Proceedings of the 27th Conference
on Decision and Control
Austin, Texas • December 1988

ASYMPTOTICALLY EFFICIENT ADAPTIVE ALLOCATION SCHEMES FOR CONTROLLED MARKOV CHAINS: FINITE PARAMETER SPACE

Rajeev Agrawal, Demosthenis Teneketzis

Department of Electrical Engineering and Computer Science
and
Communications and Signal Processing Laboratory
University of Michigan
Ann Arbor, MI 48109-2122

and

Venkatachalam Anantharam

School of Electrical Engineering
Cornell University
Ithaca, NY 14853

Abstract

We consider a controlled Markov chain whose transition probabilities and initial distribution are parametrized by an unknown parameter θ belonging to some known parameter space Θ. There is 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 optimal.

1. Introduction

Consider the following stochastic adaptive control problem: The system is modelled by a controlled Markov chain with an unknown parameter, i.e.

$$\mathcal{P}_\theta\{X_{n+1} = y | X_n = x, X_{n-1}, \ldots, X_0, U_n, \ldots, U_0\} = P(x, y; U_n, \theta) \tag{1.1}$$

where $X_0, U_0, X_1, U_1, \ldots, X_n, U_n, X_{n+1}, \ldots$ is the chronological sequence of states and control actions, and θ is an unknown parameter belonging to some known parameter space Θ; and

$$\mathcal{P}_\theta(X_0 = x) = p(x; \theta) \tag{1.2}$$

where θ is the same as in (1.1). There is a one-step reward $r(X_n, U_n)$, associated with each pair $(X_n, U_n), n \geq 0$. The objective is to find an adaptive control scheme which maximizes, in some sense, the expected value of the sum of one-step rewards

$$E_\theta J_n = E_\theta \sum_{i=0}^{n-1} r(X_n, U_n), \text{ as } n \to \infty \ . \tag{1.3}$$

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 and is therefore also optimal for an average reward per unit time criterion (cf [1]). The reward criterion described by (1.3) 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 used in [1] - [7]. For the criterion (1.3) it is no longer clear that the Certainty Equivalent Control with Forcing is optimal.

The same reward criterion as (1.3) was previously used in [8] for the study of the controlled i.i.d. process problem. This criterion was initially used by Lai and Robbins [9], [10] for the multi-armed bandit problem. Various extensions of the Lai and Robbins formulation of the multi-armed bandit problems have been reported in [11] and [12]. In this paper we show that the adaptive control problem of Markov chains can be viewed as bandit problem with Markovian rewards. Such a relation provides a convenient way of analyzing the problem, and allows us to develop an "efficient" adaptive control scheme. (We shall precisely define what we mean by efficient in Section 3.)

2. The Problem

2.1 The System Model

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

$$P(u, \theta) := \{P(x, y; u, \theta) | x, y \in \mathcal{X}\} \tag{2.1}$$

and initial probability mass function

$$p(\theta) := \{p(x; \theta) | x \in \mathcal{X}\} \ . \tag{2.2}$$

The parameter θ is unknown, but belongs to a known set Θ. Assume that $\mathcal{X}$,$\mathcal{U}$ and Θ are all finite. Further assume that for

$$x, y \in \mathcal{X};\ u \in \mathcal{U};\ \theta, \theta' \in \Theta,\ P(x, y; u, \theta) > 0 \Rightarrow P(x, y; u, \theta') > 0 \ ; \tag{2.3}$$

for every stationary control law $g : \mathcal{X} \to \mathcal{U}$

$$P^g(\theta) := \{P(x, y; g(x), \theta) | x, y \in \mathcal{X}\} \tag{2.4}$$

is irreducible and aperiodic for all $\theta \in \Theta$, and

$$p(x; \theta) > 0 \text{ for all } x \in \mathcal{X} \text{ and } \theta \in \Theta \ . \tag{2.5}$$

Let

$$\pi^g(\theta) := \{\pi^g(x; \theta) | x \in \mathcal{X}\} \tag{2.6}$$

be the stationary distribution corresponding to $P^g(\theta)$ and let

$$\mu^g(\theta) := \sum_{x \in \mathcal{X}} \pi^g(x; \theta) r(x, g(x)) \tag{2.7}$$

be the mean reward under that stationary distribution.

88CH2531-2/88/0000-1198$1.00 © 1988 IEEE

An "adaptive control scheme" γ is a sequence of random variables $\{U_n\}_{n=0}^{\infty}$ taking values in the set $\mathcal{U}$ such that the event $\{U_n = u\}$ belongs to the σ-field $\mathcal{F}_n$ generated by $X_0, U_0, X_1, U_1, \ldots, U_{n-1}, X_n$. Let $r(X_i, U_i)$ represent the one step reward at time i, where $r : \mathcal{X} \times \mathcal{U} \to \boldsymbol{R}$. Further define $J_n := \sum_{i=0}^{n-1} r(X_i, U_i)$ the total reward at time n as the sum of the one-step rewards upto time n.

Our objective is to find an adaptive control scheme γ which maximizes, in some sense, $E_\theta J_n$ as $n \to \infty$. We shall clarify this notion of optimality in Section 2.4. To achieve our objective we would like to express approximately $E_\theta J_n$ 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 under the invariant distribution corresponding to each g. For this purpose we need to translate any adaptive control scheme γ to an equivalent adaptive control scheme γ' with the following features:

(F1) The control scheme γ' chooses a stationary control law g_n (instead of a control action U_n) at each time n.

(F2) Whenever a fixed but arbitrary stationary control law g, chosen by γ', is used, the sequence of states observed in Markovian. Moreover the sequence of states corresponding to the different stationary control laws, chosen by γ', are independent conditioned on the initial state.

In Section 2.2 we identify a set of conditions which if satisfied, lead to a control scheme γ' that has the above features, and we construct such an equivalent control scheme. In section 2.3 we define the probability space $(\Omega', \mathcal{F}', P'_\theta)$ which allows us to define a sequence of states which for each stationary law g is Markovian, and independent of the sequence of states of any other stationary law g', conditioned on all their initial states. Using $(\Omega', \mathcal{F}', P'_\theta)$ and γ' we can define a control problem that is equivalent to the original one, and we can express $E_\theta J_n$ in terms of the expected number of times each of the stationary control laws g is used up to n, and the expected one-step reward under the invariant distribution corresponding to each g. Such an expression for $E_\theta J_n$ allows us to precisely define the sense in which we want to maximize it.

2.2 The Translation Scheme

Lemma 2.1 Given a controlled Markov chain on a finite state space $\mathcal{X}$ and with a finite control set $\mathcal{U}$, for any adaptive control scheme γ (as defined earlier) there exists an "equivalent adaptive control scheme" γ' taking values on the set $\mathcal{G} := \{g : \mathcal{X} \to \mathcal{U}\}$ of stationary control laws with the following properties.

(i) γ' is a sequence of random variables $\{g_n\}_{n=0}^{\infty}$ taking values on the set $\mathcal{G}$ such that the event $\{g_n = g\}$ belongs to the σ-field $\mathcal{F}'_n$ generated by $X_0, g_0, X_1, g_1, \ldots, g_{n-1}, X_n$.

(ii) $U_n(\omega) = g_n(X_n)(\omega) \qquad \forall n, \omega.$

(iii) If n_k and n_{k+1} are any two successive time instants at which a stationary control law g (fixed, but arbitrary) is used, i.e. $g_{n_k} = g_{n_{k+1}} = g$ and $g_n \neq g$, $n_k < n < n_{k+1}$ then $X_{n_k+1} = X_{n_{k+1}}$.

(Notice that (i) implies $\mathcal{F}_n = \mathcal{F}'_n$.)

Proof: See [16]

2.3 Extending the Probability Space

Let $\Omega = (\mathcal{X} \times \mathcal{U})^\infty$ be the space of all $\mathcal{X} \times \mathcal{U}$ sequences (i.e. sequences of the type $X_0, U_0, X_1, U_1, \ldots$). Give $(\mathcal{X} \times \mathcal{U})^\infty$ the product σ-field $\mathcal{F} = \sigma((\mathcal{X} \times \mathcal{U})^\infty)$, namely, the smallest σ-field such that $X_0, U_0, X_1, U_1, \ldots$ are measurable. There is a unique probability $\mathcal{P}_\theta^\gamma$ on $(\Omega, \mathcal{F})$ such that for all n and all $x_0, \ldots, x_n$ in $\mathcal{X}$ and $u_0, \ldots, u_n$ in $\mathcal{U}$,

$$\begin{aligned}\mathcal{P}_\theta^\gamma\{X_i = x_i, U_i = u_i, \text{ for } i = 0, 1, \ldots n\} \\ = p(x_0; \theta) \prod_{i=0}^{n-1} P(x_i, x_{i+1}; u_i, \theta) \\ \times \prod_{i=0}^{n} 1\{\gamma_i(x_0, u_0, \ldots, x_i) = u_i\} \ . \end{aligned} \tag{2.8}$$

This triple $(\Omega, \mathcal{F}, \mathcal{P}_\theta^\gamma)$ is the minimal underlying probability space required for the description of the problem we address in this paper.

For purposes of analysis and to capture feature (F2) it is useful to extend this probability space which we shall now proceed to do as follows: Let $\mathcal{G} = \{g^1, \ldots, g^d\}$, and $\mathcal{X}^d = \{\underline{x} = (x^{g^1}, \ldots, x^{g^d}) : x^{g^i} \in \mathcal{X}\}$. Let $\Omega' = (\mathcal{X}^d)^\infty$ be the space of all $\mathcal{X}^d$ sequences (i.e. sequences of the type $\underline{X}_0, \underline{X}_1, \ldots$). Give $(\mathcal{X}^d)^\infty$ the producet σ-field $\mathcal{F}' = \sigma((\mathcal{X}^d)^\infty)$, namely, the smallest σ-field such that $\underline{X}_0, \underline{X}_1, \ldots$ are measurable. There is a unique probability $\mathcal{P}'_\theta$ on $(\Omega', \mathcal{F}')$ such that for all n and all $\underline{x}_0, \underline{x}_1, \ldots, \underline{x}_n$ in $\mathcal{X}^d$,

$$\begin{aligned}\mathcal{P}'_\theta\{\underline{X}_i = \underline{x}_i \text{ for } i = 0, 1, \ldots n\} \\ = p'_\theta(f(\underline{x}_0)) \prod_{j=1}^{d} \prod_{i=0}^{n-1} P^{g^j}(x_i^{g^j}, x_{i+1}^{g^j}; \theta)\end{aligned} \tag{2.9}$$

where $f : \mathcal{X}^d \to \mathcal{X} \cup \{\Delta\}$, Δ is an arbitrary element used to augment the state space $\mathcal{X}$ for the purposes of analysis, and f is defined as follows: For each $x \in \mathcal{X}$ left cyclically shift $\{x^{1'} \ldots x^{k'}\}$ to $\{x^1, \ldots, x^k\}$ such that $x^1 = x$. Consider $\mathcal{G}_0^i$ (from section 2.2) constructed as before on the ordering $\{x^1, \ldots, x^k\}$. Let $h : \mathcal{X} \to \mathcal{X}^d$ such that if $g^j \in \mathcal{G}_0^i$ then $h^j(x) = x^i$. Clearly, h is one-to-one, but not onto. Let $h[\mathcal{X}]$ be the range of h, and $h^{-1} : h[\mathcal{X}] \to \mathcal{X}$ be the inverse of h on its range (h^{-1} is well-defined as h is one-to-one.) Finally, let $f|_{h[\mathcal{X}]} = h^{-1}$ and $f(\underline{x}) = \Delta \ \forall \underline{x} \in \mathcal{X}^d - h[\mathcal{X}]$, and $p'_\theta|_{\mathcal{X}} = p(\theta)$ (defined by (2.2)) and $p'_\theta(\Delta) = 0$.

Now on this probability space that we have constructed (note that there is no dependence on the adaptive control scheme γ so far) we can define the random process $X_0^\gamma, U_0^\gamma, X_1^\gamma, U_1^\gamma, \ldots$ by using the equivalent adaptive control scheme γ'. To start off let $X_0^\gamma := f(\underline{X}_0)$. Now given $X_0^\gamma, U_0^\gamma, \ldots, X_n^\gamma$ choose adaptively g_n such that, $U_n^\gamma := g_n(X_n^\gamma)$ and $X_{n+1}^\gamma := X_{T_n^{g_n}+1}^{g_n}$ where $T_n^{g_n}$ is the number of times the control law g_n was used upto time n (in $X_0, U_0, \ldots, X_n$), and $X_{T_n^{g_n}+1}^{g_n}$ is the component of $\underline{X}_{T_n^{g_n}+1}$ corresponding to g_n. It can be easily verified that the random process $X_0^\gamma, U_0^\gamma, X_1^\gamma, U_1^\gamma, \ldots$ constructured above has the same distribution (in $(\Omega', \mathcal{F}', \mathcal{P}'_\theta)$) as the one given by $(\Omega, \mathcal{F}, \mathcal{P}_\theta^\gamma)$. Note that for $\underline{X}_0 \ni f(\underline{X}_0) = \Delta$ the process is undefined, but that is not important as $\mathcal{P}'_\theta\{\underline{X}_0 : f(\underline{X}_0) = \Delta\} = 0$.

Using $(\Omega', \mathcal{F}', \mathcal{P}'_\theta)$ and γ' we can now express $E_\theta J_n$ in terms of the expected number of times each stationary control law g is used and the expected one-step reward under the invariant distribution corresponding to each g.

2.4 Analysis of the Reward Criterion

Consider

$$\begin{aligned} J_n &= \sum_{i=0}^{n-1} r(X_i, U_i) \\ &= \sum_{i=0}^{n-1} r(X_i, U_i) \sum_{g \in \mathcal{G}} 1(g_i = g) \sum_{x \in \mathcal{X}} 1(X_i = x) \\ &= \sum_{g \in \mathcal{G}} \sum_{x \in \mathcal{X}} \sum_{i=0}^{n-1} r(X_i, U_i) 1(g_i = g) 1(X_i = x) \end{aligned}$$

$$= \sum_{g\in\mathcal{G}} \sum_{x\in\mathcal{X}} r(x,g(x)) N^g(x,T_n^g) \qquad (2.10)$$

where

$$N^g(x,T_n^g) = \sum_{i=0}^{T_n^g-1} 1(X_i^g = x)$$
$$= \sum_{i=0}^{n-1} 1(X_i = x, g_i = g)$$

and

$$T_n^g = \sum_{i=0}^{n-1} 1(g_i = g) \ . \qquad (2.11)$$

Note that in the extended probability space $(\Omega', \mathcal{F}', P'_\theta)$ T_n^g is a stopping w.r.t. the increasing family of σ-algebras $\{(\bigvee_{\substack{g'\in\mathcal{G}\\ g'\neq g}} \mathcal{F}_\infty^{g'}) \bigvee \mathcal{F}_n^g\}$ where $\mathcal{F}_n^g = \sigma(X_0^g, X_1^g, \ldots, X_n^g)$ and $\mathcal{F}_\infty^g = \bigvee_n \mathcal{F}_n^g$.

To express $EN^g(x,T_n^g)$ in terms of the invariant distribution under g and ET_n^g we use the following result:

Lemma 2.2 Let $X_0, X_1, \ldots$ be Markovian with finite state space $\mathcal{X}$, transition matrix P-irreducible and aperiodic and stationary distribution π. Let $\mathcal{F}_n$ denote the σ-algebra generated by $X_0, X_1, \ldots, X_n$. Let $\mathcal{G}$ be another σ-algebra and A an event such that $A \in \mathcal{F}_0 \vee \mathcal{G}$ and $\{X_0 = x\} \cap A = \begin{cases} A & A \subset \{X_0 = x\} \\ \phi & \text{otherwise} \end{cases}$. Furthermore let $\mathcal{G}$ be independent of $\mathcal{F}_\infty$ conditioned on the event A. Let τ be a stopping of $\{\mathcal{G} \vee \mathcal{F}_n\}$ such that $E[\tau|A] < \infty$. Let

$$N(x,\tau) = \sum_{i=0}^{\tau-1} 1(X_i = x)$$

Then, for some fixed constant K , independent of A, x and τ.

$$|E[N(x,\tau)|A] - \pi(x)E[\tau|A]| \leq K \qquad (2.12)$$

Proof: Follows from Lemma 2.1 in [11].

Notice that $\bigvee_{\substack{g'\in\mathcal{G}\\ g'\neq g}} \mathcal{F}_\infty^{g'}$ and $\mathcal{F}_\infty^g$ are independent conditioned on the event $A_{\underline{x}} = \{\underline{X}_0 = \underline{x}\}, \underline{x} \in \mathcal{X}^d$. Moreover,

$$A_{\underline{x}} \in \bigvee_{g\in\mathcal{G}} \mathcal{F}_0^g \subset ((\bigvee_{\substack{g'\in\mathcal{G}\\ g'\neq g}} \mathcal{F}_\infty^{g'}) \bigvee \mathcal{F}_0^g), \text{ and}$$

$$\{X_0^g = x\} \bigcap \{\underline{X}_0 = \underline{x}\} = \begin{cases} \{\underline{X}_0 = x\}; \{\underline{X}_0 = \underline{x}\} \subset \{X_0^g = x\} \\ \phi \quad \text{otherwise} \end{cases} .$$

Therefore by Lemma 2.2 it follows that

$$|E_\theta[N^g(x,T_n^g)|A_{\underline{x}}] - \pi^g(x;\theta)E_\theta[T_n^g|A_{\underline{x}}]| \leq K$$

for some fixed constant K independent of $\underline{x}, x$ and n.

Thus,

$$|E_\theta[N^g(x,T_n^g)] - \pi^g(x,\theta)E_\theta[T_n^g]| \leq K \qquad (2.13)$$

From (2.10) and (2.13) it follows that

$$|E_\theta J_n - \sum_{g\in\mathcal{G}} \mu^g(\theta) E_\theta T_n^g| \leq K' \qquad (2.14)$$

where K' is independent of n and $\mu^g(\theta)$ is as defined by (2.7). Let $g^*(\theta) = \arg\max_{g\in\mathcal{G}}(\mu^g(\theta))$, and for simplicity assume that it is unique for each $\theta \in \Theta$. 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

$$|E_\theta J_n - n\mu^{g^*(\theta)}(\theta)| \leq K'.$$

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

$$L_n(\theta) := n\mu^{g^*(\theta)}(\theta) - E_\theta J_n \qquad (2.15)$$

By (2.14) it follows that

$$|L_n(\theta) - \sum_{\substack{g\in\mathcal{G}\\ g\neq g^*(\theta)}} (\mu^{g^*(\theta)}(\theta) - \mu^g(\theta)) E_\theta T_n^g| \leq \text{ const.} \qquad (2.16)$$

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, logrithmic, linear etc.). Thus, this is a stronger criterion for optimality than the average reward per unit time criterion (used in [1] - [7]) which only requires the *Loss* to be $o(n)$. In view of (2.16) the above problem is reduced to one of minimizing the rate at which $E_\theta T_n^g$ increases for $g \in \mathcal{G}, g \neq g^*(\theta)$.

Note that it is impossible to minimize $L_n(\theta)$ uniformly over all parameters $\theta \in \Theta$. For example the stationary control scheme $g_n = g^*(\theta)$ for all n, will have a finite *Loss* where the true parameter is θ. However, when the true parameter is θ' such that $g^*(\theta') \neq g^*(\theta)$, then this scheme will have a *Loss* proportional to n. Having made this observation we call a scheme "uniformly good" if for every parameter $\theta \in \Theta$

$$L_n(\theta) = o(n^\alpha) \text{ for every } \alpha > 0 \qquad (2.17)$$

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

3. 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

$$\begin{aligned} B(\theta) &:= \{\theta' \in \Theta : P^{g^*(\theta)}(\theta') = P^{g^*(\theta)}(\theta) \text{ and } g^*(\theta') \neq g^*(\theta)\} \ , \\ \mathcal{G}_\theta &:= \mathcal{G} - \{g^*(\theta)\} \ , \\ \mathcal{A}_\theta &:= \left\{(\alpha^g, g \in \mathcal{G}_\theta) : \alpha^g \geq 0, \sum_{g\in\mathcal{G}_\theta} \alpha^g = 1\right\} \ , \qquad (3.1) \\ d_\theta(g) &:= (\mu^{g^*(\theta)}(\theta) - \mu^g(\theta)) \text{ and} \\ I^g(\theta,\theta') &:= \sum_{x\in\mathcal{X}} \pi^g(x;\theta) \sum_{y\in\mathcal{X}} P^g(x,y;\theta) \log \frac{P^g(x,y;\theta)}{P^g(x,y;\theta')} . \end{aligned}$$

Note that $I^g(\theta,\theta')$ is just the expectation with respect to the invariant measure of $P^g(\theta)$ of the Kulback Liebler numbers between the individual rows of $P^g(\theta)$ and $P^g(\theta')$ thought of as probability distributions on $\mathcal{X}$.

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 non-empty. Then for any uniformly good control scheme ϕ, under the parameter θ,

$$1) \lim_{n\to\infty} P_\theta \left\{ \sum_{\mathcal{G}_\theta} T_n^g d_\theta(g) < \frac{\log n}{1+2\rho} \cdot \frac{1}{\max_{\alpha\in\mathcal{A}_\theta} \min_{\theta'\in B(\theta)} \frac{\sum_{\mathcal{G}_\theta} \alpha^g I^g(\theta,\theta')}{\sum_{\mathcal{G}_\theta} \alpha^g d_\theta(g)}} \right\}$$
$$= 0 \ \forall \ \rho > 0 \ . \qquad (3.2)$$

Consequently,

$$2)\ \liminf_{n\to\infty} \frac{L_n(\theta)}{\log n} \geq \min_{\alpha\in\mathcal{A}_\theta} \max_{\theta'\in B(\theta)} \frac{\sum_{\mathcal{G}_\theta} \alpha^g d_\theta(g)}{\sum_{\mathcal{G}_\theta} \alpha^g I^g(\theta,\theta')} \ . \tag{3.3}$$

Proof

The proof can easily be obtained from that of Theorem 3.1 of [8] by subsituting g for u and $\mathcal{G}_\theta$ for $\mathcal{U}_\theta$ and by invoking the ergodic theorem instead of the strong law of large numbers. □

Note that we do not have a lower bound for those values of θ for which $B(\theta)$ is empty. In view of this observation and the above lower bound we call a scheme "efficient" if

$$\begin{aligned} \limsup_{n\to\infty} \frac{L_n(\theta)}{\log n} &\leq \min_{\alpha\in\mathcal{A}_\theta} \max_{\theta'\in B(\theta)} \frac{\sum_{\mathcal{G}_\theta} \alpha^g d_\theta(g)}{\sum_{\mathcal{G}_\theta} \alpha^g I^g(\theta,\theta')} \\ &\qquad \text{if } B(\theta) \text{ is non-empty,} \\ L_n(\theta) &< \infty \qquad \text{if } B(\theta) \text{ is empty.} \end{aligned} \tag{3.4}$$

4. The Control Scheme

4.1 Preliminaries

Let $M^{(2)}$ be the unit simplex in $\mathbb{R}^{|\mathcal{X}|^2}$ identified with the space of probability measures on $\mathcal{X}^2$.

Let

$$\nu_\theta^g(x,y) := \pi^g(x;\theta)\, P^g(x,y;\theta);\ \ x,y\in\mathcal{X} \tag{4.1}$$

Then $\nu_\theta^g = \{\nu_\theta^g(x,y)\ :\ x,y\in\mathcal{X}\} \in M^{(2)}$. Since Θ and $\mathcal{G}$ are finite ν_θ^g take on only a finite number of points in $M^{(2)}$. Therefore it is possible to find an $\epsilon > 0$ such for all values of ν_θ^g we can identify ϵ-neighborhoods ("ϵ-nbd of ν_θ^g") of the type:

$$\epsilon\text{-nbd}(\nu_\theta^g) := \{\nu\in M^{(2)} : \max_{x,y\in\mathcal{X}} |\nu(x,y) - \nu_\theta^g(x,y)| < \epsilon\} \tag{4.2}$$

which are disjoint for distinct values of ν_θ^g.

Also define

$$S(\theta) := \{\theta'\in\Theta : P^{g^*(\theta)}(\theta') = P^{g^*(\theta)}(\theta) \text{ and } g^*(\theta') = g^*(\theta)\} \tag{4.3}$$

This is the set of parameters for which the optimal control laws are the *same* as that for θ, and the transition probabilities under the optimal control law are also identical. Let

$$\mathcal{G}(S(\theta)) := \{g : P^g(\theta') \neq P^g(\theta),\ \ \theta'\in S(\theta)\}. \tag{4.4}$$

Recall from Section 3 that

$$B(\theta) := \{\theta'\in\Theta : P^{g^*(\theta)}(\theta') = P^{g^*(\theta)}(\theta) \text{ and } g^*(\theta') \neq g^*(\theta)\} \ . \tag{4.5}$$

This is the set of parameters for which the optimal control laws are *better* than the optimal control law for θ, and the transition probabilities under the optimal control law for θ are identical.

Let

$$\alpha(\theta) = \{\alpha^g(\theta)\ :\ g\in\mathcal{G}_\theta\} \tag{4.6}$$

achieve the minimum in the lower bound for the *Loss* in (3.2), where $\mathcal{G}_\theta = \mathcal{G} - \{g^*(\theta)\}$ and

$$T^g_{\theta,x_0} = E^g_\theta[inf\{n\geq 1 | X_n = x_0\} | X_0 = x_0] \ , \tag{4.7}$$

be the expected reccurrence time of the state x_0 under the control law g. On the basis of these define,

$$\beta(\theta) = \{\beta^g(\theta) : g\in\mathcal{G}_\theta\} \text{ with } \beta^g(\theta) = \frac{\alpha^g(\theta)/T^g_{\theta,x_0}}{\sum_{g\in\mathcal{G}_\theta} \alpha^g(\theta)/T^g_{\theta,x_0}}. \tag{4.8}$$

4.2 Description of the Control Scheme

Let $x_0\in\mathcal{X}$ be an arbitrary but fixed state. Define the $\{\mathcal{F}_t = \sigma(X_0,U_0,X_1,\ldots,X_{t-1},U_{t-1},X_t)\}$ stopping times $\tau_0,\tau_1,\ldots$ by $\tau_m := inf\{t > \tau_{m-1} | X_t = x_0\}, m\geq 1$, and $\tau_0 = inf\{t | X_t = x_0\}$. The control scheme we construct chooses a stationary control law at times $0,\tau_0,\tau_1,\ldots$ adaptively on the basis of all the past observations and past actions, and use this control law till $\tau_0-1,\tau_1-1,\tau_2-1,\ldots$ 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 that 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 $t=0$, choose a fixed but arbitrary control law g_0 and use it till time τ_0-1. Then to start off, use each of the control laws $g\in\mathcal{G}$ once each. From then at each recurrrence point, compute the empirical pair measure $\rho_n^g := \{\rho_n^g(x,y) | x,y\in\mathcal{X}\} \in M^{(2)}$ corresponding to each $g\in\mathcal{G}$ as

$$\rho_n^g(x,y) := \frac{1}{T_n^g - \tau_0} \sum_{i=\tau_0}^{n-1} 1\{g_i = g, X_i = x, X_{i+1} = y\} \tag{4.9}$$

where n is the actual time

Define the conditions

$C1(\theta)$: $\rho_n^g\in\epsilon$-nbd (ν_θ^g) $\forall\ g\in\mathcal{G}$ and $B(\theta)$ is empty

$C2(\theta)$: $\rho_n^g\in\epsilon$-nbd (ν_θ^g) $\forall g\in\mathcal{G}$ and $B(\theta)$ is non-empty.

$C3$: there does not exist $\theta\in\Theta$ such that $\rho_n^g\in\epsilon$-nbd (ν_θ^g) $\forall g\in\mathcal{G}$.

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

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

2) If $C2(\theta)$ is satisfied for some $\theta\in\Theta$ then do the following: Maintain a count of the number of instances condition $C2(\theta)$ is satisfied. Of these, for the first instance choose among those control laws $g\in\mathcal{G}_\theta$ randomly with probabilities $\beta^g(\theta)$. Refer to this process as "randomization". For those instances when this count is even (call this situation $C2(\theta)$ a) use $g^*(\theta)$. For other instances when the count is odd (call this situation $C2(\theta)$ b) compute the likelihood ratio

$$\Lambda_n(\theta) := \lambda_{T^r_n}(\theta) \ := \min_{\theta'\in B(\theta)} \prod_{t=0}^{T^r_n-1} \frac{P^{g^r_t}(X^r_t, X^r_{t+1};\theta)}{P^{g^r_t}(X^r_t, X^r_{t+1};\theta')}$$

of θ vs $B(\theta)$, where $X^r_0, g^r_0, X^r_1, \ldots, g^r_{T^r_n-1}, X^r_{T^r_n}$ is the sequence of pairs of control laws used and states observed upto time n when "randomization" is done with $\beta(\theta)$. If $\Lambda_n > K_{n+1}$ (say $C2(\theta)b1$), where $K_n = n(\log n)^p$ for some fixed $p>1$, the use $g^*(\theta)$. If $\Lambda_n \leq K_{n+1}$ (say $C2(\theta)b2$) then do the following: Maintain a count of the number of instances this condition ($C2(\theta)b2$) is satisfied. If this count is a perfect square (say $C2(\theta)b2a$) then use round robin amongst $g\in\mathcal{G}(S(\theta))$. If this count is not a perfect square (say $C2(\theta)b2b$) then do "randomization" using $\beta(\theta)$.

3) If $C3$ is satisfied then use round-robin amongst $g\in\mathcal{G}$.

4.3 Upper Bound on the *Loss*

In this section we derive an upper bound on the *Loss* associated with the adaptive control scheme γ^* constructed in Section 4.2. 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 $X_0, X_1, \ldots$ be Markovian with finite state space $\mathcal{X}$, transition matrix P, invariant distribution π, and initial distribution p. Let $M^{(2)}$ be the unit simplex on $R^{|X|^2}$ identified with the space of probability measures on $\mathcal{X}^2$, and let $K \subset M^{(2)}$, closed, such that $\pi P \notin K$. Let $\rho_n := \{\rho_n(x,y) | x,y \in \mathcal{X}\}$ where $\rho_n(x,y) := \frac{1}{n}\sum_{i=0}^{n-1} 1\{X_i = x, X_{i+1} = y\}$. Then

(i) $P(\rho_n \in K) < A\epsilon^{-an}$ for all $n \geq 1$ for some positive constants A, a.

Let $N := \sum_{n=1}^{\infty} 1(\rho_n \in K)$. Then

(ii) $EN < \infty$

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

(iii) $EL < \infty$

Proof: See [16].

Lemma 4.2: Let $S_n = X_1 + \ldots + X_n$ where $X_1, X_2, \ldots$ are i.i.d., $EX_1 > 0$ and let $N = \sum_{n=1}^{\infty} 1(S_n \leq 0), L = \sum_{n=1}^{\infty} 1(\inf_{t \geq n} S_t \leq 0)$. Then the following are equivalent:

(a) $E(|X_1|^2\, 1(X_1 \leq 0)) < \infty$.

(b) $EN \quad < \infty$.

(c) E L $\quad < \infty$.

Proof: See Hogan [15].

Lemma 4.3: Let $X_1, X_2, \ldots$ be i.i.d. Let f^i be a real valued Borel function such that $0 < Ef^i(X_1) < \infty, i \in I$, finite. Let $S_n^i = f^i(X_1) + f^i(X_2) \ldots + f^i(X_n), L_A^i = \sum_{n=1}^{\infty} 1(\inf_{t \geq n} S_t^i \leq A)$, and $L_A = \max_{i \in I} L_A^i$. If $E(|f^i(X_1)|^2\, 1(f^i(X_1) \leq 0)) < \infty$ for all $i \in I$, then

$$\limsup_{A \to \infty} \frac{EL_A}{A} \leq \frac{1}{\min_{i \in I}(Ef^i(X_1))} \tag{4.10}$$

Proof: See [16].

Theorem 4.1 Let $\theta \in \Theta$ be such that $B(\theta)$ is non-empty. Then,

(1) $\limsup_{n \to \infty} [E_\theta [\sum_{m=1}^{\infty} 1(\lambda_{\tau_m}(\theta) \leq K_{n+1})] / \log n]$

$$\leq \frac{1}{\min_{\theta' \in B(\theta)} \sum_{\mathcal{G}_\theta} \beta^g(\theta) T^g_{\theta, x_0} I^g(\theta, \theta')} \tag{4.11}$$

(2) $P_{\theta'}\{\lambda_i(\theta) > K_{n+1} \text{ for some } 1 \leq i \leq n\} \leq \frac{1}{K_{n+1}}$ for $\theta' \in B(\theta)$. (4.12)

Proof: See [16].

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

$$\text{(i)} \quad E_\theta T_n^g \leq \left(\frac{\alpha^g(\theta)}{\min_{\theta' \in B(\theta)} \sum_{\mathcal{G}_\theta} \alpha^g(\theta) I^g(\theta, \theta')} + o(1) \right) \log n \quad \text{if } B(\theta) \text{ is non-empty}$$

$$E_\theta T_n^g < \infty \quad \text{if } B(\theta) \text{ is empty} . \tag{4.13}$$

Consequently

$$\text{(ii)} \quad L_n(\theta) \leq \left(\frac{\sum_{\mathcal{G}_\theta} \alpha^g(\theta)\, d_\theta(g)}{\max_{\theta' \in B(\theta)} \sum_{\mathcal{G}_\theta} \alpha^g(\theta) I^g(\theta, \theta')} + o(1) \right) \log n \quad \text{if } B(\theta) \text{ is non-empty}$$

$$L_n(\theta) < \infty \quad \text{if } B(\theta) \text{ is empty} \tag{4.14}$$

where $\alpha(\theta) = \{\alpha^g(\theta) : g \in \mathcal{G}_\theta\}$ is defined by (4.6).

Proof: See [16].

5. Conclusions

In this paper we considered the problem of adaptive control of Markov Chains. The optimality criterion used, namely minimizing the rate at which the Loss increases is stronger than the average reward per unit time criterion. Multi-armed bandit problems with *"Loss"* as the optimality criterion is one class of stochastic adaptive control problems that has previously been analyzed. Therefore one way to proceed with our problem is to relate it to the multi-armed bandit problem, like was done in [8] for the controlled i.i.d. process problem. The translation scheme and the extended probability space are crucial in allowing us to view the adaptive control of Markov chains as a multi-armed bandit problem. The stationary control laws correspond to the "arms", and the sequence of states observed when any particular stationary control law is used are Markovian. The formulation then resembles that of the multi-armed bandit problem in [11], part II. One very important difference between our problem and that of [11] is that the parametrization of the "arms" in our problem is not independent. This difference is reflected in the lower bound on the Loss we obtain in Section 3, and also needs to be kept in mind when designing an optimal scheme like the one of Section 4. The control scheme presented in Section 4 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.

Acknowledgements

The research of Rajeev Agrawal and Demosthenis Teneketzis was supported in part by NSF Grant No. ECS-8517708 and ONR Grant No. N00014-87-K-0540.

References

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

[2] P. Mandl,"Estimation and Control in Markov Chains", Adv. Appl. Prob., 6 (1974), pp. 40-60.

[3] V. Borkar and P. Varaiya, "Adaptive Control of Markov Chains, I: finite parameter set", *IEEE Transactions on Automatic Control*, AC-24, 1979, pp. 953-958.

[4] V. Borkar and P. Varaiya, "Identification and Adaptive Control of Markov Chains", *SIAM J. on Control and Optimization*, 20, 1982, pp. 470-489.

[5] P. R. Kumar and A. Becker, "A new family of optimal adaptive controllers for Markov chains", *IEEE Transactions on Automatic Control*, AC-27, 1982, pp. 137-146.

[6] P. R. Kumar and W. Lin, "Optimal adaptive controllers for unknown Markov chains", *IEEE Transactions on Automatic Control*, AC-27, 1982, pp. 765-774.

[7] R. A. Milito and J. B. Cruz, "An optimization oriented approach to the adaptive control of Markov chains", *IEEE Transactions on Automatic Control*, Vol AC-32, No. 9, September 1987.

[8] R. Agrawal, D. Teneketzis, V. Anantharam, "Asymptotically Efficent Allocation Schemes for Controlled I.I.D. Processes: Finite Parameter Space," Technical Report No. 253, Communications and Signal Processing Lab, Univ. of Michigan, January, 1988.

[9] T.L. Lai and H. Robbins, "Asymptotically Efficient Adaptive Allocation Rules", em Advances in Applied Mathematics, 1984.

[10] T.L. Lai and H. Robbins, "Asymptotically Optimal Allocation of Treatments in Sequential Experiments," In 'Design of Experiments' (eds. T.J. Santuer and A.C. Tamhane), Marcel-Dekker, New York.

[11] 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 Transaction on Automatic Control*, Vol 1 AC-32, No. 11, November 1987, pp 968-982.

[12] R. Agrawal, M. Hegde, and D. Teneketzis, "Asymptotically Efficient Adaptive Allocation Rules for the Multi-armed Bandit Problem with Switching Cost", Technical Report No. 246, Communications and Signal Processing Lab, Univ. of Michigan, April, 1987.

[13] S. Ross, "Stochastic Processes", Wiley, 1983.

[14] R.S. Ellis, "Entropy, Large Deviations, and Statistical Mechanics", Springer-Verlag, 1985.

[15] M. Hogan, "Moments of the Minimum of a Random Walk and Complete Convergence," Technical Report No. 21, Department of Statistics, Stanford University, Jan 1983.

[16] R. Agrawal, D. Teneketzis, V. Anantharam, "Asymptotically Efficent Allocation Schemes for Controlled Markov Chains: Finite Parameter Space," Technical Report No. 254, Communications and Signal Processing Lab, Univ. of Michigan, February, 1988.