Asymptotically Efficient Adaptive Allocation Schemes for Controlled Markov Chains: Finite Parameter Space

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

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. The *loss* associated with a control scheme at a parameter value is the function of time giving the difference between the maximum reward that could have been achieved if the parameter were known, and the reward achieved by the scheme. Since it is impossible to uniformly minimize the *loss* for all parameter values 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 for *every* parameter value, and is therefore asymptotically efficient.**

I. Introduction

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

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

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

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

where θ is the same as in (1.1). At each time i we choose a control action U_i (on the basis of the entire past $X_0, U_0, X_1, U_1, \cdots, X_i$) and collect a one-step reward $r(X_i, U_i)$. The idea is to maximize, in some sense, the expected value of the sum of one-step rewards up to time n [i.e., $E_\theta \sum_{i=0}^{n-1} r(X_i, U_i)$] as $n \to \infty$. In particular, let $J_n^*(\theta)$ be the supremum of $E_\theta \sum_{i=0}^{n-1} r(X_i, U_i)$ over all control schemes, and define the *loss* $L_n(\theta)$ as

$$L_n(\theta) = J_n^*(\theta) - E_\theta \sum_{i=0}^{n-1} r(X_i, U_i). \quad (1.3)$$

Manuscript received March 1, 1988; revised May 5, 1989. 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 is with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53706-1691.

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 8931407.

The problem is to find an adaptive control scheme that minimizes the rate at which the *loss* $L_n(\theta)$ increases $n \to \infty$. This criterion will be clarified further in Section II, and we shall henceforth refer to it as the *loss* criterion.

One of the current approaches to stochastic adaptive control problems is the so-called "certainty equivalence control with forcing" (cf. [1]). This scheme has the following features: i) at *almost* every instant of time the unknown parameter θ is estimated first and then the corresponding optimal control law is used (certainty equivalence); ii) every once in a while experimentation with various control actions (forcing) is done in order to escape false identification traps. Even though "certainty equivalence control with forcing" is self-tuning in the Cesaro sense and is therefore optimal for the average-reward-per-unit-time criterion (cf. [1]), there remains the problem of how much reward such a strategy sacrifices. How often is experimentation needed in order to avoid false identification while still achieving the maximum possible reward for all parameter values? This issue can be investigated by introducing the *loss* criterion. The *loss* criterion used in this paper is stronger than the average-reward-per-unit-time criterion, used in [1]–[7], which just requires $L_n(\theta)$ to be $o(n)$. For the *loss* criterion it is no longer clear that certainty-equivalence-control-with-forcing is optimal.

The *loss* criterion was first used by Lai and Robbins in the context of the multiarmed bandit problem, and a solution methodology was developed for bandits with independent identically distributed arms in their seminal papers [9], [10]. Various extensions of the Lai–Robbins formulation of the multiarmed bandit problems have been reported in [11], [12]. A crucial aspect of the bandit problem (see [9] for an introduction) is that the arms of the bandit are independent. This results in a clear definition of the role of experimentation: experimenting with one arm gives information only about the parameter of that arm; knowledge about the parameters of the other arms is unchanged. In the context of controlled Markov chains, experimentation corresponds to using a control strategy that does not appear optimal. However, use of such a strategy changes the state of knowledge for *all* parameter values simultaneously.

In spite of these difficulties, we are able to address the issue of optimal experimentation for controlled Markov chains in this paper, by treating the problem as a kind of multiarmed bandit problem in a manner similar to what we did for the controlled i.i.d. process in [8]. The crucial new idea is the "translation scheme" (Section II-B) which along with the construction of an "extended probability space" (Section II-C) allow us to convert the original control problem into one of "playing" stationary control strategies. The subsequent analysis is also more delicate than that of [9]–[12] because of the mixed role of experimentation discussed above.

The rest of the paper is organized as follows. In Section II it is shown that the original control problem can be converted into one of "playing" stationary control strategies. More specifically, it is shown that the *loss* $L_n(\theta)$ can be expressed in terms of the expected number of times each stationary control law g is used up to time n and the expected one-step reward under the invariant

0018-9286/89/1200-1249$01.00

distribution corresponding to each g. A lower bound on the *loss* $L_n(\theta)$ is developed in Section III. An adaptive control scheme is proposed in Section IV; it is shown that the *loss* associated with the proposed control scheme is equal to the lower bound on $L_n(\theta)$, thus proving that the proposed control scheme is "asymptotically efficient."

II. The Problem

A. The System Model

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

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

and initial probability mass function

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

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

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

For every stationary control law $g: \mathfrak{X} \to \mathfrak{U}$

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

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

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

Let

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

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

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

to be the mean reward under the stationary distribution $\pi^g(\theta)$. 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 up to time n.

An "adaptive control scheme" γ is a sequence of random variables $\{U_n\}_{n=0}^{\infty}$ taking values in the set $\mathfrak{U}$ such that the event $\{U_n = u\}$ belongs to the σ-field $\mathfrak{F}_n$ generated by $X_0, U_0, X_1, U_1, \cdots, U_{n-1}, X_n$.

Our objective is to find an adaptive control scheme γ which maximizes, in some sense, $E_\theta^\gamma J_n$ as $n \to \infty$. We shall now clarify this notion of optimality. For each $\theta \in \Theta$, and each $n \geq 1$, let $J_n^*(\theta)$ be the supremum of $E_\theta^\gamma J_n$ over all control schemes γ. In most cases of interest, this supremum will not be attained by the same control scheme for different values of θ and n. Thus, for any control scheme γ we define the *loss*

$$L_n^\gamma(\theta) := J_n^*(\theta) - E_\theta^\gamma J_n \geq 0 \tag{2.8}$$

which represents the shortfall from the best possible. Minimizing the *loss* is then equivalent to maximizing the expected sum of rewards. The objective is to find one control scheme γ that works well for all $\theta \in \Theta$ and for large n. In particular, we want to restrict attention to (asymptotically) *uniformly good* control schemes, i.e., those for which

$$L_n^\gamma(\theta) = o(n^\alpha), \qquad \forall \alpha > 0,\ \theta \in \Theta. \tag{2.9}$$

Such schemes do not allow the *loss* to increase too rapidly for any $\theta \in \Theta$. We would like to find a control scheme that minimizes the rate at which the *loss* increases within the class of *uniformly good* schemes.

Note that optimality with respect to the average-reward (cost)-per-unit-time criterion requires the weaker condition

$$L_n^\gamma(\theta) = o(n^1), \qquad \forall \theta \in \Theta. \tag{2.10}$$

Thus, the notion of optimality we are using here is clearly stronger than the average-reward (cost)-per-unit-time criterion.

In order to evaluate the performance of any control scheme we would like to view this adaptive control problem as a multiarmed bandit problem where the arms now correspond to stationary control laws. The motivation for doing this is to express $E_\theta^\gamma J_n$ and thus $L_n^\gamma(\theta)$ in terms of the expected number of times each stationary control law (arm) g is used up to time n, and the expected one-step reward under the invariant distribution corresponding to each g.

To relate our problem to a multiarmed bandit problem we note that if we have a multiarmed bandit problem with Markovian observations (rewards), then the sequence of observations can be realized by appropriate interleaving of the sequences of Markovian observations (rewards) corresponding to different arms. We want to "imitate" this feature of the multiarmed bandit problem in the controlled Markov chain problem in the following manner. First by a *translation scheme* (Section II-B) we identify for any adaptive control scheme an "equivalent adaptive control scheme" that chooses a stationary control law (arm) g_n at each state n. Then, we extend the probability space (Section II-C) so that we now start with sequences of Markovian observations (rewards) corresponding to different stationary control laws g, i.e., with transition probabilities $P^g(\theta)$. By using the extended probability space and the *translation scheme*, we can construct a sequence of observations (and actions) that has the same statistics as the original controlled Markov chain.

After we relate our problem to the multiarmed bandit problem, we show by analysis of the reward criterion (Section II-D) that we can express $E_\theta^\gamma J_n$, and thus $L_n^\gamma(\theta)$ in terms of the expected number of times each stationary control law g is used up to time n and the expected one-step reward under the invariant distribution corresponding to each g.

B. Translation Scheme

In this section, by means of Theorem 2.1, we show that given any adaptive control scheme we can find another adaptive control scheme which at each stage n chooses stationary control laws g_n instead of control actions U_n. This scheme is equivalent to the original control scheme in the sense that $U_n = g_n(X_n)$ for each n [Theorem 2.1 ii)]. Furthermore, the successive times at which any particular stationary control law g is used are such that the corresponding sequence of observations is Markovian with transition probability $P^g(\theta)$. The Markovian property of the sequence of observations is achieved by ensuring that the successive states (observations) in the process corresponding to each stationary control law continue each other, in the sense that for any two successive time instants n_k and n_{k+1} at which the same stationary control law is used, the states X_{n_k+1} and $X_{n_{k+1}}$ are the same (Theorem 2.1 iii)].

The translation scheme developed in this section is the first step in our effort to relate our problem to the multiarmed bandit problem.

Theorem 2.1: Given a controlled Markov chain on a finite state-space $\mathfrak{X}$ and with a finite control set $\mathfrak{U}$, for any adaptive control scheme γ (as defined earlier) there exists an "equivalent adaptive control scheme" γ' taking values on the set $\mathcal{G} := \{g: \mathfrak{X} \to \mathfrak{U}\}$ 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 $\mathfrak{F}'_n$ generated by $X_0, g_0, X_1, g_1, \cdots, g_{n-1}, X_n$.

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 (by Construction): Let $\#\mathfrak{X} = k$ and let $x^{1'}, x^{2'}, \cdots, x^{k'}$ be a prior (but arbitrary) ordering of x. Similarly, let $\#\mathfrak{U} = l$ and $\mathfrak{U} = \{u^1, u^2, \cdots, u^l\}$. To start off, observe X_0 and then reorder $\mathfrak{X}$ as $x^1, x^2, \cdots, x^k$ by a left cyclic shift of the prior ordering, such that $x^1 = X_0$. Define $\mathcal{G}_0^i$, $i = 1, \cdots, k$ inductively as follows:

$$\mathcal{G}_0^1 = \{g \in \mathcal{G} : g(x^j) = u^1, 1 < j \leq k\}$$

$$\mathcal{G}_0^i = \{g \in \mathcal{G} : g(x^j) = u^1, i < j \leq k\} - \bigcup_{j=1}^{i-1} \mathcal{G}_0^j; \qquad i = 2, \cdots, k.$$

Notice that $\mathcal{G}_0^i$; $1 = 1, \cdots, k$ defines a partition of $\mathcal{G}$, i.e., $\cup_{i=1}^k \mathcal{G}_0^i = \mathcal{G}$ and $i \neq j \Rightarrow \mathcal{G}_0^i \cap \mathcal{G}_0^j = \phi$.

Now suppose at time $n \geq 0$, i.e., after observing X_n, we have a partition $\mathcal{G}_n^i : i = 1, \cdots, k$ of $\mathcal{G}$ with the following five properties.

P1) $\mathcal{G}_n^i$, $i = 1, \cdots, k$ is determined by $\mathcal{F}'_n$.

P2) $\forall 1 \leq i \leq k \quad \forall g \in \mathcal{G}_n^i$, the last time up to time $n-1$ that the control g was used (if any) was followed by the state x^i.

Let

$$X_n = x^{j_n} \qquad \text{for some } j_n = 1, \cdots, k. \tag{2.11}$$

Then,

P3) $\forall j_n \leq m \leq k$ and for any $f_m : \{x^1, \cdots, x^m\} \to \mathfrak{U}$ there exists a unique

$$g \in \bigcup_{i=1}^m \mathcal{G}_n^i \ni g|_{\{x^1, \cdots, x^m\}} = f_m.$$

P4) $\forall 1 \leq m < j_n$ there exists a unique

$$f'_m : \{x^1, \cdots, x^m\} \to \mathfrak{U} \ni \forall g \in \bigcup_{i=1}^m \mathcal{G}_n^i, \; g|_{\{x^1, \cdots, x^m\}} \neq f'_m.$$

P5) $\forall 1 < m < j_n$ the above found f'_m's satisfy $f'_{m-1} = f'_m|_{\{x^1, \cdots, x^{m-1}\}}$.

Also assume the following.

P6) g_j, $0 \leq j < n$ satisfy properties i), ii), and iii) of Theorem 2.1.

We shall now show that we can choose a g_n satisfying property P6) on the basis of $\mathcal{F}'_n$ and construct a new partition $\mathcal{G}_{n+1}^i$; $i = 1, \cdots, k$ satisfying properties P1)–P5) assumed true for time n. Choose $g_n \in \mathcal{G}_n^{j_n}$ [j_n as determined by (2.11)] such that

$$g_n|_{\{x^1, \cdots, x^{j_n-1}\}} = f'_{j_n-1} \text{ and } g_n(x^{j_n}) = g_n(X_n) = U_n. \tag{2.12}$$

Such a choice is clearly possible by the above induction hypothesis [properties P3) and P4)]. By noting the fact that U_n is determined by $\mathcal{F}_n = \mathcal{F}'_n$ and by the induction hypothesis [properties P1), P2), and P6)] it follows that P6) is satisfied for $n+1$. Next, let $X_{n+1} = x^{j_{n+1}}$ for some $j_{n+1} = 1, \cdots, k$. If $j_{n+1} = j_n$, then $\mathcal{G}_{n+1}^i := \mathcal{G}_n^i \; \forall i = 1, \cdots, k$, and it trivially follows that $\mathcal{G}_{n+1}^i, i = 1, \cdots, k$ satisfy P1)–P5). Else, if $j_{n+1} \neq j_n$, $\mathcal{G}_{n+1}^{j_n} := \mathcal{G}_n^{j_n} - \{g_n\}$, $\mathcal{G}_{n+1}^{j_{n+1}} := \mathcal{G}_n^{j_{n+1}} + \{g_n\}$, and $\forall i \neq j_n, j_{n+1}$, $\mathcal{G}_{n+1}^i := \mathcal{G}_n^i$. In this case also it is easy to check that $\mathcal{G}_{n+1}^i$ satisfy P1) and P2). To show that $\mathcal{G}_{n+1}^i$ satisfy P3)–P5) consider two cases.

Case 1 $j_{n+1} > j_n$:

- $\forall j_{n+1} \leq m \leq k$, $\cup_{i=1}^m \mathcal{G}_{n+1}^i = \cup_{i=1}^m \mathcal{G}_n^i - \{g_n\} + \{g_n\} = \cup_{i=1}^m \mathcal{G}_n^i$. Thus, P3) is satisfied.
- $\forall 1 \leq m < j_n$, $\cup_{i=1}^m \mathcal{G}_{n+1}^i = \cup_{i=1}^m \mathcal{G}_n^i$. Thus, P4) and P5) are satisfied for $1 \leq m < j_n$ and $1 < m < j_n$, respectively.
- $\forall j_n \leq m < j_{n+1}$, $\cup_{i=1}^m \mathcal{G}_{n+1}^m = \cup_{i=1}^m \mathcal{G}_n^m - \{g_n\}$. Consider the $f'_m = g_n|_{\{x^1, \cdots, x^m\}}$. By the induction hypothesis P3) it then follows that P4) is satisfied for $j_n \leq m < j_{n+1}$.

Clearly, this construction of f'_m also satisfies

$$f'_{m-1} = f'_m|_{\{x^1, \cdots, x^{m-1}\}} \; \forall j_n < m < j_{n+1}$$

and by (2.12) it also follows that

$$\underset{\text{(old)}}{f'_{j_n-1}} = \underset{\text{(new)}}{f'_{j_n}}|_{\{x^1, \cdots, x^{j_n-1}\}}.$$

Thus, P5) is satisfied for $j_n \leq m < j_{n+1}$.

Case 2 $j_{n+1} < j_n$:

- $\forall j_n \leq m \leq k$, $\cup_{i=1}^m \mathcal{G}_{n+1}^i = \cup_{i=1}^m \mathcal{G}_n^i - \{g_n\} + \{g_n\} = \cup_{i=1}^m \mathcal{G}_n^i$. Thus, P3) is satisfied for $j_n \leq m \leq k$.
- $\forall j_{n+1} \leq m < j_n$ $\cup_{i=1}^m \mathcal{G}_{n+1}^i = \cup_{i=1}^m \mathcal{G}_n^i + \{g_n\}$. And since $f'_m = g_n|_{\{x^1, \cdots, x^m\}}$ was the unique one missing from $\cup_{i=1}^m \mathcal{G}_n^i$ [by (2.12) and the induction hypothesis on P4), P5)] it follows that P3) is now satisfied for $j_{n+1} \leq m < j_n$.
- $\forall 1 \leq m < j_{n+1}$ $\cup_{i=1}^m \mathcal{G}_{n+1}^i = \cup_{i=1}^m \mathcal{G}_n^i$ and thus P4) and P5) are satisfied.

The proof of Theorem 2.1 is now complete (using induction) by checking that the induction hypothesis is satisfied at $n = 0$.

C. Extending the Probability Space

In this section we construct an underlying probability space which is defined in terms of sequences of observations, corresponding to each stationary control law g, that are Markovian with transition probabilities $P^g(\theta)$ and are also independent of each other conditioned on the initial state. This construction along with the *translation scheme* developed in Section II-B allow us to construct a sequence of Markovian observations (and actions) that has the same statistics as the original controlled Markov chain. Thus, combining the *translation scheme* developed in Section II-B with the results of this section we manage to "imitate" the feature of the multiarmed bandit with Markovian rewards, discussed in Section II-A (namely, that in a multiarmed bandit problem with Markovian observations, the sequence of observations can be realized by appropriate interleaving of the sequences of Markovian observations corresponding to different arms).

We proceed with the construction by first specifying the minimal underlying probability space needed to describe the controlled Markov chain, and then by "extending" it to the abovementioned probability space.

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

$$\mathcal{P}_\theta^\gamma\{X_i = x_i, U_i = u_i, \qquad \text{for } i = 0, 1, \cdots, n\}$$

$$= p(x_0; \theta) \prod_{i=0}^{n-1} P(x_i, x_{i+1}; u_i, \theta)$$

$$\cdot \prod_{i=0}^{n} 1\{\gamma_i(x_0, u_0, \cdots, x_i) = u_i\}. \tag{2.13}$$

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.

We now construct the extended probability space as follows. Let $\mathcal{G} = \{g^1, \cdots, g^d\}$, and $\mathfrak{X}^d = \{\boldsymbol{x} = (x^{g^1}, \cdots, x^{g^d}) : x^{g^i} \in \mathfrak{X}\}$. Let $\Omega' = (\mathfrak{X}^d)^\infty$ be the space of all $\mathfrak{X}^d$ sequences (i.e., sequences of the type $\boldsymbol{X}_0, \boldsymbol{X}_1 \cdots$). Give $(\mathfrak{X}^d)^\infty$ the product σ-field $\mathcal{F}' = \sigma((\mathfrak{X}^d)^\infty)$, namely, the smallest σ-field such that $\boldsymbol{X}_0, \boldsymbol{X}_1, \cdots$ are measurable. There is a unique probability $\mathcal{P}'_\theta$ on

$(\Omega, \mathcal{F}')$ such that for all n and all $x_0, x_1, \cdots, x_n$ in $\mathfrak{X}^d$,

$$\mathcal{P}'_\theta\{X_i = x_i \text{ for } i = 0, 1, \cdots, n\} = p'_\theta(f(x_0)) \prod_{j=1}^{d} \prod_{i=0}^{n-1} P^{g^j}(x_i^{g^j}, x_{i+1}^{g^j}; \theta) \quad (2.14)$$

where $f: \mathfrak{X}^d \to \mathfrak{X} \cup \{\Delta\}$, Δ is an arbitrary element used to augment the state space $\mathfrak{X}$ for the purposes of analysis, and f is defined as follows. For each $x \in \mathfrak{X}$ left cyclically shift $\{x^1 \cdots x^k\}$ to $\{x^1, \cdots, x^k\}$ such that $x^1 = x$. Consider $\mathcal{G}_0^i$ (from Section II-B) constructed as before on the ordering $\{x^1, \cdots, x^k\}$. Let $h: \mathfrak{X} \to \mathfrak{X}^d$ such that if $g^i \in \mathcal{G}_0^i$, then $h^j(x) = x^i$. Clearly, h is one-to-one, but not onto. Let $h[\mathfrak{X}]$ be the range of h, and $h^{-1}: h[\mathfrak{X}] \to \mathfrak{X}$ be the inverse of h on its range (h^{-1} is well-defined as h is one-to-one). Finally, let $f|_{h[\mathfrak{X}]} = h^{-1}$ and $f(x) = \Delta\ \forall x \in \mathfrak{X}^d - h[\mathfrak{X}]$, and $p'_\theta|_{\mathfrak{X}} = p(\theta)$ [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 $X_0^\gamma, U_0^\gamma, X_1^\gamma, U_1^\gamma, \cdots$ by using the equivalent adaptive control scheme γ' developed in Theorem 2.1. To start off let $X_0^\gamma := f(X_0)$. Now given $X_0^\gamma, U_0^\gamma, \cdots, X_n^\gamma$ 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 up to time n (in $X_0, U_0, \cdots, X_n$), and $X_{T_n^{g_n}+1}^{g_n}$ is the component of $X_{T_n^{g_n}+1}$ corresponding to g_n. It can be easily verified that the random process $X_0^\gamma, U_0^\gamma, X_1^\gamma, U_1^\gamma, \cdots$ constructed 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 $X_0 \ni f(X_0) = \Delta$ the process is undefined, but that is not important as $\mathcal{P}'_\theta\{X_0: f(X_0) = \Delta\} = 0$.

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 up to time n and the expected one-step reward under the invariant distribution corresponding to each g.

D. Analysis of the Reward Criterion

In this section we show, by analysis of the reward criterion, that we can express $E_\theta^\gamma J_n$ and thus $L_n^\gamma(\theta)$ in terms of the expected number of times each stationary control law g is used up to time n and the expected one-step reward under the invariant distribution corresponding to each g.

Consider

$$\begin{aligned} J_n &= \sum_{i=0}^{n-1} r(X_i, U_i) \\ &= \sum_{i=0}^{n-1} r(X_i, U_i) \sum_{g \in \mathcal{G}} 1(g_i = g) \sum_{x \in \mathfrak{X}} 1(X_i = x) \\ &= \sum_{g \in \mathcal{G}} \sum_{x \in \mathfrak{X}} \sum_{i=0}^{n-1} r(X_i, U_i) 1(g_i = g) 1(X_1 = x) \\ &= \sum_{g \in \mathcal{G}} \sum_{x \in \mathfrak{X}} r(x, g(x)) N^g(x, T_n^g) \end{aligned} \quad (2.15)$$

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

$$T_n^g = \sum_{i=0}^{n-1} 1(g_i = g). \quad (2.16)$$

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

To express $E_\theta N^g(x, T_n^g)$ in terms of the invariant distribution under g and $E_\theta T_n^g$ we use the following result.

Lemma 2.1: Let $X_0, X_1, \cdots$ be Markovian with finite state space $\mathfrak{X}$, transition matrix $\mathcal{P}$, irreducible and aperiodic, and stationary distribution π. Let $\mathcal{F}_n$ denote the σ-algebra generated by $X_0, X_1, \cdots, X_n$. Let $\hat{\mathcal{F}}$ be another σ-algebra and A an event such that $A \in \mathcal{F}_0 \vee \hat{\mathcal{F}}$ and $\{X_0 = x\} \cap A = \begin{cases} A & A \subset \{X_0 = x\} \\ \phi & \text{otherwise} \end{cases}$. Furthermore, let $\hat{\mathcal{F}}$ be independent of $\mathcal{F}_\infty$ conditioned on the event A. Let τ be a stopping time of $\{\hat{\mathcal{F}} \vee \mathcal{F}_n\}$ 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,

$$|E[N(x, \tau)|A] - \pi(x) E[\tau|A]| \leq K. \quad (2.17)$$

Proof: Follows from [11, Lemma 2.1].

Notice that $\vee_{\substack{g' \in \mathcal{G} \\ g' \neq g}} \mathcal{F}_\infty^{g'}$ and $\mathcal{F}_\infty^g$ are independent conditioned on the event $A_x = \{X_0 = x\}$, $x \in \mathfrak{X}^d$. Moreover, $A_x \in \vee_{g \in \mathcal{G}} \mathcal{F}_0^g \subset ((\vee_{\substack{g' \in \mathcal{G} \\ g' \neq g}} \mathcal{F}_\infty^{g'}) \vee \mathcal{F}_0^g)$ and

$$\{X_0^g = x\} \cap \{X_0 = x\} = \begin{cases} \{X_0 = x\};\ \{X_0 = x\} \subset \{X_0^g = x\}. \\ \phi \qquad \text{otherwise} \end{cases}$$

Therefore, by Lemma 2.1 it follows that

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

for some fixed constant K independent of x, x, and n.

Thus,

$$|E_\theta[N^g(x, T_n^g)] - \pi^g(x, \theta) E_\theta[T_n^g]| \leq K. \quad (2.18)$$

From (2.15), (2.16), and (2.18) it follows that

$$|E_\theta J_n - \sum_{g \in \mathcal{G}} \mu^g(\theta) E_\theta T_n^g| \leq K' \quad (2.19)$$

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 (up to a constant) for all n, and for this scheme

$$|E_\theta J_n - n\mu^{g^*(\theta)}(\theta)| \leq K'. \quad (2.20)$$

By (2.8), (2.19), and (2.20) it follows that the *loss* $L_n(\theta)$ associated with any adaptive control scheme γ satisfies

$$|L_n(\theta) - \sum_{\substack{g \in \mathcal{G} \\ g \neq g^*(\theta)}} (\mu^{g^*(\theta)}(\theta) - \mu^g(\theta)) E_\theta T_n^g| \leq \text{const.} \quad (2.21)$$

Consequently, the *loss* can be expressed in terms of the expected

number of times each stationary control law g is used up to time n and the expected one-step reward under the invariant distribution corresponding to each g. In view of (2.21), our problem is reduced to one of minimizing the rate at which $E_\theta T_n^g$ increases for $g \in \mathcal{G}$, $g \neq g^*(\theta)$, within the class of *uniformly good* control schemes.

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 concepts. Let

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

$$\mathcal{G}_\theta := \mathcal{G} - \{g^*(\theta)\},$$

$$\mathcal{C}_\theta := \left\{(\alpha^g, g \in \mathcal{G}_\theta) : \alpha^g \geq 0, \sum_{g \in \mathcal{G}_\theta} \alpha^g = 1\right\},$$

$$d_\theta(g) := (\mu^{g^*(\theta)}(\theta) - \mu^g(\theta)) \text{ and}$$

$$I^g(\theta, \theta') := \sum_{x \in \mathcal{X}} \pi^g(x; \theta) \sum_{y \in \mathcal{X}} P^g(x, y; \theta) \log \frac{P^g(x, y; \theta)}{P^g(x, y; \theta')}. \tag{3.1}$$

Thus, $B(\theta)$ is the set of *bad* parameter values associated with θ, namely those parameter values θ' for which the matrix of transition probabilities is the same under θ and θ' when the optimal control law for θ, $g^*(\theta)$ is used, but such that the optimal control law for θ', $g^*(\theta')$ is different from $g^*(\theta)$. The point is that if the true parameter were θ' and we were led to believe it was θ, we would end up trapped into believing $g^*(\theta)$ is the optimal stationary control law to use unless we experiment. $\mathcal{C}_\theta$ should be thought of as a set of *averaging vectors* over $\mathcal{G}_\theta$. Note that $I^g(\theta, \theta')$ is just the expectation with respect to the invariant measure of $P^g(\theta)$ of the Kulback–Leibler numbers between the individual rows of $P^g(\theta)$ and $P^g(\theta')$ thought of as probability distributions on $\mathcal{X}$.

The lower bound on the *loss* 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 ϕ, under the parameter θ,

1)

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

Consequently,

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

Proof: The proof can be easily obtained from that of [8, Theorem 3.1] by substituting 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. The main point to keep in mind is the interpretation of the quantity on the right-hand side of (3.3). It is the minimum, over all averaging vectors associated with θ, of the maximum per unit information cost over all *bad* parameter values associated with θ. □

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 "asymptotically efficient" if

$$\limsup_{n \to \infty} \frac{L_n(\theta)}{\log n} \leq \min_{\alpha \in \mathcal{C}_\theta} \max_{\theta' \in B(\theta)} \frac{\sum\limits_{\mathcal{G}_\theta} \alpha^g d_\theta(g)}{\sum\limits_{\mathcal{G}_\theta} \alpha^g I^g(\theta, \theta')} \quad \text{if } B(\theta) \text{ is nonempty}$$

$$L_n(\theta) < \infty \quad \text{if } B(\theta) \text{ is empty.} \tag{3.4}$$

IV. The Control Scheme

In this section we describe an asymptotically efficient adaptive control scheme. The control scheme presented here 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 optimally the conflict between learning and control. In fact, it will be seen that, roughly, experimentation will be done using the optimal *averaging vector* on the right-hand side of (3.3) to get out of the identification traps of *bad* parameter values.

A. 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

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

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 that for all values of ν_θ^g we can identify ϵ-neighborhoods ("ϵ-nbd of ν_θ^g") of the type

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

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 III that

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

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 (3.3),

where $\mathcal{G}_\theta = \mathcal{G} - \{g^*(\theta)\}$. Let

$$\mathfrak{I}^g_{\theta,x_0} = E^g_\theta [\inf \{n \geq 1 | X_n = x_0\} | X_0 = x_0] \quad (4.7)$$

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

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

B. Description of the Control Scheme

Let $x_0 \in \mathfrak{X}$ be an arbitrary but fixed state. Define the $\{\mathcal{F}_t = \sigma(X_0, U_0, X_1, \cdots, X_{t-1}, U_{t-1}, X_t)\}$ stopping times $\tau_0, \tau_1, \cdots$ 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 ϕ^* we construct chooses a stationary control law at times $0, \tau_0, \tau_1, \cdots$ adaptively on the basis of all the past observations and past actions, and use this control law until $\tau_0 - 1, \tau_1 - 1, \tau_2 - 1, \cdots$, 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 recurrence points as instances. Initially, i.e., at $t = 0$, choose a fixed but arbitrary control law g_0 and use it until time $\tau_0 - 1$. Then to start off, use each of the control laws $g \in \mathcal{G}$ once each. From then at each recurrence point, compute the empirical pair measure $\rho^g_n := \{\rho^g_n(x, y) | x, y \in \mathfrak{X}\} \in M^{(2)}$ corresponding to each $g \in \mathcal{G}$ as

$$\rho^g_n(x, y) := \frac{1}{T^g_n - \tau_0} \sum_{i=\tau_0}^{n-1} 1\{g_i = g, X_i = x, X_{i+1} = y\} \quad (4.9)$$

where n is the actual time.

Define the following conditions.

C1(θ): $\rho^g_n \in \epsilon$-nbd (ν^g_θ) $\forall g \in \mathcal{G}$ and $B(\theta)$ is empty.

C2(θ): $\rho^g_n \in \epsilon$-nbd (ν^g_θ) $\forall g \in \mathcal{G}$ and $B(\theta)$ is nonempty.

C3: There does not exist $\theta \in \Theta$ such that $\rho^g_n \in \epsilon$-nbd (ν^g_θ) $\forall g \in \mathcal{G}$.

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

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

2) If C2(θ) is satisfied for some $\theta \in \Theta$, then do the following. Maintain a count of the number of instances condition C2(θ) 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 the count is even (call this situation C2(θ)a) use $g^*(\theta)$. For other instances when the count is odd (call this situation C2(θ)b) compute the likelihood ratio

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

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

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

C. Upper Bound on the Loss

In this section we derive an upper bound on the *loss* associated with the adaptive control scheme ϕ^* 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 $X_0, X_1, \cdots$ be Markovian with finite state space $\mathfrak{X}$, transition matrix P, invariant distribution π, and initial distribution p. Let $M^{(2)}$ be the unit simplex on $\mathbb{R}^{|X|^2}$ identified with the space of probability measures on $\mathfrak{X}^2$, and let $K \subset M^{(2)}$, closed, such that $\pi P \notin K$. Let $\rho_n := \{\rho_n(x, y) | x, y \in \mathfrak{X}\}$ 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: Part i) follows from the theory of large deviations. See [14, Problem IX.6.12].

$$\begin{aligned} EN &= \sum_{n=1}^{\infty} P(\rho_n \in K) \\ &\leq \sum_{n=1}^{\infty} Ae^{-an} \\ &< \infty \qquad \text{which proves ii).} \end{aligned}$$

$$\begin{aligned} EL &= E \sum_{n=1}^{\infty} 1(\exists i \geq n, \rho_i \in K) \\ &= E \sum_{n=1}^{\infty} 1 \left(\bigcup_{i \geq n} (\rho_i \in K) \right) \\ &\leq \sum_{n=1}^{\infty} \sum_{i=n}^{\infty} P(\rho_i \in K) \\ &\leq \sum_{n=1}^{\infty} \sum_{i=n}^{\infty} Ae^{-ai} \\ &< \infty \qquad \text{which proves iii).} \end{aligned}$$

□

Lemma 4.2: Let $S_n = X_1 + \cdots + X_n$ where $X_1, X_2, \cdots$ 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) $EN < \infty$;

c) $EL < \infty$.

Proof: See Hogan [15].

Lemma 4.3: Let $X_1, X_2, \cdots$ 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 $S^i_n = f^i(X_1) + f^i(X_2) \cdots + f^i(X_n)$, $L^i_A = \sum_{n=1}^{\infty} 1(\inf_{t \geq n} S^i_t \leq A)$, and $L_A = \max_{i \in I} L^i_A$. If $E(|f^i(X_1)|^2 1(f^i(X_1) \leq 0)) < \infty$ for all $i \in I$, then

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

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

$$L^i_A \leq \frac{A(1 + \epsilon)}{Ef^i(X_1)} + L^i \quad (4.11)$$

where

$$L^i = \sum_{n=1}^{\infty} 1 \left(\inf_{t \geq n} \left(S^i_t - \frac{tEf^i(X_1)}{1 + \epsilon} \right) \leq 0 \right). \quad (4.12)$$

Consider the i.i.d. r.v.'s

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

We have

$$\begin{aligned}
&E\{|Z_1^i|^2 1(Z_1^i \le 0)\} \\
&\le 2E\left\{\left[|f^i(X_1)|^2 + \left(\frac{Ef^i(X_1)}{1+\epsilon}\right)^2\right] \cdot 1\left(f^i(X_1) \le \frac{Ef^i(X_1)}{1+\epsilon}\right)\right\} \\
&\le 2E\{|f^i(X_1)|^2 1(f^i(X_1) \le 0)\} \\
&\quad + 2E\left\{|f^i(X_1)|^2 1\left(0 < f^i(X_1) \le \frac{Ef^i(X_1)}{1+\epsilon}\right)\right\} \\
&\quad + 2\left(\frac{Ef^i(X_1)}{1+\epsilon}\right)^2 \\
&< \infty.
\end{aligned}$$

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

Therefore,

$$E\left(\max_{i\in I} L^i\right) \le E\left(\sum_{i\in I} L^i\right) = \sum_{i\in I} EL^i = k(\epsilon) < \infty \quad (4.13)$$

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

Now,

$$\begin{aligned}
L_A &= \max_{i\in I} L_A^i \le \max_{i\in I}\left(\frac{A(1+\epsilon)}{Ef^i(X_1)} + L_i\right) \\
&\le \frac{A(1+\epsilon)}{\min_{i\in I}(Ef^i(X_1))} + \max_{i\in I} L^i. \qquad (4.14)
\end{aligned}$$

By (4.11) and (4.12) it follows that

$$\begin{aligned}
EL_A &\le \frac{A(1+\epsilon)}{\min_{i\in I}(Ef^i(X_1))} + k(\epsilon) \\
\limsup_{A\to\infty} \frac{EL_A}{A} &\le \frac{1+\epsilon}{\min_{i\in I}(Ef^i(X_1))}.
\end{aligned}$$

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

$$1)\ \limsup_{n\to\infty}\left[E_\theta\left[\sum_{m=1}^{\infty} 1(\lambda_{\tau_m}(\theta) \le K_{n+1})\right]/\log n\right] \le \frac{1}{\min_{\theta'\in B(\theta)} \sum_{\mathcal{G}_\theta} \beta^g(\theta)\mathfrak{I}^g_{\theta,x_0} I^g(\theta,\theta')} \quad (4.15)$$

$$2)\ P_{\theta'}\{\lambda_i(\theta) > K_{n+1} \quad \text{for some } 1 \le i \le n\} \le \frac{1}{K_{n+1}} \text{ for } \theta' \in B(\theta). \quad (4.16)$$

Proof: Let $X_0^r, X_1^r, \cdots$ be the sequence of observed states when "randomization" is used with $\alpha(\theta)$. Let $\mathfrak{X}^* = \cup_{t\ge 1}\mathfrak{X}^t$, with the Borel σ-algebra of the discrete topology, i.e., all subsets are measurable. The process $\{X_t\}_{t\ge 0}$ allows us to define $\mathfrak{X}^*$-valued random variables $B_1, B_2, \cdots$ called blocks as follows. Define the $\{\mathfrak{F}_t\}$ stopping times τ_k, $k \ge 1$ by

$$\tau_k = \inf\{t > \tau_{k-1} | X_t^r = X_0^r = x_0\}$$

with $\tau_0 = 0$. (Note that $\tau_k < \infty$ a.s.) Then

$$B_k = (X^r_{\tau_{k-1}}, X^r_{\tau_{k-1}+1}, \cdots, X^r_{\tau_k}).$$

Let $B_k' = (B_k, g_k)$. Since the same control law is used over the entire block, and the choice of the specific law for each block is made by independent randomizations at the beginning of the block, it can be easily shown that $\{B_k'\}$ are i.i.d.

Let

$$f^{\theta'}(B_k') = \log \frac{P^{g_k}(X^r_{\tau_{k-1}}, X^r_{\tau_{k-1}+1}; \theta)\cdots P^{g_k}(X^r_{\tau_k-1}, X^r_{\tau_k}; \theta)}{P^{g_k}(X^r_{\tau_{k-1}}, X^r_{\tau_{k-1}+1}; \theta')\cdots P^{g_k}(X^r_{\tau_k-1}, X^r_{\tau_k}; \theta')}.$$

Then

$$\begin{aligned}
&E_\theta[f^{\theta'}(B_k')|X_0 = x_0] \\
&= \sum_{\mathcal{G}_\theta} \beta^g(\theta) E_\theta\left[\sum_{t=\tau_{k-1}}^{\tau_k - 1} \log \frac{P^g(X_t, X_{t+1}; \theta)}{P^g(X_t, X_{t+1}; \theta')} | X_0 = x_0\right] \\
&= \sum_{\mathcal{G}_\theta} \beta^g(\theta) E_\theta\left[\sum_{x,y\in\mathfrak{X}} N(x, y, B_k) \log \frac{P^g(x, y; \theta)}{P^g(x, y; \theta')} | X_0 = x_0\right] \\
&= \sum_{\mathcal{G}_\theta} \beta^g(\theta) \sum_{x,y\in\mathfrak{X}} \pi^g(x; \theta) P^g(x, y; \theta)\mathfrak{I}^g_{\theta,x_0} \log \frac{P^g(x, y; \theta)}{P^g(x, y; \theta')} \\
&= \sum_{\mathcal{G}_\theta} \beta^g(\theta)\mathfrak{I}^g_{\theta,x_0} I^g(\theta, \theta')
\end{aligned}$$

and

$$\begin{aligned}
&E_\theta[(f^{\theta'}(B_k'))^2 1(f^{\theta'}(B_k') \le 0)|X_0 = x_0] \\
&= \sum_{\mathcal{G}_\theta} \beta^g(\theta) E_\theta[(f^{\theta'}((B_k, g)))^2 1(f^{\theta'}((B_k, g)) \le 0)|X_0 = x_0] \\
&= \sum_{\mathcal{G}_\theta} \beta^g(\theta) \sum_{B_k\in X^*} P^g(B_k; \theta|X_0 = x_0) \\
&\quad \cdot \left(\log \frac{P^g(B_k; \theta|X_0 = x_0)}{P^g(B_k; \theta'|X_0 = x_0)}\right)^2 \\
&\quad \cdot 1\left(\log \frac{P^g(B_k; \theta|X_0 = x_0)}{P^g(B_k; \theta'|X_0 = x_0)} \le 0\right) \\
&= \sum_{\mathcal{G}_\theta} \beta^g(\theta) \sum_{B_k\in X^*} P^g(B_k; \theta'|X_0 = x_0) \frac{P^g(B_k; \theta|X_0 = x_0)}{P^g(B_k; \theta'|X_0 = x_0)} \\
&\quad \cdot \left(\log \frac{P^g(B_k; \theta|X_0 = x_0)}{P^g(B_k; \theta'|X_0 = x_0)}\right)^2 \\
&\quad \cdot 1\left(\frac{P^g(B_k; \theta|X_0 = x_0)}{P^g(B_k; \theta'|X_0 = x_0)} \le 1\right) \\
&\le \sum_{\mathcal{G}_\theta} \beta^g(\theta) \sum_{B_k\in X^*} P^g(B_k; \theta'|X_0 = x_0) \frac{4}{e^2}, \\
&\quad \text{as } x(\log x)^2 \le \frac{4}{e^2} \text{ on } 0 \le x \le 1, \\
&= \frac{4}{e^2} < \infty.
\end{aligned}$$

Thus, by Lemma 4.3 we have the desired result i).

To prove ii) note that

$$\{\Lambda_i(\theta) > K_{n+1} \quad \text{for some } 1 \le i \le n\}$$

$$= \left\{ \min_{\theta' \in B(\theta)} \prod_{t=0}^{i-1} \frac{P^{g_t^r}(X_t^r, X_{t+1}^r; \theta)}{P^{g_t^r}(X_t^r, X_{t+1}^r; \theta')} > K_{n+1} \quad \text{for some } 1 \le i \le n \right\}$$

$$\subseteq \left\{ \prod_{t=0}^{i-1} \frac{P^{g_t^r}(X_t^r, X_{t+1}^r; \theta)}{P^{g_t^r}(X_t^r, X_{t+1}^r; \theta')} > K_{n+1} \quad \text{for some } 1 \le i \le n \right\}$$

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

$$\left\{ \prod_{t=0}^{i-1} \frac{P^{g_t^r}(X_t^r, X_{t+1}^r; \theta)}{P^{g_t^r}(X_t^r, X_{t+1}^r; \theta')} \right\}_{i \ge 1}$$

is an $\mathcal{F}_i$ martingale under θ' with mean 1.

Thus, the result follows by the submartingale inequality (see [13, p. 243]). □

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

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

$$E_\theta T_n^g < \infty \quad \text{if } B(\theta) \text{ is empty.} \tag{4.17}$$

Consequently,

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

$$L_n(\theta) < \infty \quad \text{if } B(\theta) \text{ is empty} \tag{4.18}$$

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

Proof: As in Section IV-B define the $\{\mathcal{F}_t (= \sigma(X_0, U_0, X_1, \cdots, X_{t-1}, U_{t-1}, X_t))\}$ stopping times $\tau_0, \tau_1, \cdots$ by $\tau_m := \inf\{t > \tau_{m-1} | X_t = x_0\}$ with $\tau_0 = \inf\{n | X_n = x_0\}$. Then $\tau_m < \infty$ a.s. Then for any $n \ge 0$, any $g \in \mathcal{G}_\theta$, we have

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

$$\le \sum_{i:\tau_i < n} 1(g_{\tau_i} = g)(\tau_{i+1} - \tau_i) + \tau_0$$

since the choice of g's is only made at the stopping times τ_i. So

$$E_\theta T_n^g \le E_\theta \sum_{i=0}^{\infty} 1(g_{\tau_i} = g)(\tau_{i+1} - \tau_i) 1(\tau_i < n) + E_\theta \tau_0$$

$$= \sum_{i=0}^{\infty} E_\theta [E_\theta [1(g_{\tau_i} = g) 1(\tau_i < n)(\tau_{i+1} - \tau_i) | \mathcal{F}_{\tau_i}]] + E_\theta \tau_0$$

$$= \sum_{i=0}^{\infty} E_\theta [1(g_{\tau_i} = g) 1(\tau_i < n) E_\theta [(\tau_{i+1} - \tau_i) | \mathcal{F}_{\tau_i}]] + E_\theta \tau_0$$

$$= \sum_{i=0}^{\infty} E_\theta [1(g_{\tau_i} = g) 1(\tau_i < n) \mathfrak{I}_{\theta, x_0}^g] + E_\theta \tau_0$$

$$= \mathfrak{I}_{\theta, x_0}^g E_\theta \sum_{i:\tau_i < n} 1(g_{\tau_i} = g) + E_\theta \tau_i.$$

Let us now examine the term $\sum_{i:\tau_i<n} 1(G_i = g)$, where $G_i = g_{\tau_i}$.

$$\sum_{i:\tau_i<n} 1(G_i = g)$$

$$= 1 + \sum_{i \ge d:\tau_i<n} 1(G_i = g)$$

$$= 1 + \sum_{i \ge d:\tau_i<n} 1\{G_i = g,\ \text{C1}(\theta') \text{ is satisfied at stage } i \text{ for some } \theta' \in \Theta\}$$

$$+ \sum_{i \ge d:\tau_i<n} 1\{G_i = g,\ \text{C2}(\theta') \text{ is satisfied at stage } i \text{ for some } \theta' \in \Theta\}$$

$$+ \sum_{i \ge d:\tau_i<n} 1\{G_i = g,\ \text{C3 is satisfied at stage } i\}$$

$$= 1 + \text{Term 1} + \text{Term 2} + \text{Term 3 (say)} \tag{4.19}$$

where C1(θ'), C2(θ'), and C3 are defined in Section IV-B and d is the cardinality of the set $\mathcal{G}$ of stationary controls. Let us now examine each term separately. Defining $\mathfrak{L}^g$ by

$$\mathfrak{L}^g := \sup_{T_n^g \ge 1} \{\rho_n^g \notin \epsilon\text{-nbd } (\nu_\theta^g)\} \tag{4.20}$$

and noting that $E_\theta \mathfrak{L}^g < \infty$ by Lemma 4.1 ii), we get Term 3 $\le \sum_{g \in \mathcal{G}} \mathfrak{L}^g$, thus,

$$E_\theta \text{ Term 3} \le \sum_{g \in \mathcal{G}} E_\theta \mathfrak{L}^g < \infty \tag{4.21}$$

and Term 1 $\le \mathfrak{L}^g$, thus,

$$E_\theta \text{ Term 1} \le E_\theta \mathfrak{L}^g < \infty. \tag{4.22}$$

$$\text{Term 2} = \sum_{i \geq d:\tau_i < n} 1\{G_i = g,\ \text{C2}(\theta') \text{ is satisfied at stage } i \text{ for some } \theta' \in \Theta \text{ such that } \nu_{\theta'}^{g^*(\theta')} \neq \nu_{\theta}^{g^*(\theta')}\}$$
$$+ \sum_{i \geq d:\tau_i < n} 1\{G_i = g,\ \text{C2}(\theta') \text{ is satisfied at stage } i \text{ for some } \theta' \in \Theta \text{ such that } \theta \in B(\theta')\}$$
$$+ \sum_{i \geq d:\tau_i < n} 1\{G_i = g,\ \text{C2}(\theta') \text{ is satisfied at stage } i \text{ for some } \theta' \in \Theta \text{ such that } \theta \in S(\theta')\}$$
$$+ \sum_{i \geq d:\tau_i < n} 1\{G_i = g,\ \text{C2}(\theta) \text{ is satisfied at stage } i\}$$
$$= \text{Term 2a} + \text{Term 2b} + \text{Term 2c} + \text{Term 2d (say)}. \tag{4.23}$$

Next we upperbound each of the Terms 2a–2d separately.

Term 2a

$$= \sum_{\substack{\theta':B(\theta') \text{ is not empty and} \\ \nu_{\theta'}^{g^*(\theta')} \neq \nu_{\theta}^{g^*(\theta')}}} \sum_{i \geq d:\tau_i < n} 1\{G_i = g,\ \text{C2}(\theta') \text{ is satisfied at stage } i\}$$
$$\leq \sum_{\substack{\theta':B(\theta') \text{ is not empty and} \\ \nu_{\theta'}^{g^*(\theta')} \neq \nu_{\theta}^{g^*(\theta')}}} \left[1 + \sum_{i \geq d:\tau_i < n} 1\{G_i = g^*(\theta'),\ \text{C2}(\theta') \text{ is satisfied at stage } i\}\right]$$
$$\leq \sum_{\substack{\theta':B(\theta') \text{ is not empty and} \\ \nu_{\theta'}^{g^*(\theta')} \neq \nu_{\theta}^{g^*(\theta')}}} (\mathfrak{L}^{g^*(\theta')} + 1). \tag{4.24}$$

The first of the inequalities of (4.24) holds because under C2(θ'), $g^*(\theta')$ is chosen on all the even instances, therefore, on at least as many instances as any other control minus one. The second of the inequalities of (4.24) holds because the sum on the left-hand side counts a subset of the times when $g^*(\theta')$ is used and $\rho_n(g^*(\theta')) \notin \epsilon\text{-nbd}\ (\nu_{\theta}^{g^*(\theta')})$ where θ is the true parameter.

By Lemma 4.1 ii) it follows that

$$E_\theta \text{ Term 2a} \leq \sum_{\substack{\theta':B(\theta') \text{ is empty and} \\ \nu_{\theta'}^{g^*(\theta')} \neq \nu_{\theta}^{g^*(\theta')}}} (1 + E_\theta \mathfrak{L}^{g^*(\theta')}) < \infty. \tag{4.25}$$

Term 2b

$$\leq \sum_{\theta':\theta \in B(\theta')} \sum_{i \geq d:\tau_i < n} 1\{\text{C2}(\theta') \text{ is satisfied at stage } i\}$$
$$\leq \sum_{\theta':\theta \in B(\theta')} 2\left[1 + \sum_{i \geq d:\tau_i < n} 1\{\text{C2}(\theta')\text{b is satisfied at stage } i\}\right]$$
$$= \sum_{\theta':\theta \in B(\theta')} 2\left[1 + \sum_{i \geq d:\tau_i < n} 1\{\text{C2}(\theta')\text{b1 is satisfied at stage } i\} + \sum_{i \geq d:\tau_i < n} 1\{\text{C2}(\theta')\text{b2 is satisfied at stage } i\}\right]$$
$$\leq \sum_{\theta':\theta \in B(\theta')} 2\left[1 + \sum_{i \geq d:\tau_i < n} 1\{\Lambda_{\tau_i}(\theta') > K_{\tau_i} + 1\} + \sum_{i \geq d:\tau_i < n} 1\{\text{C2}(\theta')\text{b2 is satisfied at stage } i\}\right]$$
$$\leq \sum_{\theta':\theta \in B(\theta')} 2\left[1 + \sum_{i=d}^{\infty} 1\{\lambda_j(\theta') > K_i \quad \text{for some } j \leq i-1\} + \sum_{i \geq d:\tau_i < n} 1\{\text{C2}(\theta')\text{b2 is satisfied at stage } i\}\right]. \tag{4.26}$$

The first of the inequalities of (4.26) results by removing the condition $G_i = g$. The second one results by observing that the total number of time instants that C2(θ') is satisfied is upperbounded by twice the odd instants that C2(θ') holds, and by noting that the first time we randomize and the other odd times we call C2(θ')b. The third inequality results because {C2(θ')b2 is satisfied at stage i} implies $\{\Lambda_{\tau_i}(\theta') > K_{\tau_i+1}\}$.

Consider now the term $\sum_{i \geq d:\tau_i < n} 1\{\text{C2}(\theta')\text{b2 is satisfied at stage } i\}$.

$$\sum_{i \geq d:\tau_i < n} 1\{\text{C2}(\theta')\text{b2 is satisfied at stage } i\}$$
$$= \sum_{i \geq d:\tau_i < n} 1\{\text{C2}(\theta')\text{b2a is satisfied at stage } i\} + \sum_{i \geq d:\tau_i < n} 1\{\text{C2}(\theta')\text{b2b is satisfied at stage } i\}$$
$$\leq 1 + 2 \sum_{i \geq d:\tau_i < n} 1\{\text{C2}(\theta')\text{b2b is satisfied at stage } i\}$$
$$= 1 + 2 \sum_{i \geq d:\tau_i < n} 1\{\text{C2}(\theta')\text{b2b is satisfied at stage } i;$$
of the number of instances that C2(θ')b2b has been satisfied so far, the fraction of instances that g' is chosen $\in (\beta^{g'}(\theta') - \epsilon, \beta^{g'}(\theta') + \epsilon)\}$
$$+ 2 \sum_{i \geq d:\tau_i < n} 1\{\text{C2}(\theta')\text{b2b is satisfied at stage } i;$$
of the number of instances that C2(θ')b2b has been satisfied so far, the fraction of instances that g' is chosen $\notin (\beta^{g'}(\theta') - \epsilon, \beta^{g'}(\theta') + \epsilon)\}$
$$\leq 1 + 2 \sum_{j=1}^{\infty} 1\{\rho_i(g') \notin \epsilon\text{-nbd}\ (\nu_{\theta}^{g'}) \text{ for some } i > (\beta^{g'}(\theta') - \epsilon)j\}$$
$$+ 2 \sum_{j=1}^{\infty} 1\{\text{Of } j \text{ the fraction of instances } g' \text{ is chosen } \notin (\beta^{g'}(\theta') - \epsilon, \beta^{g'}(\theta') + \epsilon)\} \tag{4.27}$$

where $g' \in \mathcal{G}_{\theta'}$ is such that $\nu_{\theta}^{g'} \neq \nu_{\theta'}^{g'}$.

The first of the inequalities of (4.27) results by observing that the number of instances when condition C2(θ')b2a is satisfied (i.e., the count of the number of instances C2(θ')b2 is satisfied is a perfect square) is upperbounded by the number of

instances when condition C2(θ')b2b is satisfied plus one. Consider now changing the index of summation to the instances when randomization is done. Then the condition C2(θ')b2b, along with the condition that the fraction of instances that g' is chosen $\in (\beta^{g'}(\theta') - \epsilon, \beta^{g'}(\theta') + \epsilon)$ at stage i, imply that $\rho_i(g') \notin \epsilon$-nbd $(\nu_\theta^{g'})$ for some $i > (\beta^{g'}(\theta') - \epsilon)j$. By extending the summation to the infinity together with the above observation establishes the last of the inequalities of (4.27).

Thus, by Lemma 4.1 i) and (4.16) it follows that

$$E_\theta \text{ Term 2b} \le \sum_{\theta':\theta\in B(\theta')} 2\left[1+\sum_{i=d}^{\infty}(i(\log i)^p)^{-1}+1 + 2\sum_{j=1}^{\infty}\sum_{i>(\beta^{g'}(\theta')-\epsilon)j} A_1 e^{-a_1 i} + 2\sum_{j=1}^{\infty} A_2 e^{-a_2 j}\right] < \infty \quad (4.28)$$

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

Term 2c

$$= \sum_{\theta':\theta\in S(\theta')}\sum_{i\ge d:\tau_i<n} 1\{G_i = g,\ \text{C2}(\theta') \text{ is satisfied at stage } i\}$$

$$\le \sum_{\theta':\theta\in S(\theta')}\left[1+\sum_{i\ge d:\tau_i<n} 1\{G_i = g, \times \text{C2}(\theta')\text{b2 is satisfied at stage } i\}\right]$$

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

$$\le \sum_{\theta':\theta\in S(\theta')}\left[1+l^2+\sum_{j=1}^{\infty} 1\{\rho_j(g')\notin \epsilon\text{-nbd }(\nu_\theta^{g'})\}(2j+1)l^2\right] \quad (4.29)$$

where $g' \in \mathcal{G}(S(\theta'))$ is such that $\nu_\theta^{g'} \neq \nu_{\theta'}^{g'}$ and $\#\mathcal{G}(S(\theta')) = l$.

The first inequality of (4.29) results by noting that since $\theta \in S(\theta')$, $g \neq g^*(\theta') = g^*(\theta)$ can be chosen only when condition C2(θ')b2 is satisfied, or at the first instance when C2(θ') is true. The second inequality results by removing the requirement $G_i = g$. The third inequality results by upperbounding the number of instances condition C2(θ')b2 is satisfied. This can be achieved as follows. First restrict attention to those instances that are perfect squares and the control g' is used. At these instances since C2(θ') is satisfied $\rho_n(g') \in \epsilon$-nbd $(\nu_{\theta'}^{g'})$, thus, by the choice of $g' \in \mathcal{G}(S(\theta'))$, $\rho_n(g') \notin \epsilon$-nbd $(\nu_\theta^{g'})$. Consider the sum of the intervals between the above instances. (Note that the length of the jth interval is upperbounded by $[(j+1)^2 - j^2]l^2 = (2j+1)l^2$.) Then the number of instances condition C2(θ')b2 is satisfied cannot exceed this sum. Finally, the inequality results by changing the summation index to all the times when g' is used and upperbounding the interval following the time $\rho_j(g') \notin \epsilon$-nbd $(\nu_\theta^{g'})$ by $(2j+1)l^2$. Again, by using Lemma 4.1 i) we get

$$E_\theta \text{ Term 2c} \le \sum_{\theta':\theta\in S(\theta')}\left[1+l^2+\sum_{j=1}^{\infty} Ae^{-aj}\cdot(2j+1)l^2\right] < \infty. \quad (4.30)$$

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

$$\text{Term 2d} = 0. \quad (4.31)$$

Otherwise,

$$\text{Term 2d} = \sum_{i\ge d:\tau_i<n} 1\{G_i = g,\ \text{C2}(\theta) \text{ is satisfied at stage } i\}$$

$$\le 1 + \sum_{i\ge d:\tau_i<n} 1\{G_i = g,\ \text{C2}(\theta)\text{b2 is satisfied at stage } i\}$$

$$= 1 + \sum_{i\ge d:\tau_i<n} 1\{G_i = g,\ \text{C2}(\theta)\text{b2a is satisfied at stage } i\} + \sum_{i\ge d:\tau_i<n} 1\{G_i = g,\ \text{C2}(\theta)\text{b2b is satisfied at stage } i\}$$

$$\le 2 + \sum_{i\ge d:\tau_i<n} 1\{G_i = g,\ \text{C2}(\theta)\text{b2b is satisfied at stage } i\} + \left(\sum_{i\ge d:\tau_i<n} 1\{\text{C2}(\theta)\text{b2b is satisfied at stage } i\}\right)^{1/2}. \quad (4.32)$$

The first of the inequalities of (4.32) is obtained by noting $g \neq g^*(\theta)$ can be chosen only at the first instance when C2(θ) is satisfied (in which case randomization is done) or when C2(θ)b2 is satisfied. The last of the inequalities of (4.32) results because the number of instances condition C2(θ)b2a is satisfied is upperbounded by one plus the square root of the number of instances C2(θ)b2b is satisfied.

To upperbound E_θ Term 2d we use (4.32), Jensen's inequality, and the following fact. At each instance i when condition C2(θ)b2b is satisfied, the choice of the control action $G_i \in \mathcal{G}_\theta$ is made by an independent randomization $\beta(\theta)$. Then,

$$E_\theta \text{ Term 2d} \le 2 + \sum_{i\ge d:\tau_i<n} P_\theta\{\text{C2}(\theta)\text{b2b is satisfied at stage } i\}\cdot\beta^g(\theta) + \left(\sum_{i\ge d:\tau_i<n} P_\theta\{\text{C2}(\theta)\text{b2b is satisfied at stage } i\}\right)^{1/2}$$

$$\le 2 + \beta^g(\theta)E_\theta[\sup\{1\le i\le n | \lambda_i(\theta)\le K_{n+1}\}] + (E_\theta[\sup\{1\le i\le n|\lambda_i(\theta)\le K_{n+1}\}])^{1/2}. \quad (4.33)$$

Using (4.15) we get

$$\limsup_{n\to\infty} E_\theta \text{ Term 2d}/\log n \le \frac{\beta^g(\theta)}{\min_{\theta'\in B(\theta)}\sum_{\mathcal{G}_\theta}\beta^g(\theta)\mathfrak{I}^g_{\theta,x_0}I^g(\theta,\theta')}. \quad (4.34)$$

Combining (4.19), (4.21)–(4.23), (4.25), (4.28), (4.30), (4.31), and (4.34) we get (4.17). Equation (4.18) follows easily from (4.17) and (2.21). □

In view of Theorems 3.1 and 4.2, the adaptive control scheme ϕ^* that we constructed in Section IV-B is asymptotically efficient, i.e.,

$$L_n(\theta) \sim \min_{\alpha\in\mathcal{G}_\theta}\max_{\theta'\in B(\theta)}\frac{\sum_{\mathcal{G}_\theta}\alpha^g d_\theta(g)}{\sum_{\mathcal{G}_\theta}\alpha^g I^g(\theta,\theta')}\log n \quad \text{if } B(\theta) \text{ is nonempty}$$

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

IV. 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. Multiarmed 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 multiarmed bandit problem, like it 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 multiarmed 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 multiarmed 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 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 optimally the conflict between learning and control.

References

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

[2] P. Mandl, "Estimation and control in Markov chains," *Adv. Appl. Prob.*, vol. 6, pp. 40–60, 1974.

[3] V. Borkar and P. Varaiya, "Adaptive control of Markov chains, I: Finite parameter set," *IEEE Trans. Automat. Contr.*, vol. AC-24, pp. 953–958, 1979.

[4] ——, "Identification and adaptive control of Markov chains," *SIAM J. Contr. Optimiz.*, vol. 20, pp. 470–489, 1982.

[5] P. R. Kumar and A. Becker, "A new family of optimal adaptive controllers for Markov chains," *IEEE Trans. Automat. Contr.*, vol. AC-27, pp. 137–146, 1982.

[6] P. R. Kumar and W. Lin, "Optimal adaptive controllers for unknown Markov chains," *IEEE Trans. Automat. Contr.*, vol. AC-27, pp. 765–774, 1982.

[7] R. A. Milito and J. B. Cruz, "An optimization oriented approach to the adaptive control of Markov chains," *IEEE Trans. Automat. Contr.*, vol. AC-32, Sept. 1987.

[8] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotically efficient allocation schemes for controlled I.I.D. processes: Finite parameter space," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 258–267, Mar. 1989.

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

[10] ——, "Asymptotically optimal allocation of treatments in sequential experiments," in *Design of Experiments,* T. J. Santuer and A. C. Tamhane, Eds. New York: Marcel-Dekker.

[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 Trans. Automat. Contr.*, vol. AC-32, pp. 968–982, Nov. 1987.

[12] 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.

[13] S. Ross, *Stochastic Processes*. New York: Wiley, 1983.

[14] R. S. Ellis, *Entropy, Large Deviations, and Mechanics*. New York: Springer-Verlag, 1985.

[15] M. Hogan, "Moments of the minimum of a random walk and complete convergence," Dep. Statistics, Stanford Univ., Stanford, CA, Tech. Rep. 21, Jan. 1983.

Rajeev Agrawal (M'89), for a photograph and biography, see p. 267 of the March 1989 issue of this Transactions.

Demosthenis Teneketzis, for a photograph and biography, see p. 267 of the March 1989 issue of this Transactions.

Venkatachalam Anantharam (M'86), for a photograph and biography, see p. 267 of the March 1989 issue of this Transactions.