\}$. Letting $n\to\infty$ by Lemma 4.3 we obtain

$$-\varepsilon+J^{\theta^{0,j}}(u_k)\le J^{\theta^{0,l}}(u_i)+\varepsilon \quad \text{for } l\in\{1,\dots,N\} \text{ and } i\in\{1,\dots,r(\varepsilon)\}.$$

Therefore

$$-\varepsilon+J^{\theta^{0,j}}(u_k)\le \min_{l=1,\dots,N}\min_{i=1,\dots,r(\varepsilon)} J^{\theta^{0,l}}(u_i)+\varepsilon \quad \text{for } l\in\{1,\dots,N\} \text{ and } i\in\{1,\dots,r(\varepsilon)\}.$$

Hence $(j,k)\in Z$. ■

Finally, we have the following theorem.

THEOREM 4.1. *There exists C such that $P(C)=0$ and for $\omega\in\Omega\setminus C$,*

$$\limsup_{t\to\infty}\Big|J(t)-\min_{j=1,\dots,N}\min_{k=1,\dots,r(\varepsilon)} J^{\theta^{0,j}}(u_k)\Big|\le\varepsilon(2\|c\|+3). \tag{4.5}$$

Proof. By Corollary 4.2 we have

$$\limsup_{t\to\infty}\Big|J(t)-t^{-1}\sum_{i=0}^{t-1}\sum_{j=1}^{N}\sum_{k=1}^{r(\varepsilon)} J^{\theta^{0,j}}(u_k)\chi_{(j,k)=\eta([J(t)])}\Big|\le\varepsilon(2\|c\|+1).$$

By Lemma 4.4 it remains to estimate

$$\limsup_{t\to\infty} t^{-1}\sum_{i=0}^{t-1}\sum_{j=1}^{N}\sum_{k=1}^{r(\varepsilon)} J^{\theta^{0,j}}(u_k)\chi_{(j,k)=\eta([J(t)])}\chi_Z(j,k).$$

For this purpose we repeat the arguments of the proof of Theorem 2.1, and finally obtain (4.5). ■

5. Numerical examples. Below we present some simulation results for the controlled multiarmed bandit problem with the evolution of arms described by the equation

$$x^j_{i+1}=f(x^j_i,u_i,\theta^j)+g(x^j_i)w_i,\qquad x^j_0=x,$$

where $f(x,u,\theta)=\min\{(u(x)\cdot x-\theta)^2+\theta+1,\text{const}\}$ for $\theta\in[-1,1]$, a compact set of unknown parameters, and $w_i\in N(0,1)$ is a white noise. For simplicity assume that $g(x^j_i)=c$ and the cost function is

$$c(x,u(x))=\min\{x^2,\text{const1}\},\quad \text{const}:=100,\quad \text{const1}:=100.$$

It can easily be shown that for a given θ the optimal control function u_θ is

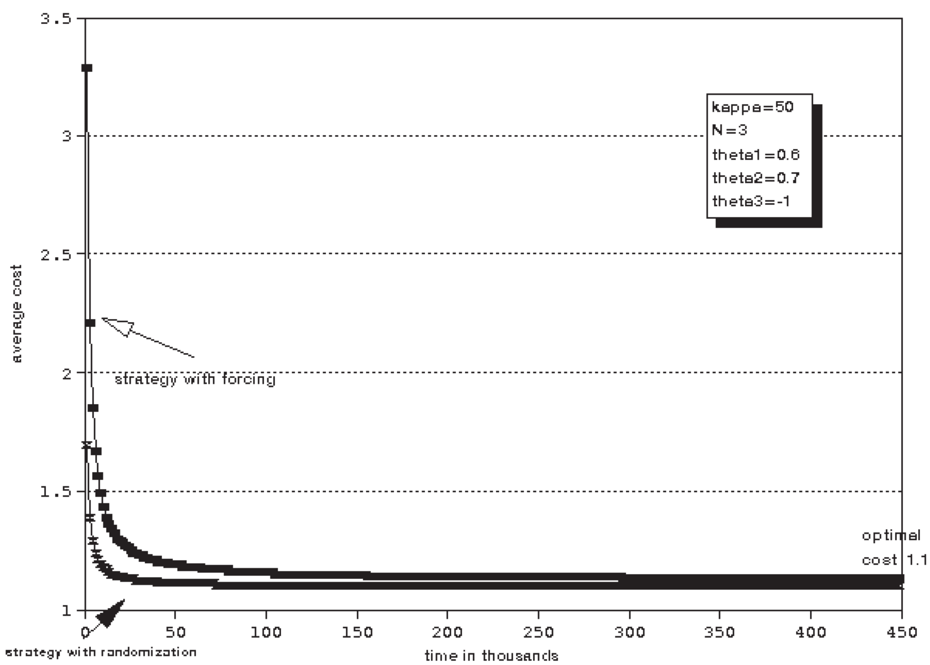
$$u_\theta(x) = \begin{cases} 1 & \text{if } \theta/x \geq 1, \\ \theta/x & \text{if } -1 < \theta/x < 1, \\ -1 & \text{if } \theta/x \leq -1. \end{cases}$$

Therefore we consider the class of admissible control functions

$$\vartheta(\varepsilon) = \{u_\theta : \theta = -1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}.$$

Below we show the graphs obtained by simulations of the above example.

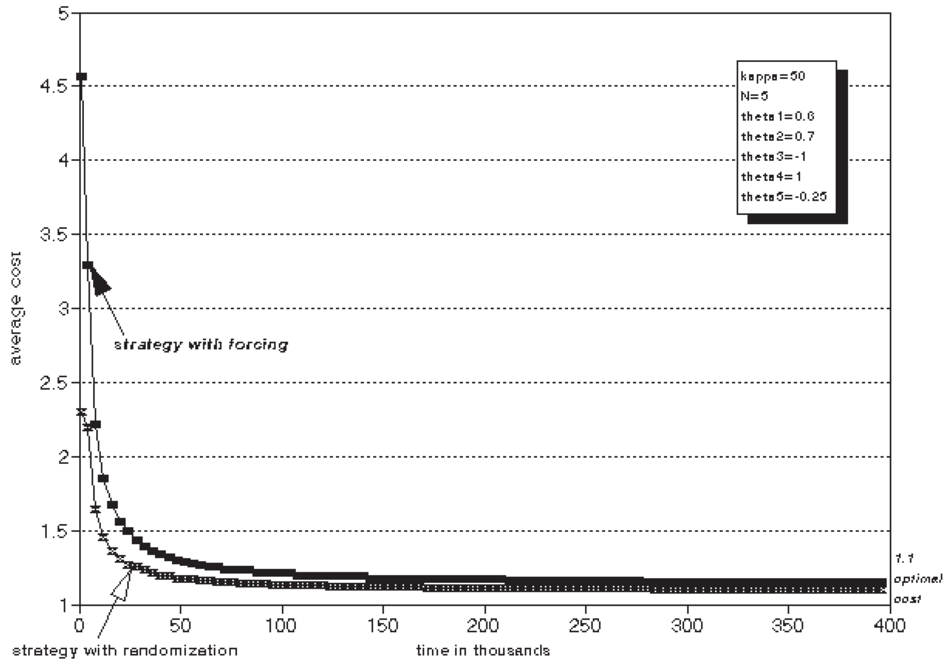
The first graph presents simulation results for the model with forcing and constant time decision horizon, and for the strategy with randomization, for the values of κ, N and θ as indicated.



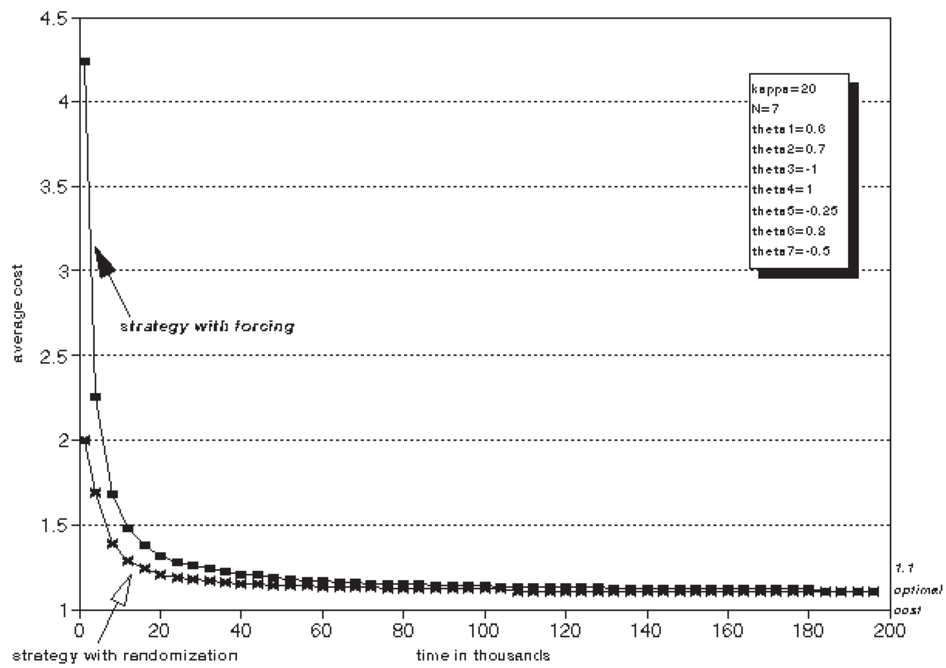
Graph 1

The optimal average cost for the first arm is $J_1 \approx 3.11$, for the second arm it is $J_2 \simeq 3.86$ and for the third arm it is $J_3 \approx 1.1$. The optimal cost for the bandit problem is therefore $J_3 \approx 1.1$ and it indicates that arm 3 should be played.

As is clear from the graph, the strategy with randomization and the strategy with forcing (for large t) come close to the optimal cost $J_3 \approx 1.1$. It should also be noticed that randomization provides faster convergence than forcing.



Graph 2



Graph 3

Similar convergence properties are also obtained for other data. The second graph presents simulation results for the model with forcing and constant decision horizon and for the strategy with randomization, with $\kappa = 50$, $N = 5$. The values of the true parameters θ are as indicated. Notice that the optimal value for the multiarmed bandit problem is equal to 1.1 and corresponds to the 5th arm.

In the third case we also have the optimal cost value for the multiarmed bandit problem equal to about 1.1 and corresponding to the third arm with the values of κ, N and θ as indicated.

The numerical results show that both strategies converge to the optimal cost.

References

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

[2] R. Agrawal and D. Teneketzis, *Certainty equivalence control with forcing: revisited*, Systems Control Lett. 13 (1989), 405–412.

[3] V. Anantharam, P. Varaiya and J. Warland, *Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—Part I: i.i.d. rewards*, IEEE Trans. Automat. Control AC-32 (11) (1987), 969–977.

[4] —, —, —, *Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—Part II: Markovian rewards*, ibid., 977–983.

[5] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. II, Wiley, New York, 1966.

[6] J. C. Gittins, *Multi-armed Bandit Allocation Indices*, Wiley, 1989.

[7] K. D. Glazebrook, *On a sufficient condition for superprocesses due to Whittle*, J. Appl. Probab. 19 (1982), 99–110.

[8] O. Hernández-Lerma, *Adaptative Markov Control Processes*, Springer, 1989.

[9] Ł. Stettner, *On nearly self-optimizing strategies for a discrete-time uniformly ergodic adaptative model*, Appl. Math. Optim. 27 (1993), 161–177.

Ewa Drabik
Institute of Computer Science
Białystok Technical University
Wiejska 45a
15-351 Białystok, Poland

Received on 21.3.1995;
revised version on 23.11.1995