# Asymptotically Efficient Adaptive Allocation Schemes for Controlled I.I.D. Processes: Finite Parameter Space

RAJEEV AGRAWAL, DEMOSTHENIS TENEKETZIS, AND VENKATACHALAM ANANTHARAM, MEMBER, IEEE

*Abstract*—**We consider a controlled i.i.d. process whose distribution is parametrized by an unknown parameter $\theta$ belonging to some known parameter space $\Theta$, and a one-step reward associated with each pair of control and the following state of the process. The objective is to maximize the expected value of the sum of one-step rewards over an infinite horizon. By introducing the loss associated with a control scheme, we show that our problem is equivalent to minimizing this loss. We define uniformly good adaptive control schemes and restrict attention to these schemes. We develop a lower bound on the loss associated with any uniformly good control scheme. Finally, we construct an adaptive control scheme whose loss equals the lower bound, and is therefore asymptotically efficient.**

## I. Introduction

CONSIDER the following stochastic adaptive control problem. The system is modeled as a controlled i.i.d. process with an unknown parameter, i.e., the state $X_n$ at time $n$ is distributed as $p(X_n; U_n, \theta)$ where $U_n$ is the control preceding $X_n$ and $\theta$ is an unknown parameter belonging to some known parameter space $\Theta$. There is a one-step reward associated with each pair of control and the following state: $r(X_n, U_n)$. The objective is to find an adaptive control scheme which maximizes, in some sense, the expected value of the sum of one-step rewards

$$E_\theta J_n = E_\theta \sum_{i=1}^{n} r(X_i, U_i), \qquad \text{as } n \to \infty. \tag{1.1}$$

One of the current approaches to stochastic adaptive control problems is the so-called certainty equivalent control with forcing (cf. [1]). This scheme is self-tuning in the Cesaro sense (cf. [1]) and is therefore also optimal for an average reward per unit time criterion. The reward criterion described by (1.1) suggests that we need to determine the maximum rate of increase of $E_\theta J_n$ as $n \to \infty$. This requirement introduces a notion of optimality that is stronger than the one suggested by the average reward per unit time criterion. For the criterion (1.1) it is no longer clear that the certainty equivalent control with forcing is optimal.

The same reward criterion as (1.1) was previously used by Lai and Robbins [2], [3] in their study of multiarmed bandit problems. Various extensions of the Lai and Robbins formulation of multiarmed problems have been reported in [4] and [5]. In this paper we show that the adaptive control problem of i.i.d. processes can be interpreted as a bandit problem. Such an interpretation provides a convenient way of analyzing the problem, and allows us to develop an efficient adaptive control scheme.

The paper is organized as follows. In Section II we give a precise formulation of the problem and relate it to the multiarmed bandit problem. In Section III we obtain an asymptotic lower bound on the loss associated with any adaptive control scheme. In Section IV we construct an adaptive control scheme which achieves the lower bound.

Manuscript received February 29, 1988; revised August 4, 1988. Paper recommended by Associate Editor, A. Haurie. 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.

R. Agrawal was with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122. He is now with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53705.

D. Teneketzis is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122.

V. Anantharam is with the School of Electrical Engineering, Cornell University, Ithaca, NY 14853.

IEEE Log Number 8825853.

## II. The Problem

### A. Problem Formulation

Consider a stochastic system described by a controlled i.i.d. process on the state-space $\mathfrak{X}$, with control set $\mathfrak{U}$, and the probability mass function (of $X_n$, $n = 1, 2, \ldots$)$p(x; u, \theta)$. The parameter $\theta$ is unknown, but belongs to a known set $\Theta$. Assume that $\mathfrak{X}$, $\mathfrak{U}$, and $\Theta$ are all finite.

An adaptive control scheme $\phi$ is a sequence of random variables $\{U_n\}_{n=1}^{\infty}$ taking values in the set $\mathfrak{U}$ such that the event $\{U_n = u\}$ belongs to the $\sigma$-field $\mathfrak{F}_{n-1}$ generated by $U_1, X_1, U_2, X_2, \ldots, U_{n-1}, X_{n-1}$. Let $r(X_i, U_i)$ represent the one-step reward, where $r: \mathfrak{X} \times \mathfrak{U} \to \mathbb{R}$. Further, define $J_n := \Sigma_{i=1}^{n} r(X_i, U_i)$, the total reward at stage $n$ as the sum of the one-step rewards up to stage $n$.

Our objective is to find an adaptive control scheme $\phi$ which maximizes, in some sense, $E_\theta J_n$ as $n \to \infty$. We shall shortly clarify this notion of optimality.

Let $R_\theta(u) := E_\theta r(X, u)$. For any given $\theta \in \Theta$, let $u^*(\theta)$ be the control that maximizes $R_\theta(u)$ over all $u \in \mathfrak{U}$. Assume for simplicity that $u^*(\theta)$ is unique for each $\theta \in \Theta$.

Define

$$T_n(u) := \sum_{i=1}^{n} 1(U_i = u)$$

as the number of times the control $u$ was used up to stage $n$. Then it follows that:

$$\begin{aligned} E_\theta J_n &= E_\theta \sum_{i=1}^{n} r(X_i, U_i) \\ &= \sum_{i=1}^{n} \sum_{u \in \mathfrak{U}} E_\theta[r(X_i, u) 1(U_i = u)] \\ &= \sum_{u \in \mathfrak{U}} \sum_{i=1}^{n} E_\theta[E_\theta[r(X_i, u) 1(U_i = u) | \mathfrak{F}_{i-1}]] \\ &= \sum_{u \in \mathfrak{U}} R_\theta(u) E_\theta T_n(u). \end{aligned} \tag{2.1}$$

This clearly implies that if we know $\theta$, the control policy that would maximize $E_\theta J_n$ for all $n$ is the stationary (in fact, constant) control $U_i = u^*(\theta)$, in which case the expected total reward is $nR_\theta(u^*(\theta))$. In the absence of the knowledge of the true parameter, it is desirable to approach this performance as closely as possible. For this purpose we define the loss

$$L_n(\theta) := nR_\theta(u^*(\theta)) - E_\theta J_n. \tag{2.2}$$

Maximizing $E_\theta J_n$ 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.). Thus, this is a stronger criterion for optimality than the average reward per unit time criterion which only requires the loss to be $o(n)$. In view of (2.1), it follows that:

$$L_n(\theta) = \sum_{u \in \mathfrak{U}} (R_\theta(u^*(\theta)) - R_\theta(u)) E_\theta T_n(u) \tag{2.3}$$

which reduces the above problem to minimizing the rate at which $E_\theta T_n(u)$ increases for $u \in \mathfrak{U}$, $u \neq u^*(\theta)$.

Note that it is impossible to minimize $L_n(\theta)$ uniformly over all parameters $\theta \in \Theta$. For example, the constant control scheme $U_i = u^*(\theta)$ for all $i$ will have zero loss when the true parameter is $\theta$. However, when the true parameter is $\theta'$ such that $u^*(\theta') \neq u^*(\theta)$, then this scheme will have a loss proportional to $n$. Having made this observation we call a scheme uniformly good if for every parameter $\theta \in \Theta$

$$L_n(\theta) = o(n^a) \qquad \text{for every } a > 0. \tag{2.4}$$

Such schemes do not allow the loss to increase very rapidly for any $\theta \in \Theta$. We restrict attention to the class of uniformly good schemes, and consider any others as uninteresting.

The loss as a performance measure has previously been used in the classical multiarmed bandit problem. In the formulation of the bandit problem appearing in [2]–[5], the objective, like in our problem, is to minimize the loss over the class of uniformly good schemes. The formulation of the bandit problem proposed in [8] and [9] considers a minimax loss criterion.

To complete the problem formulation, we make the following technical assumption: for all

$$x \in \mathfrak{X},\ u \in \mathfrak{U},\ \theta, \theta' \in \Theta,\ p(x; u, \theta) > 0 \Rightarrow p(x; u, \theta') > 0. \tag{2.5}$$

The above assumption means that for any $u \in \mathfrak{U}$, the distributions of the state under any two parameters $\theta$ and $\theta'$ are mutually absolutely continuous. This in turn implies that the Kullback–Leibler number

$$I^u(\theta, \theta') := \sum_{x \in \mathfrak{X}} p(x; u, \theta) \log \frac{p(x; u, \theta)}{p(x; u, \theta')} \tag{2.6}$$

which is a well-known distance measure between two distributions, is finite. If under $\theta$, $I^u(\theta, \theta')$ was infinite for some $u \in \mathfrak{U}$ and $\theta' \in \Theta$, then distinguishing between $\theta$ and $\theta'$ would be a trivial identification task. We shall not treat this case here as it does not contribute to the conceptual understanding of our adaptive control problem.

### *B. Relation to the Multiarmed Bandit Problem*

The controlled i.i.d. process problem formulated in Section II-A can be viewed as a multiarmed bandit problem. Consider each control action $u \in \mathfrak{U}$ as an arm, and $r(X_1, U_1)$, $r(X_2, U_2)$, ..., as the sequence of rewards obtained by choosing arms $U_1$, $U_2$, ..., etc. Since the set $\mathfrak{U}$ is finite, we have a finite number of arms which is a usual condition in multiarmed bandit problems. Then, the problem of finding an adaptive control scheme so as to maximize $E_\theta \Sigma_{i=1}^n r(X_i, U_i) = E_\theta J_n$ as $n \to \infty$ is essentially the same as finding an adaptive allocation rule in the multiarmed bandit problem as formulated by Lai and Robbins [2], [3]. However, there is one *major* difference between our formulation of the controlled i.i.d. process problem and the multiarmed bandit problem in [2] and [3]. This is in the way the parameter space $\Theta$ is defined in the two problems. In the controlled i.i.d. sequence problem the parameter $\theta \in \Theta$ parametrizes *all* the arms $u \in \mathfrak{U}$, whereas in the multiarmed bandit problem the parameter $\theta_i \in \Theta$ parametrizes the *individual* arms. Thus, $\Theta$ in the controlled i.i.d. sequence problem corresponds to $\Theta^p$ in the multiarmed bandit problem, where $p$ is the number of arms. Therefore, the parameter space of the multiarmed bandit problem has the following special structure. Each of the arms is parametrized on the same set of distributions, and any combination of parameters for individual arms is allowed. This structure is not necessarily present in the parameter space of the controlled i.i.d. process problem. As a consequence, the minimum loss over the class of uniformly good control schemes is different for the two problems.

Having given a precise formulation of the problem and having discussed its relation to the multiarmed bandit problem, we now proceed in two main steps. First in Section III we present an asymptotic lower bound on the loss. Then in Section IV we construct a scheme which achieves the lower bound.

## III. A Lower Bound on the Loss

In this section we obtain a lower bound on the loss $L_n(\theta)$ for certain values of the parameter $\theta \in \Theta$. Before we present the bound, we introduce the necessary notation. Let

$$B(\theta) = \{\theta' \in \Theta : p(\cdot; u^*(\theta), \theta') = p(\cdot; u^*(\theta), \theta) \text{ and } u^*(\theta') \neq u^*(\theta)\},$$

$$\mathfrak{U}_\theta = \mathfrak{U} - \{u^*(\theta)\},$$

$$\mathfrak{A}_\theta = \left\{(\alpha^u, u \in \mathfrak{U}_\theta) : \alpha^u \geq 0, \sum_{u \in \mathfrak{U}_\theta} \alpha^u = 1\right\},$$

$$d_\theta(u) = R_\theta(u^*(\theta)) - R_\theta(u). \tag{3.1}$$

The bound is now presented in the form of Theorem 3.1 below.

*Theorem 3.1:* Let $\theta \in \Theta$ be such that $B(\theta)$ is nonempty. Then for any uniformly good control scheme $\phi$, under the parameter $\theta$,

$$1)\ \lim_{n \to \infty} P_\theta \left\{ \sum_{\mathfrak{U}_\theta} T_n(u) d_\theta(u) < \frac{\log n}{1+2\rho} \cdot \frac{1}{\max\limits_{\alpha \in \mathfrak{A}_\theta} \min\limits_{\theta' \in B(\theta)} \dfrac{\sum\limits_{\mathfrak{U}_\theta} \alpha^u I^u(\theta, \theta')}{\sum\limits_{\mathfrak{U}_\theta} \alpha^u d_\theta(u)}} \right\} = 0 \quad \forall\, \rho > 0. \tag{3.2}$$

Consequently,

$$2)\ \liminf_{n \to \infty} \frac{L_n(\theta)}{\log n} \geq \min_{\alpha \in \mathfrak{A}_\theta} \max_{\theta' \in B(\theta)} \frac{\sum\limits_{\mathfrak{U}_\theta} \alpha^u d_\theta(u)}{\sum\limits_{\mathfrak{U}_\theta} \alpha^u I^u(\theta, \theta')}. \tag{3.3}$$

[Note that $I^u(\theta, \theta')$ is the Kullback–Leibler number that was defined by (2.6).]

*Discussion:* The key steps in the proof of Theorem 3.1 can be

described as follows. Define the event

$$A_n := \left\{ \sum_{\mathcal{U}_\theta} T_n(u) d_\theta(u) < \frac{\log n}{1+2\rho} \cdot \frac{1}{\max_{\alpha \in \mathcal{A}_\theta} \min_{\theta' \in B(\theta)} \frac{\sum_{\mathcal{U}_\theta} \alpha^u I^u(\theta, \theta')}{\sum_{\mathcal{U}_\theta} \alpha^u d_\theta(\mathcal{U})}} \right\}.$$

In order to show that $\lim_{n\to\infty} P_\theta(A_n) = 0$, we consider any $\theta' \in B(\theta)$ and relate $P_\theta(A_n)$ to $P_{\theta'}(A_n)$ by a measure transformation (Radon–Nikodym derivative). The measure transformation is the likelihood ratio of $\theta'$ versus $\theta$. Since under $\theta$ and $\theta' \in B(\theta)$ the distribution under $u^*(\theta)$ coincides, this likelihood ratio reduces to the likelihood ratio of samples obtained when $u \neq u^*(\theta)$ is used. An upper bound is then obtained on $P_\theta(A_n)$ by getting an upper bound on $P_{\theta'}(A_n)$ and a lower bound on the measure transformation (from $\theta$ to $\theta'$). Actually, we use a prior distribution $\Pi(\theta')$ on $B(\theta)$ as an artifice to obtain a tighter upper bound on $P_\theta(A_n)$ than the one that would be obtained by considering each $\theta' \in B(\theta)$ separately. Finally, we get (3.2) by showing that in the limit the upper bound on $P_\theta(A_n)$ tends to 0 as $n \to \infty$.

We now present two lemmas before we prove Theorem 3.1. Lemma 3.1 is useful in getting the lower bound on the measure transformation (from $\theta$ to $\theta'$) and it uses the fact that the normalized log likelihood ratio of samples obtained under a fixed control $u$ converges to $I^u(\theta, \theta')$ a.s. $P_\theta$ by the strong law of large numbers. Lemma 3.2 obtains the required upper bound on $P_{\theta'}(A_n)$ by exploiting the fact that the rule in consideration is uniformly good.

*Lemma 3.1:* For each $u \in \mathcal{U}$, let $X_1^u, X_2^u, \cdots$, be the sequence of states observed when the preceding control action is $u$. Then for every $\epsilon > 0$, $\rho > 0$, $u \in \mathcal{U}$, and $\theta' \in \Theta$ there is a constant $K^u(\epsilon, \rho, \theta') < \infty$, and an event $A^u(\epsilon, \rho, \theta')$ with $P_\theta(A^u(\epsilon, \rho, \theta')) > 1 - \epsilon$, such that

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{p(X_i^u; u, \theta)}{p(X_i^u; u, \theta')} < (1+\rho) I^u(\theta, \theta') + \frac{K^u(\epsilon, \rho, \theta')}{n}$$

$$\text{for all } n \geq 1 \text{ on } A^u(\epsilon, \rho, \theta'). \quad (3.4)$$

*Proof:* For $\theta' \in \Theta$ and $u \in \mathcal{U}$ such that $p(\cdot; u, \theta') = p(\cdot; u, \theta)$, the result is trivial. Otherwise, note that under $\theta$, $X_1^u, X_2^u, \cdots$, are i.i.d. r.v.'s with marginal $p(x; u, \theta)$. By the strong law of large numbers it follows that

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{p(X_i^u; u, \theta)}{p(X_i^u; u, \theta')} \to I^u(\theta, \theta') \quad \text{a.s. } P_\theta \text{ as } n \to \infty. \quad (3.5)$$

Thus,

$$T := \max \left\{ n \geq 1 : \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(X_i^u; u, \theta)}{p(X_i^u; u, \theta')} \geq (1+\rho) I^u(\theta, \theta') \right\} < \infty \text{ a.s. } P_\theta \quad (3.6)$$

and consequently by (2.5) it follows that:

$$\max_{1 \leq n \leq T} \sum_{i=1}^{n} \log \frac{p(X_i^u; u, \theta)}{p(X_i^u; u, \theta')} < \infty \quad \text{a.s. } P_\theta. \quad (3.7)$$

Equation (3.4) follows from (3.6) and (3.7). □

*Lemma 3.2:* For any uniformly good rule, under $P_{\theta'}$, $\theta' \in B(\theta)$, we have

$$P_{\theta'}\left\{ \sum_{\mathcal{U}_\theta} T_n(u) d_\theta(u) < k \log n \right\} = o(n^{a-1}) \quad \text{for every } a,\ k > 0. \quad (3.8)$$

*Proof:*

$$P_{\theta'}\left\{ \sum_{\mathcal{U}_\theta} T_n(u) d_\theta(u) < k \log n \right\}$$

$$= P_{\theta'}\left\{ \sum_{\mathcal{U}} T_n(u) d_\theta(u^*(\theta')) - \sum_{\mathcal{U}} T_n(u) d_\theta(u) > n d_\theta(u^*(\theta')) - k \log n \right\}$$

$$= P_{\theta'}\left\{ \sum_{\mathcal{U}_{\theta'}} T_n(u)[R_\theta(u) - R_\theta(u^*(\theta'))] > n d_\theta(u^*(\theta')) - k \log n \right\}$$

$$\leq P_{\theta'}\left\{ \sum_{\mathcal{U}_{\theta'}} T_n(u) d_{\theta'}(u) \geq \frac{\min_{u \in \mathcal{U}_{\theta'}} d_{\theta'}(u)}{d_\theta(u^*(\theta'))} \cdot (n d_\theta(u^*(\theta')) - k \log n) \right\}$$

$$= P_{\theta'}\left\{ \sum_{\mathcal{U}_{\theta'}} T_n(u) d_{\theta'}(u) \geq bn - c \log n \right\}$$

for appropriate constants $b$, $c > 0$,

$$\leq \frac{L_n(\theta')}{bn - c \log n} \quad \text{by Markov's inequality}$$

$$= o(n^{a-1}) \quad \text{by the definition of uniformly good.} \quad \square$$

*Proof of Theorem 3.1:* Fix a distribution $\Pi(\theta')$ on $B(\theta)$ with $\Pi(\theta') > 0$ for every $\theta' \in B(\theta)$. Let $P_\Pi$ denote the corresponding distribution on the sequence of controls and observed states. Note that, conditional on knowing $\theta'$, this is just $P_{\theta'}$. Then, for any event $A_n \in \mathcal{F}_n = \sigma(U_1, X_1, U_2, X_2, \cdots, U_n, X_n)$

$$P_\Pi(A_n) = \sum_{\theta' \in B(\theta)} \Pi(\theta') P_{\theta'}(A_n)$$

$$= \sum_{\theta' \in B(\theta)} \Pi(\theta') \int_{A_n} \prod_{u \in \mathcal{U}} \prod_{i=1}^{T_n(u)} \frac{p(X_i^u; u, \theta')}{p(X_i^u; u, \theta)}\, dP_\theta$$

$$= \int_{A_n} \sum_{\theta' \in B(\theta)} \Pi(\theta') \prod_{u \in \mathcal{U}_\theta} \prod_{i=1}^{T_n(u)} \frac{p(X_i^u; u, \theta')}{p(X_i^u; u, \theta)}\, dP_\theta. \quad (3.9)$$

This gives us the previously discussed measure transformation between $P_\theta$ and $P_\Pi$. Now on $\bigcap_{\theta' \in B(\theta)} \bigcap_{u \in \mathcal{U}_\theta} A^u(\epsilon, \rho, \theta')$, where

$A^u(\epsilon, \rho, \theta')$ are as obtained in Lemma 3.1, we have

$$\sum_{\theta' \in B(\theta)} \Pi(\theta') \prod_{u \in \mathfrak{U}_\theta} \prod_{i=1}^{T_n(u)} \frac{p(X_i^u; u, \theta')}{p(X_i^u; u, \theta)}$$

$$= \sum_{\theta' \in B(\theta)} \Pi(\theta') \exp\left( - \sum_{u \in \mathfrak{U}_\theta} \sum_{i=1}^{T_n(u)} \log \frac{p(X_i^u; u, \theta)}{p(X_i^u; u, \theta')} \right)$$

$$\geq \sum_{\theta' \in B(\theta)} \Pi(\theta') \exp\left( -(1+\rho) \sum_{u \in \mathfrak{U}_\theta} T_n(u) I^u(\theta, \theta') \right)$$

$$\cdot \exp\left( - \sum_{u \in \mathfrak{U}_\theta} K^u(\epsilon, \rho, \theta') \right)$$

$$\geq \min_{\theta' \in B(\theta)} \left( \Pi(\theta') \exp\left( - \sum_{u \in \mathfrak{U}_\theta} K^u(\epsilon, \rho, \theta') \right) \right)$$

$$\cdot \exp\left( -(1+\rho) \min_{\theta' \in B(\theta)} \left( \sum_{u \in \mathfrak{U}_\theta} T_n(u) d_\theta(u) \right) \right.$$

$$\left. \cdot \frac{\sum_{u \in \mathfrak{U}_\theta} \frac{T_n(u)}{T_n} I^u(\theta, \theta')}{\sum_{u \in \mathfrak{U}_\theta} \frac{T_n(u)}{T_n} d_\theta(u)} \right)$$

$$\geq c(\epsilon, \rho) \cdot \exp\left( -(1+\rho) \left( \sum_{u \in \mathfrak{U}_\theta} T_n(u) d_\theta(u) \right) \right.$$

$$\left. \cdot \max_{\alpha \in \mathfrak{A}_\theta} \min_{\theta' \in B(\theta)} \frac{\sum_{u \in \mathfrak{U}_\theta} \alpha^u I^u(\theta, \theta')}{\min_{\theta' \in B(\theta)} \sum_{u \in \mathfrak{U}_\theta} \alpha^u d_\theta(u)} \right)$$

where

$$T_n = \sum_{u \in \mathfrak{U}_\theta} T_n(u) \text{ and}$$

$$c(\epsilon, \rho) = \min_{\theta' \in B(\theta)} \left( \Pi(\theta') \exp\left( - \sum_{u \in \mathfrak{U}_\theta} K^u(\epsilon, \rho, \theta') \right) \right). \quad (3.10)$$

Let

$$A_n := \left\{ \sum_{\mathfrak{U}_\theta} T_n(u) d_\theta(u) < \frac{\log n}{1+2\rho} \right.$$

$$\left. \cdot \frac{1}{\max_{\alpha \in \mathfrak{A}_\theta} \min_{\theta' \in B(\theta)} \frac{\sum_{\mathfrak{U}_\theta} \alpha^u I^u(\theta, \theta')}{\sum_{\mathfrak{U}_\theta} \alpha^u d_\theta(u)}} \right\}.$$

Using (3.8)–(3.10) it follows that

$$o(n^{a-1}) = \sum_{\theta' \in B(\theta)} \Pi(\theta') P_{\theta'}(A_n) = P_\Pi(A_n)$$

$$\geq P_\Pi\left( A_n \cap \left( \bigcap_{\theta' \in B(\theta)} \bigcap_{u \in \mathfrak{U}_\theta} A^u(\epsilon, \rho, \theta') \right) \right)$$

$$\geq c(\epsilon, \rho) n^{-(1+\rho)/(1+2\rho)}$$

$$\cdot P_\theta\left( A_n \cap \left( \bigcap_{\theta' \in B(\theta)} \bigcap_{u \in \mathfrak{U}_\theta} A^u(\epsilon, \rho, \theta') \right) \right). \quad (3.11)$$

Therefore,

$$\lim_{n \to \infty} P_\theta\left( A_n \cap \left( \bigcap_{\theta' \in B(\theta)} \bigcap_{u \in \mathfrak{U}_\theta} A^u(\epsilon, \rho, \theta') \right) \right) = 0. \quad (3.12)$$

Letting $\epsilon \to 0$ for fixed $\rho > 0$, we get (3.2)

$$\lim_{n \to \infty} P_\theta(A_n) = 0.$$

By using Markov's inequality and letting $\rho \to 0$, we get (3.3)

$$\liminf_{n \to \infty} \frac{L_n(\theta)}{\log n} \geq \min_{\alpha \in \mathfrak{A}_\theta} \max_{\theta' \in B(\theta)} \frac{\sum_{\mathfrak{U}_\theta} \alpha^u d_\theta(u)}{\sum_{\mathfrak{U}_\theta} \alpha^u I^u(\theta, \theta')}. \quad \square$$

Notice that for those $\theta \in \Theta$ for which $B(\theta)$ is empty, Theorem 3.1 does not provide a lower bound. This in fact allows us to construct uniformly good schemes which have a finite loss for those values of $\theta$. This will become clear in Section IV.

## IV. The Control Scheme

We first introduce some notation in Section IV-A. Then, in Section IV-B we describe in detail the adaptive control scheme, and in Section IV-C we give a heuristic interpretation of its key features. In Section IV-D we derive an upper bound on the loss associated with it. This bound is the same as the lower bound derived in Section III. Thus, the proposed control scheme is asymptotically efficient.

### *A. Notation*

In this section we introduce the notation used for the description of the control scheme we construct in Section IV-B.

Let $M^{(1)}$ be the unit simplex in $\mathbb{R}^{|\mathfrak{X}|}$ identified with the space of probability measures on $\mathfrak{X}$. Then $p(u, \theta) := \{p(x; u, \theta) : x \in \mathfrak{X}\} \in M^{(1)}$. Since $\Theta$ is finite, $p(u, \theta)$ takes on only a finite number of values for each $u \in \mathfrak{U}$. Therefore, for each value of $u \in \mathfrak{U}$, it is possible to find an $\epsilon^u > 0$ such that for all values of $p(u, \theta)$ we can identify $\epsilon$-neighborhoods

$$\epsilon\text{-nbd}(p(u, \theta)) := \{p \in M^{(1)} : \max_{x \in \mathfrak{X}} |p(x) - p(x; u, \theta)| < \epsilon^u\} \quad (4.1)$$

which are disjoint for distinct values of $p(u, \theta)$.

Also, define

$$S(\theta) := \{\theta' \in \Theta : p(u^*(\theta), \theta') = p(u^*(\theta), \theta) \text{ and } u^*(\theta') = u^*(\theta)\}. \quad (4.2)$$

This is the set of parameters for which the optimal control actions

are the same as that for $\theta$, and the distributions under the optimal control action are also identical. Let

$$\mathcal{U}(S(\theta)) := \{u : p(u, \theta') \neq p(u, \theta), \theta' \in S(\theta)\}. \quad (4.3)$$

Recall from Section III that

$$B(\theta) := \{\theta' \in \Theta : p(u^*(\theta), \theta') = p(u^*(\theta), \theta) \text{ and } u^*(\theta') \neq u^*(\theta)\}. \quad (4.4)$$

This is the set of parameters for which the optimal control actions are better than the optimal control action for $\theta$, and the distributions under the optimal control action for $\theta$ are identical. Also let

$$\alpha(\theta) = \{\alpha^u(\theta) : u \in \mathcal{U}_\theta\} \quad (4.5)$$

achieve the minimum in the lower bound for the loss in (3.2), where $\mathcal{U}_\theta = \mathcal{U} - \{u^*(\theta)\}$. Let $k = \#\mathcal{U}$.

### B. Description of the Control Scheme

To start off, use each of the control actions $u \in \mathcal{U}$ once each. From then on at each stage $n$, compute the empirical measure

$$q_n(u) := \{q_n(u)(x); x \in \mathcal{X}\} \in M^{(1)}$$

corresponding to each $u \in \mathcal{U}$ as

$$q_n(u)(x) := \frac{1}{T_n(u)} \sum_{i=1}^{n} 1(\phi_i = u, X_i = x). \quad (4.6)$$

Define the following conditions.

*C1(θ):* $q_n(u) \in \epsilon$-nbd $(p(u, \theta))$ $\forall\, u \in \mathcal{U}$ and $B(\theta)$ is empty.

*C2(θ):* $q_n(u) \in \epsilon$-nbd $(p(u, \theta))$ $\forall\, u \in \mathcal{U}$ and $B(\theta)$ is nonempty.

*C3:* There does not exist $\theta \in \Theta$ such that $q_n(u) \in \epsilon$-nbd $(p(u, \theta))$ $\forall\, u \in \mathcal{U}$.

(Note that $C3 = (\cup_{\theta \in \Theta}(C1(\theta) \cup C2(\theta)))'$.) Proceed as follows.

1) If $C1(\theta)$ is satisfied for some $\theta \in \Theta$, then use $u^*(\theta)$.

2) If $C2(\theta)$ is satisfied for some $\theta \in \Theta$, then do the following: maintain a count of the number of times condition $C2(\theta)$ is satisfied. Of these, for the first time choose among the control actions $u \in \mathcal{U}_\theta$ randomly with probabilities $\alpha^u(\theta)$. Refer to this process as randomization. For those instants when this count is even (say $C2(\theta)a$), use $u^*(\theta)$. For other instants when the count is odd (call this situation $C2(\theta)b$), compute the likelihood ratio

$$\Lambda_n(\theta) := \lambda_{T_n^r}(\theta) := \min_{\theta' \in B(\theta)} \prod_{t=1}^{T_n^r} \frac{\alpha^{U_t^r}(\theta) p(X_t^r; U_t^r, \theta)}{\alpha^{U_t^r}(\theta) p(X_t^r; U_t^r, \theta')}$$

$$= \min_{\theta' \in B(\theta)} \prod_{t=1}^{T_n^r} \frac{p(X_t^r; U_t^r, \theta)}{p(X_t^r; U_t^r, \theta')} \quad (4.7)$$

of $\theta$ versus $B(\theta)$, where $(U_1^r, X_1^r), \cdots, (U_{T_n^r}^r, X_{T_n^r}^r)$ is the sequence of pairs of control actions and observed states up to stage $n$ when randomization is done with $\alpha(\theta)$. If $C2(\theta)b1$:$\Lambda_n > K_{n+1}$, where $K_n = n(\log n)^p$ for some fixed $p > 1$, then use $u^*(\theta)$. If $C2(\theta)b2$:$\Lambda_n \leq K_{n+1}$ then do the following. Maintain a count of the number of times this condition ($C2(\theta)b2$) is satisfied. If this count is a perfect square (say $C2(\theta)b2a$), then use round robin among $u \in \mathcal{U}(S(\theta))$. If this count is not a perfect square (say $C2(\theta)b2b$), then do randomization using $\alpha(\theta)$.

3) If $C3$ is satisfied, then use round-robin among $u \in \mathcal{U}$.

### C. Discussion of the Control Scheme

In this section we give a heuristic interpretation of the control scheme constructed in Section IV-B and some of the key underlying ideas.

The entire control scheme is based on an identification strategy that makes use of empirical measures. From the theory of large deviations it follows that the empirical measure of an i.i.d. process should converge to any $\epsilon$-neighborhood of its true distributions in finite time. Using this idea, at each stage we compute the empirical measure corresponding to states observed when the various control actions are used, and identify three types of situations $C1(\theta)$, $C2(\theta)$, and $C3$ which need to be treated differently. Whenever condition $C1(\theta)$ is satisfied there is no conflict between learning and control, in the sense that the apparent best action is also appropriate for information gathering. This condition is akin to the closed-loop identifiability condition. However, when condition $C2(\theta)$ is satisfied there is a conflict between learning and control. The apparent best action in this situation does not probe the system adequately, and consequently we need to repeatedly use forcing controls to learn about the system. The amount of information already available is quantified by means of certain likelihood ratios, and by comparing these to an appropriately time-varying threshold, we decide: 1) we have sufficient information and we can go ahead and use the apparent best control action; or 2) we have insufficient information and we need to use one of the forcing controls to learn more about the system. This scheme quantifies the available and required information very precisely, and uses the various controls in proportion that is least expensive and most effective for learning, and is consequently able to achieve the optimal rate of learning. Finally, when condition $C3$ is true, then we clearly have insufficient information about the system and we need to concern ourselves with identification alone.

### D. Upper Bound on the Loss

In this section we derive an upper bound on the loss associated with the adaptive control scheme $\phi^*$ constructed in Section IV-B. The bound is given by the main Theorem 4.2. Lemmas 4.1, 4.2, 4.3, and Theorem 4.1 are needed for the proof of the main theorem.

*Lemma 4.1:* Let $X_1, X_2, \cdots$, be i.i.d. r.v.'s on the finite state-space $\mathcal{X}$, with marginal $p(x)$. Let $M^{(1)}$ be the unit simplex in $\mathbb{R}^{|\mathcal{X}|}$ identified with the space of probability measures on $\mathcal{X}$, and let $K \subset M^{(1)}$, closed, such that $p \notin K$. Let

$$q_n := \{q_n(x) | x \in \mathcal{X}\} \text{ where } q_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1(X_i = x).$$

Then

i) $P(q_n \in K) < Ae^{-an}$ for all $n \geq 1$

for some positive constants $A$, $a$.

Let $L := \sup\{n \geq 1 | q_n \in K\}$. Then

ii) $EL < \infty$.

*Proof:* i) follows from the theory of large deviations (see [7, Theorem II.4.3]).

To prove ii) note that

$$EL = E \sum_{n=1}^{\infty} 1(\exists i \geq n, q_i \in K)$$

$$= E \sum_{n=1}^{\infty} 1\left(\bigcup_{i \geq n} (q_i \in K)\right)$$

$$\leq \sum_{n=1}^{\infty} \sum_{i=n}^{\infty} P(q_i \in K)$$

$$\leq \sum_{n=1}^{\infty} \sum_{i=n}^{\infty} Ae^{-ai}$$

$$< \infty. \qquad \square$$

*Lemma 4.2:* Let $X_1, X_2, \cdots$, be i.i.d. r.v.'s on the finite state-space $\mathfrak{X}$ with marginals $p(x)$. Let $f:\mathfrak{X} \to \mathbb{R}$ be such that $Ef(X_i) = \Sigma_{x\in\mathfrak{X}} p(x)f(x) > 0$. For $n \geq 1$, let $S_n := \Sigma_{i=1}^n f(X_i)$.
Then

i) $P(S_n \leq 0) < Ae^{-an}$ for all $n \geq 1$

for some positive constants $A$, $a$.

Let $L := \sup\{n \geq 1 | S_n \leq 0\}$. Then

ii) $EL < \infty$.

*Proof:* Let $K = \{q \in M^{(1)} | \Sigma_{x\in\mathfrak{X}} f(x)q(x) \leq 0\}$. Then $K \subset M^{(1)}$ is closed and $p \notin K$. Also since

$$S_n = \sum_{x\in\mathfrak{X}} nq_n(x)f(x)$$

$$\{S_n \leq 0\} = \left\{n \sum_{x\in\mathfrak{X}} q_n(x)f(x) \leq 0\right\}$$

$$= \left\{\sum_{x\in\mathfrak{X}} q_n(x)f(x) \leq 0\right\}$$

$$= \{q_n \in K\}. \tag{4.8}$$

The proof of Lemma 4.2 follows from (4.8) and Lemma 4.1.

*Lemma 4.3:* Let $X_1, X_2, \cdots$, be i.i.d. r.v.'s on the finite state-space $\mathfrak{X}$, with marginals $p(x)$. Let $f^i:\mathfrak{X} \to \mathbb{R}$ be such that $0 < Ef^i(X_1) < \infty$, $i \in I$, finite. Let $S_n^i = f^i(X_1) + f^i(X_2) + \cdots + f^i(X_n)$, $L_A^i = \Sigma_{n=1}^\infty 1(\inf_{t\geq n} S_t^i \leq A)$, and $L_A = \max_{i\in I} L_A^i$. Then

$$\limsup_{A\to\infty} \frac{EL_A}{A} \leq \frac{1}{\min_{i\in I}(Ef^i(X_1))}. \tag{4.9}$$

*Proof:* For $\epsilon > 0$, and for any fixed $i \in I$

$$L_A^i \leq \frac{A(1+\epsilon)}{Ef^i(X_1)} + L^i \tag{4.10}$$

where

$$L^i = \sum_{n=1}^\infty 1\left(\inf_{t\geq n}\left(S_t^i - \frac{tEf^i(X_1)}{1+\epsilon}\right) \leq 0\right). \tag{4.11}$$

Consider the i.i.d. r.v.'s $X_1, X_2, \cdots$, and $f:\mathfrak{X} \to \mathbb{R}$ given by

$$f(X_t) = f^i(X_t) - \frac{Ef^i(X_t)}{1+\epsilon}.$$

Then

$$Ef(X_t) = Ef^i(X_t) - \frac{Ef^i(X_t)}{1+\epsilon} > 0.$$

Then, by Lemma 4.2 it follows that $EL^i < \infty$.
Therefore,

$$E(\max_{i\in I} L^i) \leq E\left(\sum_{i\in I} L^i\right) = \sum_{i\in I} EL^i = k(\epsilon) < \infty \tag{4.12}$$

for some constant $k(\epsilon)$ independent of $A$.
Now,

$$L_A = \max_{i\in I} L_A^i \leq \max_{i\in I}\left(\frac{A(1+\epsilon)}{Ef^i(X_1)} + L_i\right)$$

$$\leq \frac{A(1+\epsilon)}{\min_{i\in I}(Ef^i(X_1))} + \max_{i\in I} L^i. \tag{4.13}$$

By (4.12) and (4.13) it follows that:

$$EL_A \leq \frac{A(1+\epsilon)}{\min_{i\in I}(Ef^i(X_1))} + k(\epsilon)$$

$$\limsup_{A\to\infty} \frac{EL_A}{A} \leq \frac{1+\epsilon}{\min_{i\in I}(Ef^i(X_1))}.$$

By letting $\epsilon \to 0$, we get the desired result. □

*Theorem 4.1:* Let $\theta \in \Theta$ be such that $B(\theta)$ is nonempty. Then

1) $$\limsup_{n\to\infty} [E_\theta [\sup\{1 \leq i \leq n | \lambda_i(\theta) \leq K_{n+1}\}]/\log n]$$

$$\leq \frac{1}{\min_{\theta'\in B(\theta)} \sum_{\mathfrak{U}_\theta} \alpha^u(\theta) I^u(\theta, \theta')}. \tag{4.14}$$

2) $$P_{\theta'}\{\lambda_i(\theta) > K_{n+1} \text{ for some } i \leq n\} \leq 1/K_{n+1}$$
for $\theta' \in B(\theta)$ (4.15)

where

$$\lambda_i(\theta) := \min_{\theta'\in B(\theta)} \prod_{t=1}^i \frac{\alpha^{U_t^r}(\theta)p(X_t^r; U_t^r, \theta)}{\alpha^{U_t^r}(\theta)p(X_t^r; U_t^r, \theta')}$$

$$= \min_{\theta'\in B(\theta)} \prod_{t=1}^i \frac{p(X_t^r, U_t^r, \theta)}{p(X_t^r, U_t^r, \theta')}$$

$(U_t^r, X_t^r)$ is the pair of control action and observed state at the randomizing instants (note that $(U_1^r, X_1^r)$ $(U_2^r, X_2^r)$, $\cdots$, are i.i.d. with marginal $\alpha^u(\theta)p(x; u, \theta')$ under $\theta'$) and

$$K_n = n(\log n)^p \qquad \text{for some } p > 1.$$

*Proof:*

$$\sup\{1 \leq i \leq n | \lambda_i(\theta) \leq K_{n+1}\}$$

$$= \max_{\theta'\in B(\theta)} \sup\left\{1 \leq i \leq n \left| \prod_{t=1}^i \frac{p(X_t^r; U_t^r, \theta)}{p(X_t^r; U_t^r, \theta')} \leq K_{n+1}\right.\right\}$$

$$= \max_{\theta'\in B(\theta)} \sum_{i=1}^\infty 1\left(\inf_{n\geq i} \sum_{t=1}^n \log \frac{p(X_t^r; U_t^r, \theta)}{p(X_t^r; U_t^r, \theta')} \leq \log K_{n+1}\right). \tag{4.16}$$

We can now use Lemma 4.3 by making the following translation.
Let

$$\mathfrak{X} = \mathfrak{U} \times \mathfrak{X}$$

$$X_t = (U_t^r, X_t^r)$$

$$I = B(\theta)$$

$$f^{\theta'}(X_t) = \log \frac{p(X_t^r, U_t^r, \theta)}{p(X_t^r; U_t^r, \theta')}, \quad \theta' \in B(\theta)$$

$$S_n^{\theta'} := \sum_{t=1}^n f^{\theta'}(X_t) = \sum_{t=1}^n \log \frac{p(X_t^r; U_t^r, \theta)}{p(X_t^r; U_t^r, \theta')}$$

$$A = \log K_{n+1}$$

$$L_A^{\theta'} = \sum_{i=1}^{\infty} 1\left(\inf_{n\geq i} S_n^{\theta'} \leq A\right)$$

$$= \sum_{i=1}^{\infty} 1\left(\inf_{n\geq i} \sum_{t=1}^{n} \log \frac{p(X_t^r; U_t^r, \theta)}{p(X_t^r; U_t^r, \theta')} \leq \log K_{n+1}\right)$$

$$L_A = \max_{i\in I} L_A^i$$

$$= \max_{\theta'\in B(\theta)} \sum_{i=1}^{\infty} 1\left(\inf_{n\geq i} \sum_{t=1}^{n} \log \frac{p(X_t^r; U_t^r, \theta)}{p(X_t^r; U_t^r, \theta')} \leq \log K_{n+1}\right). \tag{4.17}$$

Note that $E_\theta(f^{\theta'}(X_1)) = \Sigma_{\mathcal{U}_\theta} \alpha^u(\theta) I^u(\theta, \theta') > 0$.

Hence, (4.14) follows by a straightforward application of Lemma 4.3 to (4.16) along with the translation (4.17).

To prove (4.15), note that

$$\{\lambda_i(\theta) > K_{n+1} \text{ for some } i \leq n\}$$

$$\subseteq \left\{\prod_{t=1}^{i} \frac{p(X_t^r; U_t^r, \theta)}{p(X_t^r; U_t^r, \theta')} > K_{n+1} \text{ for some } i \leq n\right\}$$

for any $\theta' \in B(\theta)$

where

$$\left\{\prod_{t=1}^{i} \frac{p(X_t^r; U_t^r, \theta)}{p(X_t^r; U_t^r, \theta')}\right\}_{i=1}^{\infty}$$

is a martingale under $\theta'$. Therefore, (4.15) follows by the martingale inequality (cf. [6, p. 243]). □

Lemmas 4.1–4.3 and Theorem 4.1 are now used to characterize the performance of the proposed adaptive control scheme.

*Theorem 4.2:* Under the adaptive control scheme $\phi^*$, for $u \neq u^*(\theta)$

$$\text{i) } E_\theta T_n(u) \leq \left(\frac{\alpha^u(\theta)}{\min_{\theta'\in B(\theta)} \sum_{\mathcal{U}_\theta} \alpha^u(\theta) I^u(\theta, \theta')} + o(1)\right) \log n$$

if $B(\theta)$ is nonempty

$$E_\theta T_n(u) \leq M < \infty \quad \text{if } B(\theta) \text{ is empty.} \tag{4.18}$$

Consequently,

$$\text{ii) } L_n(\theta) \leq \left(\max_{\theta'\in B(\theta)} \frac{\sum_{\mathcal{U}_\theta} \alpha_\theta^u d_\theta(u)}{\sum_{\mathcal{U}_\theta} \alpha^u(\theta) I^u(\theta, \theta')} + o(1)\right) \log n$$

if $B(\theta)$ is nonempty

$$L_n(\theta) \leq M < \infty \quad \text{if } B(\theta) \text{ is empty} \tag{4.19}$$

where $\alpha(\theta) = \{\alpha^u(\theta): u \in \mathcal{U}_\theta\}$ is defined by (4.5).

*Proof:* For $u \neq u^*(\theta)$, we have

$$T_n(u) = \sum_{i=1}^{n} 1(U_i = u)$$

$$= 1 + \sum_{i=k+1}^{n} 1(U_i = u)$$

$$= 1 + \sum_{i=k+1}^{n} 1\{U_i = u, C1(\theta') \text{ is satisfied at stage } i \text{ for some } \theta' \in \Theta\}$$

$$+ \sum_{i=k+1}^{n} 1\{U_i = u, C2(\theta') \text{ is satisfied at stage } i \text{ for some } \theta' \in \Theta\}$$

$$+ \sum_{i=k+1}^{n} 1\{U_i = u, C3 \text{ is satisfied at stage } i\}$$

$$= 1 + \text{Term } 1 + \text{Term } 2 + \text{Term } 3 \text{ (say)} \tag{4.20}$$

where $C1(\theta')$, $C2(\theta')$, and $C3$ are defined in Section IV-B. Let us now examine each term separately. Defining $\mathfrak{I}_u$ by

$$\mathfrak{I}_u := \sup_{T_n(u)\geq 1} \{q_n(u) \notin \epsilon\text{-nbd}(p(u, \theta))\} \tag{4.21}$$

and noting that $E_\theta \mathfrak{I}_u < \infty$ by Lemma 4.1 ii), we get Term 3 $\leq \Sigma_{u\in\mathcal{U}} \mathfrak{I}_u$, thus

$$E_\theta \text{ Term } 3 \leq \sum_{u\in\mathcal{U}} E_\theta \mathfrak{I}_u < \infty \tag{4.22}$$

and Term 1 $\leq \mathfrak{I}_u$, thus,

$$E_\theta \text{ Term } 1 \leq E_\theta \mathfrak{I}_u < \infty. \tag{4.23}$$

$$\text{Term } 2 = \sum_{i=k+1}^{n} 1\{U_i = u, C2(\theta') \text{ is satisfied at stage } i \text{ for some } \theta' \in \Theta \text{ such that } p(u^*(\theta'), \theta') \neq p(u^*(\theta'), \theta)\}$$

$$+ \sum_{i=k+1}^{n} 1\{U_i = u, C2(\theta') \text{ is satisfied at stage } i \text{ for some } \theta' \in \Theta \text{ such that } \theta \in B(\theta')\}$$

$$+ \sum_{i=k+1}^{n} 1\{U_i = u, C2(\theta') \text{ is satisfied at stage } i \text{ for some } \theta' \in \Theta \text{ such that } \theta \in S(\theta')\}$$

$$+ \sum_{i=k+1}^{n} 1\{U_i = u, C2(\theta) \text{ is satisfied at stage } i\}.$$

$$= \text{Term } 2a + \text{Term } 2b + \text{Term } 2c + \text{Term } 2d \text{ (say).} \tag{4.24}$$

Next we upperbound Terms 2a–2d separately.

$$\text{Term } 2a = \sum_{\substack{\theta':B(\theta') \text{ is not empty and} \\ p(u^*(\theta'),\theta') \neq p(u^*(\theta'),\theta)}} \sum_{i=k+1}^{n} 1\{U_i = u,\ C2(\theta') \text{ is satisfied at stage } i\}$$

$$\leq \sum_{\substack{\theta':B(\theta') \text{ is not empty and} \\ p(u^*(\theta'),\theta') \neq p(u^*(\theta'),\theta)}} \left[1 + \sum_{i=k+1}^{n} 1\{U_i = u^*(\theta'),\ C2(\theta') \text{ is satisfied at stage } i\}\right]$$

$$\leq \sum_{\substack{\theta':B(\theta') \text{ is not empty and} \\ p(u^*(\theta'),\theta') \neq p(u^*(\theta'),\theta)}} (\mathfrak{I}_{u^*(\theta')} + 1). \tag{4.25}$$

The first of the inequalities of (4.25) holds because under $C2(\theta')$, $u^*(\theta')$ is used on all the even times, therefore, at least as many times as any other control minus one. The second of the inequalities of (4.25) holds because the sum on the left-hand side counts a subset of the time instants when $u^*(\theta')$ is used and $q_n(u^*(\theta')) \notin \epsilon\text{-nbd}\ (p(u^*(\theta'), \theta))$ where $\theta$ is the true parameter.

By Lemma 4.1 ii), it follows that

$$E_\theta \text{ Term } 2a \leq \sum_{\substack{\theta':B(\theta') \text{ is not empty and} \\ p(u^*(\theta'),\theta') \neq p(u^*(\theta'),\theta)}} (1 + E_\theta \mathfrak{I}_{u^*(\theta')}) < \infty. \tag{4.26}$$

$$\text{Term } 2b \leq \sum_{\theta':\theta \in B(\theta')} \{C2(\theta') \text{ is satisfied at stage } i\}$$

$$\leq \sum_{\theta':\theta \in B(\theta')} 2\left[1 + \sum_{i=k+1}^{n} 1\{C2(\theta')b \text{ is satisfied at stage } i\}\right]$$

$$= \sum_{\theta':\theta \in B(\theta')} 2\left[1 + \sum_{i=k+1}^{n} 1\{C2(\theta')b1 \text{ is satisfied at stage } i\} + \sum_{i=k+1}^{n} 1\{C2(\theta')b2 \text{ is satisfied at stage } i\}\right]$$

$$\leq \sum_{\theta':\theta \in B(\theta')} 2\left[1 + \sum_{i=k+1}^{n} 1\{\Lambda_{i-1}(\theta') > K_i\} + \sum_{i=k+1}^{n} 1\{C2(\theta')b2 \text{ is satisfied at stage } i\}\right]$$

$$\leq \sum_{\theta':\theta \in B(\theta')} 2\left[1 + \sum_{i=k+1}^{n} 1\{\lambda_j(\theta') > K_i \text{ for some } j \leq i-1\} + \sum_{i=k+1}^{n} 1\{C2(\theta')b2 \text{ is satisfied at stage } i\}\right]. \tag{4.27}$$

The first of the inequalities of (4.27) results by removing the condition $U_i = u$. The second one results by observing that the total number of time instants that $C2(\theta')$ is satisfied is upperbounded by twice the odd instants that $C2(\theta')$ holds, and by noting that the first time we randomize and the other odd times we call $C2(\theta')b$. The third inequality results because $\{C2(\theta')b2$ is satisfied at stage $i\}$ implies $\{\Lambda_{i-1}(\theta') > K_i\}$.

Consider now the term $\Sigma_{i=k+1}^{n} 1\{C2(\theta')b2$ is satisfied at stage $i\}$.

$$\sum_{i=k+1}^{n} 1\{C2(\theta')b2 \text{ is satisfied at stage } i\}$$

$$= \sum_{i=k+1}^{n} 1\{C2(\theta')b2a \text{ is satisfied at stage } i\} + \sum_{i=k+1}^{n} 1\{C2(\theta')b2b \text{ is satisfied at stage } i\}$$

$$\leq 1 + 2\sum_{i=k+1}^{n} 1\{C2(\theta')b2b \text{ is satisfied at stage } i\}$$

$$= 1 + 2\sum_{i=k+1}^{n} 1\{C2(\theta')b2b \text{ is satisfied at stage } i;$$

of the number of times that $C2(\theta')b2b$ has been satisfied so far, the fraction of times that $u'$ is used $\in (\alpha^{u'}(\theta') - \epsilon, \alpha^{u'}(\theta') + \epsilon)\}$

$$+ 2\sum_{i=k+1}^{n} 1\{C2(\theta')b2b \text{ is satisfied at stage } i;$$

of the number of times that $C2(\theta')b2b$ has been satisfied so far, the fraction of times that $u'$ is used $\notin (\alpha^{u'}(\theta') - \epsilon, \alpha^{u'}(\theta') + \epsilon)\}$

$$\leq 1 + 2\sum_{j=1}^{\infty} 1\{q_i(u') \notin \epsilon\text{-nbd}(p(u', \theta)) \text{ for some } i > (\alpha^{u'}(\theta') - \epsilon)j\}$$

$$+ 2\sum_{j=1}^{\infty} 1\{\text{of } j \text{ the fraction of times } u' \text{ is used} \notin (\alpha^{u'}(\theta') - \epsilon, \alpha^{u'}(\theta') + \epsilon)\} \tag{4.28}$$

where $u' \in \mathfrak{U}_{\theta'}$ is such that $p(u', \theta) \neq p(u', \theta')$.

The first of the inequalities of (4.28) results by observing that the number of times condition $C2(\theta')b2a$ is satisfied (i.e., the count of the number of times $C2(\theta')b2$ is satisfied is a perfect square) is upperbounded by the number of times condition $C2(\theta')b2b$ is satisfied plus one. Consider now changing the index of summation to the time instants when randomization is done. Then the condition $C2(\theta')b2b$, along with the condition that the fraction of times that $u'$ is used $\in (\alpha^{u'}(\theta') - \epsilon, \alpha^{u'}(\theta') + \epsilon)$ at stage $i$, imply that $q_i(u') \notin \epsilon\text{-nbd}\ (p(u'), \theta))$ for some $i > (\alpha^{u'}(\theta') - \epsilon)j$. By extending the summation to infinity together with the above observation establishes the last of the inequalities of (4.28).

Thus, by Lemma 4.1 i) and (4.15) it follows that

$$E_\theta \text{ Term } 2b \leq \sum_{\theta':\theta \in B(\theta')} 2\left[1 + \sum_{i=k+1}^{n} (i(\log i)^p)^{-1} + 1 + 2\sum_{j=1}^{\infty} \sum_{i > (\alpha^{u'(\theta')} - \epsilon)j} A_1 e^{-a_1 i} + 2\sum_{j=1}^{\infty} A_2 e^{-a_2 j}\right] < \infty \tag{4.29}$$

where $A_1$, $a_1$, $A_2$, $a_2 > 0$ are some constants.

$$\text{Term } 2c = \sum_{\theta':\theta\in S(\theta')} \sum_{i=k+1}^{n} \{U_i = u,\ C2(\theta') \text{ is satisfied at stage } i\}$$

$$\le \sum_{\theta':\theta\in S(\theta')} \left[1 + \sum_{i=k+1}^{n} 1\{U_i = u,\ C2(\theta')b2 \text{ is satisfied at stage } i\}\right]$$

$$\le \sum_{\theta':\theta\in S(\theta')} \left[1 + \sum_{i=k+1}^{n} 1\{C2(\theta')b2 \text{ is satisfied at stage } i\}\right]$$

$$\le \sum_{\theta':\theta\in S(\theta')} \left[1 + l^2 + \sum_{j=1}^{\infty} 1\{q_j(u') \notin \epsilon\text{-nbd}\,(p(u', \theta))\} \cdot (2j+1)l^2\right] \tag{4.30}$$

where $u' \in \mathcal{U}(S(\theta'))$ is such that $p(u', \theta) \neq p(u', \theta')$ and $\#\mathcal{U}(S(\theta')) = l$.

The first inequality of (4.30) results by noting that since $\theta \in S(\theta')$, $u \neq u^*(\theta') = u^*(\theta)$ can be used only when condition $C2(\theta')b2$ is satisfied, or at the first instant when $C2(\theta')$ is true. The second inequality results by removing the requirement $U_i = u$. The third inequality results by upperbounding the number of times condition $C2(\theta')b2$ is satisfied. This can be achieved as follows. First restrict attention to those time instants that are perfect squares and the control $u'$ is used. At these time instants since $C2(\theta')$ is satisfied $q_n(u') \in \epsilon\text{-nbd}\,(p(u', \theta'))$, thus, by the choice of $u' \in \mathcal{U}(S(\theta'))$, $q_n(u') \notin \epsilon\text{-nbd}\,(p(u', \theta))$. Consider the sum of the intervals between the above time instants. (Note that the length of the $j$th interval is upperbounded by $[(j+1)^2 - j^2]l^2 = (2j+1)l^2$.) Then the number of times condition $C2(\theta')b2$ is satisfied cannot exceed this sum. Finally, the inequality results by changing the summation index to all the times when $u'$ is used and upperbounding the interval following the time $q_j(u') \notin \epsilon\text{-nbd}\,(p(u', \theta))$ by $(2j+1)l^2$.

Again, by using Lemma 4.1 i) we get

$$E_\theta \text{ Term } 2c \le \sum_{\theta':\theta\in S(\theta')} \left[1 + l^2 + \sum_{j=1}^{\infty} Ae^{-aj} \cdot (2j+1)l^2\right] < \infty. \tag{4.31}$$

Now if $B(\theta)$ is empty then,

$$\text{Term } 2d = 0. \tag{4.32}$$

Otherwise,

$$\text{Term } 2d = \sum_{i=k+1}^{n} 1\{U_i = u,\ C2(\theta) \text{ is satisfied at stage } i\}$$

$$\le 1 + \sum_{i=k+1}^{n} 1\{U_i = u,\ C2(\theta)b2 \text{ is satisfied at stage } i\}$$

$$= 1 + \sum_{i=k+1}^{n} 1\{U_i = u,\ C2(\theta)b2a \text{ is satisfied at stage } i\}$$

$$+ \sum_{i=k+1}^{n} 1\{U_i = u,\ C2(\theta)b2b \text{ is satisfied at stage } i\}$$

$$\le 2 + \sum_{i=k+1}^{n} 1\{U_i = u,\ C2(\theta)b2b \text{ is satisfied at stage } i\}$$

$$+ \left(\sum_{i=k+1}^{n} 1\{C2(\theta)b2b \text{ is satisfied at stage } i\}\right)^{1/2}. \tag{4.33}$$

The first of the inequalities of (4.33) is obtained by noting that $u \neq u^*(\theta)$ can be used only at the first instant when $C2(\theta)$ is satisfied (in which case randomization is done) or when $C2(\theta)b2$ is satisfied. The last of the inequalities of (4.33) results because the number of times condition $C2(\theta)b2a$ is satisfied is upperbounded by one plus the square root of the number of times $C2(\theta)b2b$ is satisfied.

To upperbound $E_\theta$ Term $2d$ we use (4.33), Jensen's inequality, and the following fact. At each stage $i$ when condition $C2(\theta)b2b$ is satisfied, the choice of the control action $U_i \in \mathcal{U}_\theta$ is made by an independent randomization $\alpha(\theta)$. Then,

$$E_\theta \text{ Term } 2d \le 2 + \sum_{i=k+1}^{n} P_\theta\{C2(\theta)b2b \text{ is satisfied at stage } i\} \cdot \alpha^u(\theta) + \left(\sum_{i=k+1}^{n} P_\theta\{C2(\theta)b2b \text{ is satisfied at stage } i\}\right)^{1/2}$$

$$\le 2 + \alpha^u(\theta) E_\theta\,[\sup\{1 \le i \le n \mid \lambda_i(\theta) \le K_{n+1}\}] + (E_\theta\,[\sup\{1 \le k \le n \mid \lambda_i(\theta) \le K_{n+1}\}])^{1/2}. \tag{4.34}$$

Using (4.14) we get

$$\limsup_{n\to\infty} E_\theta \text{ Term } 2d/\log n \le \frac{\alpha^u(\theta)}{\min_{\theta'\in B(\theta)} \sum_{\mathcal{U}_\theta} \alpha^u(\theta) I^u(\theta, \theta')}. \tag{4.35}$$

Combining (4.20), (4.22), (4.23), (4.24), (4.26), (4.29), (4.31), (4.32), and (4.35) we get (4.18). Equation (4.19) follows easily from (4.18) and (2.3). □

## V. Conclusion

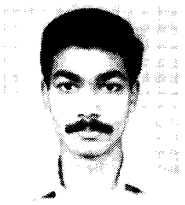
In this paper we considered the problem of adaptive control of i.i.d. processes. The optimality criterion we used, namely minimizing the rate at which the loss increases is stronger than the average reward per unit time criterion. We showed that this problem can be viewed as a multiarmed bandit problem like the one considered in [2]. However, the parametrization of arms is not independent. This difference is reflected in the lower bound on the loss we obtain in Section III, and also needs to be kept in mind when designing an optimal scheme like the one of Section IV. The control scheme presented in Section IV has an intuitively appealing structure as it clearly specifies the conditions under which there is either only identification, or only control, or identification and control, and treats each one of these conditions optimally.

## References

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

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

[3] T. L. Lai and H. Robbins, "Asymptotically optimal allocation of treatments in sequential experiments," in *Design of Experiments,* T. J. Santuer and A. C. Tamhane, Eds. New York: Marcel-Dekker.
[4] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays: Part I—IID rewards; Part II—Markovian rewards," *IEEE Trans. Automat. Contr.,* vol. AC-32, pp. 968–982, Nov. 1987.
[5] R. Agrawal, M. Hegde, and D. Teneketzis, "Asymptotically efficient adaptive allocation rules for the multi-armed bandit problem with switching cost," *IEEE Trans. Automat. Contr.,* vol. 33, pp. 899–906, Oct. 1988.
[6] S. Ross, *Stochastic Processes.* New York: Wiley, 1983.
[7] R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics.* New York: Springer-Verlag, 1985.
[8] J. A. Bather, "The minimax risk for the two-armed bandit problem," in *Mathematical Learning Models—Theory and Algorithms* (Springer-Verlag Lecture Notes in Statistics, Vol. 20). New York: Springer-Verlag, 1983, pp. 1–11.
[9] W. Vogel, "An asymptotic minimax theorem for the two-armed bandit problem," *Ann. Math. Statist.,* vol. 31, pp. 444–451, 1960.

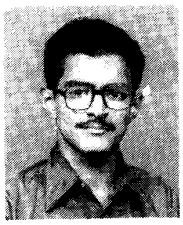


**Rajeev Agrawal** was born in Lucknow, India, on December 1, 1963. He received the B.Tech degree in electrical engineering from the Indian Institute of Technology, Kanpur, in 1985, and the M.S. and Ph.D. degrees in electrical engineering-systems from the University of Michigan, Ann Arbor, in 1987 and 1988, respectively.

In September 1988 he joined the Department of Electrical and Computer Engineering at the University of Wisconsin, where he is currently an Assistant Professor. His current research interests are in stochastic systems, stochastic adaptive control, resource allocation problems, stochastic scheduling, communication networks, and communication systems.

**Demosthenis Teneketzis** received the B.S. degree in electrical engineering from the University of Patras, Patras, Greece, in 1974, and the M.S., E.E., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1976, 1977, and 1979, respectively.

From 1979 to 1980 he worked for Systems Control Inc., Palo Alto, CA, and from 1980 to 1984 he was with Alphatech Inc., Burlington, MA. Since September 1984 he has been with the University of Michigan, Ann Arbor, where he is presently an Associate Professor of Electrical Engineering and Computer Science. His current research interests include stochastic systems and control, team theory, game theory, decentralized systems, information theory and queueing networks.



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

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

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