Systems & Control Letters 13 (1989) 405–412
North-Holland

Certainty equivalence control with forcing: revisited

Rajeev AGRAWAL

Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI 53706-1691, U.S.A.

Demosthenis TENEKETZIS

Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122, U.S.A.

Received 26 April 1989
Revised 22 August 1989

Abstract: Certainty equivalence control with forcing has been shown to be optimal for several stochastic adaptive control problems with the average cost per unit time criterion. Recently researchers have started looking at stochastic adaptive control problems with a view to minimizing the rate of increase of the *learning loss*. This criterion is stronger than the average cost per unit time criterion. Certainty equivalence control with forcing does not usually suffice for the *learning loss* criterion and one has to develop fairly complicated schemes in order to achieve optimality. The objective of this paper is to see how well one might be able to do with a certainty-equivalence-control-with-forcing type of scheme. In particular we construct a class of such schemes whose *learning loss* is $O((\log n)^{1+\delta})$ for $\delta > 0$, whereas optimal schemes typically have a $O(\log n)$ *learning loss*.

Keywords: Stochastic adaptive control; certainty equivalence control with forcing; learning loss; multi-armed bandits; Markov chains.

1. Introduction

Stochastic adaptive control problems can roughly be described as follows: There is a stochastic dynamic system whose evolution over time can be influenced by (causally) choosing certain control variables at each time. This choice has to be made so as to optimize the system behaviour with respect to some cost criterion. The exact dynamics of the system or the probabilities governing them are unknown, but are known to belong to some large parametric family. Therefore, the choice of the control variables over time cannot be based on the knowledge of the exact system model (parameter). A common sense approach to stochastic adaptive optimization problems is the so-called *certainty equivalence control* (cf. [8,9]). At each stage, the unknown system parameter is estimated, and assuming that one knows how to control the system if the true parameter were known, one chooses a control action as if the estimated parameter is the true one. A major drawback of the above mentioned approach is that it sometimes leads to the *closed-loop identification* problem, i.e., one mistakenly gets locked on to a false parameter. This happens when there is a conflict between learning and control. One way to deal with the closed-loop identification problem is to use *forcing* controls regularly at some well spread out times which are determined *a priori*. Forcing controls are control actions that are different from those of the certainty equivalence control; the forcing control probes the system and forces one to get out of false locks. The *certainty equivalence control with forcing* strategy is optimal when one is dealing with stochastic adaptive optimization problems with the average-cost-per-unit-time criterion. Such a strategy was first proposed by Robbins [12] in the context of multi-armed bandit problems, and was later used in the context of more general stochastic adaptive control problems; see [8, Chapter 12; 9] and references therein for a detailed description of stochastic adaptive control with the average-cost-per-unit-time criterion.

Recently researchers have attempted to minimize the precise rate of increase of the *learning loss*, or *regret*, or *loss*, i.e., the additional cost one incurs over time because of the inbuilt learning task in stochastic adaptive control problems. (Note that optimality with respect to the average-cost-per-unit-time criterion requires the *learning loss* to be just $o(n)$.) Such a criterion was first used for the multi-armed bandit problem by Lai and Robbins [10,11]. Subsequently, Anantharam, Varaiya and Walrand [5,6], and Agrawal, Hegde and

Teneketzis [1,2] addressed various extensions of the Lai and Robbins formulations of the multi-armed bandit problem. More recently Agrawal, Teneketzis, and Anantharam [3,4] have studied more general stochastic adaptive optimization problems such as the adaptive control of i.i.d. processes and the adaptive control of Markov chains with a view to minimizing the rate of increase of the *learning loss*.

The general approach taken in [1–6,10,11] proceeds in the following steps: First the *learning loss* is interpreted in terms of some notion of experimentation. Then, a lower bound on the asymptotic rate of increase of the *learning loss* and the underlying idea of experimentation is developed. Afterwards, an adaptive control scheme is developed by the following *on-line* procedure: Situations in which there is no conflict between learning and control, and ones in which there is, are differentiated. When there is a conflict between learning and control, the amount of information available and the amount of information required to resolve this conflict is quantified precisely. This is done by using upper confidence bounds in the multi-armed bandit problems, and likelihood ratios and time-varying thresholds in the controlled i.i.d. process and controlled Markov chain problems. The use of upper confidence bounds and likelihood ratios determine the appropriate times for experimention. Finally, an upper bound on the *learning loss* for the constructed adaptive control scheme is obtained; this upper bound equals the lower bound developed earlier, thereby establishing asymptotic efficiency of the constructed adaptive control scheme.

It is clear from the developments in [1–6,10,11] that in order for an adaptive control scheme to achieve the best possible performance it must determine its times for experimentation (or forcing) *on-line* and not *a priori*. Moreover, the schemes constructed in [1–6,10,11] are not only very complicated on that account, but also very sensitive to the parameter spaces they are designed for. This motivates us to go back and look at a simple strategy like *certainty equivalence control with forcing* and see how well we can do with it. In the studies so far [1–6,10,11] the optimal rate of experimentation has been at most O(log n), thus it seems reasonable to believe that if the rate of experimentation or forcing in the certainty equivalence control with forcing scheme is fixed *a priori* to be slightly greater than O(log n), then we should be able to achieve a performance which is close to the optimal. This conjecture is the main focus of this paper. In particular, we construct a class of such schemes for the multi-armed bandit problem in Section 2 and for the adaptive control of Markov chains in Section 3.

2. The multi-armed bandit problem

Consider the following problem: There are $p \geq 2$ arms. Successive plays of arm j, $j = 1, 2, \ldots, p$, yield i.i.d. rewards with a common distribution function $F(\cdot; j, \theta)$ where $F(\cdot; \cdot, \cdot)$ is a known function and θ is an unknown parameter that belongs to a known parameter space Θ. Assume that

$$\int |x| F(\mathrm{d}x; j, \theta) < \infty$$

for all $j = 1, 2, \ldots, p$, and $\theta \in \Theta$. An adaptive allocation rule ϕ consists of a sequence of $\{1, \ldots, p\}$-valued random variables $\{\phi_n\}_{n=1}^{\infty}$, indicating which arm has been selected for play at stage n on the basis of all the past actions and past observations. That is, ϕ_n is a function of only the past actions $\phi_1, \ldots, \phi_{n-1}$ and the past rewards $X_1, \ldots, X_{n-1}$. Let

$$J_n := \sum_{i=1}^{n} X_i \tag{2.1}$$

be the sum of rewards collected upto stage n. The problem is to find an adaptive allocation scheme ϕ which maximizes, in some sense, $\mathrm{E}_\theta J_n$ as $n \to \infty$.

It is easy to show using Wald's Lemma (cf [11]) that

$$\mathrm{E}_\theta J_n = \sum_{j=1}^{p} \mu(j, \theta) \mathrm{E}_\theta T_n(j), \tag{2.2}$$

where

$$T_n(j) := \sum_{i=1}^{n} 1\{\phi_i = j\} \tag{2.3}$$

is the number of times arm j was used upto stage n, and

$$\mu(j, \theta) := \int x F(\mathrm{d}x; j, \theta) \tag{2.4}$$

is the mean reward from arm j under the parameter θ. Let $j^*(\theta)$ denote the best arm, i.e.,

$$\mu(j^*(\theta), \theta) \geq \mu(j, \theta) \quad \text{for all } j = 1, \ldots, p. \tag{2.5}$$

Clearly, if the true parameter θ were known, then the optimal strategy would be to always use the arm $j^*(\theta)$, in which case

$$\mathrm{E}_\theta J_n = n\mu(j^*(\theta), \theta). \tag{2.6}$$

In the absence of the knowledge of θ it is desirable to approach this performance as closely as possible. For this purpose define the *regret* (or *learning loss*) as

$$\begin{aligned} R_n(\theta) &:= n\mu(j^*(\theta), \theta) - \mathrm{E}_\theta J_n \\ &= \sum_{j=1}^{p} (\mu(j^*(\theta), \theta) - \mu(j, \theta))\mathrm{E}_\theta T_n(j). \end{aligned} \tag{2.7}$$

The objective is to design adaptive allocation schemes for which the *regret* increases slowly.

2.1. The adaptive allocation scheme

In this section we construct a class of *certainty equivalence control with forcing* type adaptive allocation schemes and subsequently we upperbound its *regret*. Let $\{b_i\}_{i=1}^{\infty}$ be a positive integer valued sequence to be specified later. Define the related sequence $\{a_i\}_{i=0}^{\infty}$ as follows:

$$a_0 := 0, \tag{2.8a}$$

$$a_i := \sum_{k=1}^{i} (b_k + p) = \sum_{k=1}^{i} b_k + ip, \quad i \geq 1. \tag{2.8b}$$

At times $a_i + j$ $(1 \leq j \leq p,\ i \geq 0)$, use (force) arm j. Let $\hat{j}_i$ be the estimate for $j^*(\theta)$ based on the observations made at times $a_k + j$ $(1 \leq j \leq p,\ 0 \leq k < i)$. Use arm $\hat{j}_i$ from time $a_{i-1} + p + 1$ to the time a_i, i.e. for b_i times. Thus,

$$\phi_n = j, \quad \text{for } n = a_i + j,\ 1 \leq j \leq p,\ i \geq 0, \tag{2.9a}$$

$$\phi_n = \hat{j}_i, \quad \text{for } a_{i-1} + p + 1 \leq n \leq a_i,\ i \geq 1. \tag{2.9b}$$

For the scheme constructed above we can upperbound the expected number of times we use any arm as follows: For any time $n \geq 1$ let $g(n)$ be the smallest integer such that $a_{g(n)} \geq n$. Then, for any arm j,

$$\mathrm{E}_\theta T_n(j) \leq g(n) + \sum_{i=1}^{g(n)} \mathrm{P}_\theta(\hat{j}_i = j) b_i. \tag{2.10}$$

Consequently, for any inferior arm $j \neq j^*(\theta)$,

$$\mathrm{E}_\theta T_n(j) \leq g(n) + \sum_{i=1}^{g(n)} \mathrm{P}_\theta(\hat{j}_i \neq j^*(\theta)) b_i. \tag{2.11}$$

Notice that so far we haven't specified the rules for choosing $\hat{j}_i$, $i \geq 0$, as well as the sequence $\{b_i\}_{i=1}^{\infty}$ which determines the forcing instants. These can now be determined as follows: To choose $\hat{j}_i$, $i \geq 0$, first compute the sample mean,

$$\bar{X}_{ji} = \frac{1}{i} \sum_{k=0}^{i-1} X_{a_k + j}, \tag{2.12}$$

for each arm j, based on the rewards collected from that arm during the time instants forcing is used. Then, choose $\hat{j}_i$, $i \geq 0$, to be the arm with the largest sample mean, i.e.,

$$\bar{X}_{\hat{j}_i i} \geq \bar{X}_{ji} \quad \text{for all } j = 1, \ldots, p. \tag{2.13}$$

Let $\{b_i\}_{i=1}^{\infty}$ be given by

$$b_i := \lfloor \exp(i^{1/(1+\delta)}) \rfloor, \quad \text{for any } \delta > 0, \tag{2.14}$$

where $\lfloor x \rfloor$ denotes the largest integer that is smaller than or equal to x. Then for the above scheme we have the following result:

Theorem 2.1. *Assume that*

$$\int \exp(tx) F(\mathrm{d}x;\ j, \theta) < \infty$$

for all $t \in \mathbb{R}$, $j = 1, 2, \ldots, p$, and $\theta \in \Theta$. Then, for the above scheme, for any inferior arm $j \neq j^*(\theta)$,

$$\mathrm{E}_\theta T_n(j) \leq \mathrm{O}\big((\log n)^{1+\delta}\big). \tag{2.15}$$

Consequently,

$$R_n(\theta) \leq \mathrm{O}\big((\log n)^{1+\delta}\big). \tag{2.16}$$

Proof. From the above assumption on $F(\cdot : \cdot, \cdot)$ it follows (cf. [7], Theorem II.4.1) from the theory of large deviations that

$$\begin{aligned} &\mathrm{P}_\theta\big(\bar{X}_{ji} \notin (\mu(j, \theta) - \varepsilon, \mu(j, \theta) + \varepsilon)\big) \\ &\quad \leq A(j, \theta, \varepsilon) \exp(-\alpha(j, \theta, \varepsilon) i), \end{aligned} \tag{2.17}$$

for all $\varepsilon > 0$, $j = 1, \ldots, p$, $\theta \in \Theta$, for some $A(j, \theta, \varepsilon) \geq 0$, $\alpha(j, \theta, \varepsilon) > 0$.

Choose ε such that $\mu(j^*(\theta), \theta) - \varepsilon > \mu(j, \theta) + \varepsilon$ for all $j \neq j^*(\theta)$. Then it follows that

$$\begin{aligned} &P_\theta\big(\hat{j}_i \neq j^*(\theta)\big) \\ &\quad \leq P_\theta\big(\bar{X}_{ji} \notin (\mu(j, \theta) - \varepsilon, \mu(j, \theta) + \varepsilon) \\ &\qquad\qquad \text{for some } j\big) \\ &\quad \leq \sum_{j=1}^{p} A(j, \theta, \varepsilon) \exp(-\alpha(j, \theta, \varepsilon) i) \\ &\quad \leq A(\theta, \varepsilon) \exp(-\alpha(\theta, \varepsilon) i) \qquad (2.18) \end{aligned}$$

for some $A(\theta, \varepsilon) > 0$, $\alpha(\theta, \varepsilon) > 0$.

Thus,

$$\begin{aligned} &\sum_{i=1}^{g(n)} P_\theta\big(\hat{j}_i \neq j^*(\theta)\big) b_i \\ &\quad \leq \sum_{i=1}^{g(n)} A(\theta, \varepsilon) \exp(-\alpha(\theta, \varepsilon) i) \big|\exp(i^{1/(i+\delta)})\big| \\ &\quad \leq \sum_{i=1}^{\infty} A(\theta, \varepsilon) \exp(-\alpha(\theta, \varepsilon) i) \exp(i^{1/(1+\delta)}) \\ &\quad = K(\theta) \text{ (say)} < \infty. \qquad (2.19) \end{aligned}$$

Now, by definition, $g(n)$ is the smallest integer such that

$$a_{g(n)} = \sum_{i=1}^{g(n)} b_i + g(n) p \geq n.$$

Thus,

$$\sum_{i=1}^{g(n)-1} b_i + (g(n) - 1) p < n.$$

So for $n > a_1$,

$$\begin{aligned} &b_{g(n)-1} < n \\ &\quad \Rightarrow \exp\big((g(n) - 1)^{1/(1+\delta)}\big) < n \\ &\quad \Rightarrow (g(n) - 1)^{1/(1+\delta)} < \log n \\ &\quad \Rightarrow g(n) < (\log n)^{1+\delta} + 1. \qquad (2.20) \end{aligned}$$

Note that for $1 \leq n \leq a_1$,

$$g(n) = 1 \leq (\log n)^{1+\delta} + 1.$$

Thus, for $n \geq 1$,

$$g(n) \leq (\log n)^{1+\delta} + 1.$$

By (2.11), (2.19) and (2.20),

$$E_\theta T_n(j) \leq O\big((\log n)^{1+\delta}\big),$$

and consequently by (2.7),

$$R_n(\theta) \leq O\big((\log n)^{1+\delta}\big). \qquad \square$$

Thus we have constructed a class of adaptive allocation schemes, such that for any given $\delta > 0$ we have a scheme whose *regret* is $O((\log n)^{1+\delta})$. In the next section we extend these results to the more general setting of controlled Markov chains.

3. Adaptive control of Markov chains

Consider a stochastic system described by a controlled Markov chain on the state space $\mathcal{X}$, with control set $\mathcal{U}$, transition probability matrix

$$P(u, \theta) := \{ P(x, y; u, \theta) \mid x, y \in \mathcal{X} \} \qquad (3.1)$$

and initial probability mass function

$$p(\theta) := \{ p(x; \theta) \mid x \in \mathcal{X} \}. \qquad (3.2)$$

The parameter θ is unknown, but belongs to a known set Θ. Assume that $\mathcal{X}$ and $\mathcal{U}$ are finite. Furthermore, assume that for every stationary control law $g: \mathcal{X} \to \mathcal{U}$,

$$P^g(\theta) := \{ P(x, y; g(x), \theta) \mid x, y \in \mathcal{X} \} \qquad (3.3)$$

is irreducible and aperiodic for all $\theta \in \Theta$. Let

$$\pi^g(\theta) := \{ \pi^g(x; \theta) \mid x \in \mathcal{X} \} \qquad (3.4)$$

be the stationary distribution corresponding to $P^g(\theta)$. Let $r(X_i, U_i)$ represent the one-step reward at time i, where $r: \mathcal{X} \times \mathcal{U} \to \mathcal{R}$, and define

$$J_n := \sum_{i=0}^{n-1} r(X_i, U_i), \qquad (3.5)$$

the total reward at time n, as the sum of one-step rewards up to time n.

Our objective is to find an adaptive control scheme γ which maximizes, in some sense, $E_\theta J_n$ as $n \to \infty$.

As a result of a 'translation scheme' developed in [4], we can approximately express $E_\theta J_n$ in terms of the expected number of times each of the stationary control laws g is used up to time n, and the expected one-step reward $\mu^g(\theta)$ under the

invariant distribution corresponding to each stationary control law g as follows:

$$\left|\mathrm{E}_\theta J_n - \sum_{g\in\mathscr{G}} \mu^g(\theta)\mathrm{E}_\theta T_n^g\right| \le k', \tag{3.6}$$

where k' is independent of n,

$$\mu^g(\theta) := \sum_{x\in\mathscr{X}} \pi^g(x;\theta) r(x, g(x)), \tag{3.7}$$

and $\mathscr{G}$ is the set of stationary control laws. Let

$$g^*(\theta) := \arg\max_{g\in\mathscr{G}} (\mu^g(\theta)).$$

Thus, if we knew the true parameter the control scheme $g_n = g^*(\theta)$ gives the optimal reward (upto a constant) for all n, and for this scheme

$$\left|\mathrm{E}_\theta J_n - n\mu^{g^*(\theta)}(\theta)\right| \le K'.$$

In the absence of the knowledge of the true parameter it is desirable to approach this performance as closely as possible. For this purpose define the *loss* associated with an adaptive control scheme γ by

$$L_n(\theta) := n\mu^{g^*(\theta)}(\theta) - \mathrm{E}_\theta J_n. \tag{3.8}$$

By (3.6) it follows that

$$\left|L_n(\theta) - \sum_{\substack{g\in\mathscr{G}\\ g\neq g^*(\theta)}} \left(\mu^{g^*(\theta)}(\theta) - \mu^g(\theta)\right)\mathrm{E}_\theta T_n^g\right| \le \text{const.} \tag{3.9}$$

Maximizing $\mathrm{E}_\theta J_n$ is thus equivalent to minimizing the *loss*. More precisely we want to minimize the rate at which the *loss* increases with n (e.g. finite, logarithmic, linear etc.). In view of (3.9) the above problem is reduced to one of minimizing the rate at which $\mathrm{E}_\theta T_n^g$ increases for $g\in\mathscr{G}$, $g\neq g^*(\theta)$.

3.1. The adaptive control scheme

In this section we construct a class of *certainty equivalence control with forcing* type adaptive control schemes and subsequently we upperbound its *loss*. Let $x_0\in\mathscr{X}$ be an arbitrary but fixed state. Define the $\{\mathscr{F}_n = \sigma(X_0, U_0, X_1, \ldots, X_{n-1}, U_{n-1}, X_n)\}$-stopping times $\tau_0, \tau_1, \ldots$ by

$$\tau_m := \inf\{n > \tau_{m-1} \mid X_n = x_0\}, \quad m \ge 1,$$

and $\tau_0 = 0$. The control scheme we construct chooses a stationary control law G_i at times τ_i, $i\ge 0$, adaptively on the basis of all the past observations and past actions, and uses this control law till $\tau_{i+1} - 1$ respectively. That is, over each recurrence interval marked by the state x_0 we use the same control law which is chosen adaptively at the beginning of the corresponding block. With this in mind we now describe how the choice of control laws is made at the beginning of each block. From now on we shall refer to the actual time as time and the reccurrence points as instances. Initially, i.e. at $n = \tau_0 = 0$, choose a fixed but arbitrary control law G_0 and use it till time $\tau_1 - 1$. *By thinking of stationary control laws as arms and the decision instances as the actual time in the multi-armed bandit problem, we can construct an adaptive control scheme along the same lines as the one for the multi-armed bandit problem.* For the sake of notational convenience let $\mathscr{G} = \{1, \ldots, p\}$ and let $j^*(\theta)$ correspond to $g^*(\theta)$. Let $\{b_i\}_{i=1}^\infty$ be a positive integer valued sequence to be specified later. Define the related sequence $\{a_i\}_{i=0}^\infty$ as follows:

$$a_0 := 0, \tag{3.10a}$$

$$a_i := \sum_{k=1}^{i} (b_k + p) = \sum_{k=1}^{i} b_k + ip, \quad i \ge 1. \tag{3.10b}$$

At instances $a_i + j$ $(1\le j\le p,\ i\ge 0)$, use (force) arm j. Let $\hat{j}_i$ be estimate for $j^*(\theta)$ based on the observations made at the recurrence intervals immediately following the instances $a_k + j$ $(1\le j\le p,\ 0\le k < i)$. Use arm $\hat{j}_i$ from instance $a_{i-1} + p + 1$ to the instance a_i, i.e., for b_i instances. Thus,

$$G_m = j, \quad \text{for } m = a_i + j,\ 1\le j\le p,\ i\ge 0, \tag{3.11a}$$

$$G_m = \hat{j}_i, \quad \text{for } a_{i-1} + p + 1 \le m \le a_i,\ i \ge 1. \tag{3.11b}$$

For the scheme constructed above we can upperbound the expected number of times we use any arm as follows: For any time n let m be the largest integer such that $\tau_m < n$ and let $g(m)$ be the smallest integer such that $a_{g(m)} \ge m$. Note that $m+1$ is a stopping time of $\{\tau_i\}_{i=0}^\infty$. Then, for any arm j,

$$T_n(j) \le \sum_{l=0}^{m} 1(G_l = j)(\tau_{l+1} - \tau_l).$$

Thus,

$$\begin{aligned}
\mathrm{E}_\theta T_n(j) &\le \mathrm{E}_\theta \sum_{l=1}^{\infty} 1(G_l=j)(\tau_{l+1}-\tau_l) \\
&\quad \cdot 1(l<m+1) + \mathrm{E}_\theta \tau_1 \\
&= \sum_{l=1}^{\infty} \mathrm{E}_\theta\Big[\mathrm{E}_\theta\Big[1(G_l=j)1(l<m+1) \\
&\qquad \cdot(\tau_{l+1}-\tau_l)\,|\,\mathscr{F}_{\tau_l}\Big]\Big] + \mathrm{E}_\theta\tau_1 \\
&= \sum_{l=1}^{\infty} \mathrm{E}_\theta\Big[1(G_l=j)1(l<m+1) \\
&\qquad \cdot \mathrm{E}_\theta\Big[(\tau_{l+1}-\tau_l)\,|\,\mathscr{F}_{\tau_l}\Big]\Big] + \mathrm{E}_\theta\tau_1 \\
&= \sum_{l=1}^{\infty} \mathrm{E}_\theta\Big[1(G_l=j)1(l<m+1)\mathscr{T}^j_{\theta,x_0}\Big] \\
&\quad + \mathrm{E}_\theta\tau_1 \\
&= \mathscr{T}^j_{\theta,x_0}\mathrm{E}_\theta \sum_{l=1}^{m} 1(G_l=j) + \mathrm{E}_\theta\tau_1 \qquad (3.12)
\end{aligned}$$

where

$$\mathscr{T}^j_{\theta,x_0} = \mathrm{E}^j_\theta\big[\inf\{n\ge 1 \mid X_n = x_0\} \mid X_0 = x_0\big], \qquad (3.13)$$

is the expected reccurrence time of the state x_0 under the control law j. Let us now examine the term $\sum_{l=1}^{m} 1(G_l=j)$. Clearly, by the construction of the scheme (Eq. (3.11)), it follows that

$$\sum_{l=1}^{m} 1(G_l=j) \le g(m) + \sum_{i=1}^{n} 1\big(\hat{j}_i=j\big)b_i.$$

Thus,

$$\begin{aligned}
&\mathrm{E}_\theta \sum_{l=1}^{m} 1(G_l=j) \\
&\quad \le \mathrm{E}_\theta g(m) + \sum_{i=1}^{n} \mathrm{P}_\theta\big(\hat{j}_i=j\big)b_i. \qquad (3.14)
\end{aligned}$$

Consequently, for any inferior arm $j \ne j^*(\theta)$,

$$\begin{aligned}
&\mathrm{E}_\theta \sum_{l=1}^{m} 1(G_l=j) \\
&\quad \le \mathrm{E}_\theta g(m) + \sum_{i=1}^{n} \mathrm{P}_\theta\big(\hat{j}_i\ne j^*(\theta)\big)b_i. \qquad (3.15)
\end{aligned}$$

Notice that as for the multi-armed bandit problem, so far we haven't specified the rules for choosing $\hat{j}_i$, $i\ge 0$, as well as the sequence $\{b_i\}_{i=1}^{\infty}$ which determines the forcing instances. These can now be determined as follows: To choose $\hat{j}_i$, $i\ge 0$, first compute the sample mean,

$$\bar{r}_{ji} = \frac{\sum_{k=0}^{i-1} \sum_{l=\tau_{a_k+j}}^{\tau_{a_k+j+1}-1} r(X_l,\ j(X_l))}{\sum_{k=0}^{i-1} (\tau_{a_k+j+1} - \tau_{a_k+j})}, \qquad (3.16)$$

for each arm j, based on the rewards collected from that arm during the time forcing is used. Then choose $\hat{j}_i$, $i\ge 0$, to be the arm with the largest sample mean, i.e.,

$$\bar{r}_{\hat{j}_i i} \ge \bar{r}_{ji} \quad \text{for all } j=1,\dots,p. \qquad (3.17)$$

Let $\{b_i\}_{i=1}^{\infty}$ be given by

$$b_i := \big\lfloor \exp(i^{1/(1+\delta)}) \big\rfloor, \quad \text{for any } \delta > 0. \qquad (3.18)$$

Then, for the above scheme we have the following result:

Theorem 3.1. *For the above scheme, for any inferior arm $j \ne j^*(\theta)$,*

$$\mathrm{E}_\theta T_n(j) \le \mathrm{O}\big((\log n)^{1+\delta}\big). \qquad (3.19)$$

Consequently,

$$L_n(\theta) \le \mathrm{O}\big((\log n)^{1+\delta}\big). \qquad (3.20)$$

Proof. It follows from the theory or large deviations (cf. [7], Problem IX.6.12) that

$$\begin{aligned}
&\mathrm{P}_\theta\big(\bar{r}_{ji} \notin (\mu^j(\theta)-\varepsilon,\ \mu^j(\theta)+\varepsilon)\big) \\
&\quad \le A(j,\theta,\varepsilon)\exp(-\alpha(j,\theta,\varepsilon)i), \qquad (3.21)
\end{aligned}$$

for all $\varepsilon>0$, $j=1,\dots,p$, $\theta\in\Theta$, for some $A(j,\theta,\varepsilon)\ge 0$, $\alpha(j,\theta,\varepsilon)>0$.

Choose ε such that $\mu^{j^*(\theta)}(\theta) - \varepsilon > \mu^j(\theta)+\varepsilon$ for all $j\ne j^*(\theta)$. Then it follows that

$$\begin{aligned}
&\mathrm{P}_\theta\big(\hat{j}_i \ne j^*(\theta)\big) \\
&\quad \le \mathrm{P}_\theta\big(\bar{r}_{ji} \notin (\mu^j(\theta)-\varepsilon,\ \mu^j(\theta)+\varepsilon) \text{ for some } j\big) \\
&\quad \le \sum_{j=1}^{p} A(j,\theta,\varepsilon)\exp(-\alpha(j,\theta,\varepsilon)i) \\
&\quad \le A(\theta,\varepsilon)\exp(-\alpha(\theta,\varepsilon)i) \qquad (3.22)
\end{aligned}$$

for some $A(\theta,\varepsilon)>0$, $\alpha(\theta,\varepsilon)>0$.

Thus,

$$\sum_{i=1}^{n} P_\theta\big(\hat{j}_k \neq j^*(\theta)\big) b_i$$
$$\leq \sum_{i=1}^{n} A(\theta, \varepsilon) \exp(-\alpha(\theta, \varepsilon) i) \big[\exp(i^{1/(1+\delta)})\big]$$
$$\leq \sum_{i=1}^{\infty} A(\theta, \varepsilon) \exp(-\alpha(\theta, \varepsilon) i) \exp(i^{1/(1+\delta)})$$
$$= K(\theta) \text{ (say)} < \infty. \qquad (3.23)$$

Now, by definition, $g(m)$ is the smallest integer such that

$$a_{g(m)} = \sum_{i=1}^{g(m)} b_i + g(m) p \geq m.$$

Hence,

$$\sum_{i=1}^{g(m)-1} b_i + (g(m) - 1) p < m.$$

So for $m > a_1$,

$$b_{g(m)-1} < m$$
$$\Rightarrow \exp\big((g(m) - 1)^{1/(1+\delta)}\big) < m$$
$$\Rightarrow (g(m) - 1)^{1/(1+\delta)} < \log m$$
$$\Rightarrow g(m) < (\log m)^{1+\delta} + 1$$
$$< (\log n)^{1+\delta} + 1. \qquad (3.24)$$

Note that for $1 \leq m \leq a_1$,

$$g(m) = 1 \leq (\log m)^{1+\delta} + 1 \leq (\log n)^{1+\delta} + 1.$$

Thus, for $n \geq 1$,

$$g(m) \leq (\log n)^{1+\delta} + 1.$$

By (3.12), (3.15), (3.23) and (3.24),

$$E_\theta T_n(j) \leq O\big((\log n)^{1+\delta}\big),$$

and consequently by (3.9),

$$L_n(\theta) \leq O\big((\log n)^{1+\delta}\big). \qquad \square$$

Thus we have constructed a class of adaptive control schemes, such that for any given $\delta > 0$ we have a scheme whose *loss* is $O((\log n)^{1+\delta})$.

4. Conclusions

We have constructed a certainty-equivalence-control-with-forcing type scheme which has the following features: (i) it is very simple; (ii) it achieves a performance that can be arbitrarily close to that of asymptotically efficient control schemes; and (iii) it can be used to analyze problems for which it is very difficult to determine asymptotically efficient control schemes (e.g. the adaptive control problem of Section 3).

There is one significant difference between the proposed scheme and the certainty-equivalence-with-forcing type schemes that have appeared so far in the literature (e.g. [8,9] and references therein). In the scheme proposed in this paper, the certainty-equivalence-control is based on the estimate of the optimal control law, whereas in all other schemes the certainty-equivalence-control is based on the estimate of the true parameter, that is, on the result of identification. Thus, in the scheme proposed in this paper, attention is entirely focused on the control part of the adaptive optimization problem.

Acknowledgements

This work was supported in part by the National Science Foundation under Grant ECS-8517708 and by the Office of Naval Research under Grant N00014-87-K-0540.

References

[1] R. Agrawal, M. Hegde and D. Teneketzis, Asymptotically efficient adaptive allocation rules for the multi-armed bandit problem with switching cost, *IEEE Trans. Automat. Control* **33** (10) (1988) 899–906.

[2] R. Agrawal, M. Hegde and D. Teneketzis, Multi-armed bandit problems with multiple plays and switching cost, Technical Report No. 258, Communication and Signal Processing Laboratory, University of Michigan, Ann Arbor, MI (Dec. 1988); to appear in *Stochastics*.

[3] R. Agrawal, D. Teneketzis and V. Anantharam, Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite parameter space, *IEEE Trans. Automat. Control* **34** (3) (March 1989) 258–267.

[4] R. Agrawal, D. Teneketzis and V. Anantharam, Asymptotically efficient adaptive allocation schemes for controlled Markov chains: Finite parameter space, Technical Report No. 254, Communications and Signal Processing

Laboratory, University of Michigan, Ann Arbor, MI, February 1988; to appear in *IEEE Trans. Automat. Control.*

[5] 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. Control* **32** (11) (1987) 968–975.

[6] V. Anantharam, P. Varaiya and J. Walrand, Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays; Part II: Markovian rewards, *IEEE Trans. Automat. Control* **32** (11) (1987) 975–982.

[7] R.S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics* (Springer-Verlag, Berlin–New York, 1985).

[8] P.R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control* (Prentice-Hall, Englewood Cliffs, NJ, 1986).

[9] P.R. Kumar, A survey of some results in stochastic adaptive control, *SIAM J. Control Optim.* **23** (3) (1985) 329–380.

[10] T.L. Lai and H. Robbins, Asymptotically efficient adaptive allocation rules, *Adv. Appl. Math.* **6** (1985) 4–22.

[11] T.L. Lai and H. Robbins, Asymptotically efficient allocation of treatments in sequential experiments, in: T.J. Santner and A.C. Tamhane, Eds., *Design of Experiments* (Marcel Dekker, New York, 1984) 127–142.

[12] H. Robbins, Some aspects of the sequential design of experiments, *Bull. Amer. Math. Soc.* **55** (1952) 527–535.