SIAM J. CONTROL OPTIM.
Vol. 35, No. 3, pp. 715–743, May 1997

ASYMPTOTICALLY EFFICIENT ADAPTIVE CHOICE OF CONTROL LAWS IN CONTROLLED MARKOV CHAINS*

TODD L. GRAVES† AND TZE LEUNG LAI‡

Abstract. We consider a controlled Markov chain on a general state space whose transition probabilities are parameterized by an unknown parameter belonging to a compact metric space. There is a one-step reward associated with each pair of control and the following state of the process. Given a finite set of stationary control laws, under each of which the Markov chain is uniformly recurrent, an optimal control law in this set is one that maximizes the long-run average reward. In ignorance of the parameter value, we construct an adaptive control rule which uses the optimal control law(s) at a relative frequency of $1-O(n^{-1}\log n)$ and show that this relative frequency gives an asymptotically optimal balance between the control objective and the amount of information needed to learn about the unknown parameter. The basic idea underlying this construction is to introduce suitable "uncertainty adjustments" via sequential testing theory into the certainty-equivalence rule, thus resolving the apparent dilemma between control and information.

Key words. adaptive control of Markov chains, martingales, likelihood ratios, stationary distributions, certainty equivalence, sequential testing, multiarmed bandits

AMS subject classifications. 93C40, 93E20, 93E35, 60J20, 62L10

PII. S0363012994275440

1. Introduction and background.

We consider here a controlled Markov chain $\{X_n, n \geq 0\}$ on a measurable state space $(S, \mathcal{A})$, with a general control set U and a parametric family of transition density functions $p(x, y; u, \theta)$ with respect to some measure M on S, where θ is an unknown parameter taking values in a compact metric space Θ. Thus the transition probability measure under control action u and parameter θ is given by $P_\theta^u(X_{n+1} \in A | X_n = x) = \int_A p(x, y; u, \theta) dM(y)$. The initial distribution of X_0 under P_θ^u is also assumed to be absolutely continuous with respect to M. Let $G = \{g_1, \ldots, g_L\}$ be a finite set of stationary control laws $g_j : S \to U$ such that for every $g \in G$, the transition probability function $\{P_\theta^{g(x)}(x, A) : x \in S, A \in \mathcal{A}\}$ is irreducible with respect to some maximal irreducibility measure and has stationary distribution $\{\pi_\theta^g(A) : A \in \mathcal{A}\}$. Let $r(X_t, u_t)$ represent the one-step reward at time t, where $r : S \times U \to \mathbf{R}$, and define the long-run average reward

$$\mu_\theta(g) = \int r(x, g(x)) d\pi_\theta^g(x), \tag{1.1}$$

which will be assumed to be finite. If θ were known, then one would use the stationary control law $g_{j(\theta)}$ such that

$$\mu_\theta^* := \max_{g \in G} \mu_\theta(g) = \mu_\theta(g_{j(\theta)}). \tag{1.2}$$

*Received by the editors October 11, 1994; accepted for publication (in revised form) February 26, 1996.

http://www.siam.org/journals/sicon/35-3/27544.html

†National Institute of Statistical Sciences, P.O. Box 14162, Research Triangle Park, NC 27709-4162 (graves@niss.rti.org). The research of this author was supported by the National Science Foundation.

‡Department of Statistics, Stanford University, Stanford, CA 94305 (karola@playfair.stanford.edu). The research of this author was supported by the National Science Foundation and the National Security Agency.

In ignorance of θ, a certainty-equivalence control rule is to use the control law $g_{j(\hat{\theta}_t)}$ at time t, where $\widehat{\theta}_t$ is an estimate of θ based on the observed data $X_0, u_0, \ldots, X_{t-1}, u_{t-1}, X_t$ (in chronological order).

For the case of a finite state space S, Mandl [18] studied this certainty-equivalence rule in which $\widehat{\theta}_t$ is a minimum contrast estimate and showed that $\widehat{\theta}_t$ converges almost surely (a.s.) to θ under a restrictive "identifiability condition" and some other regularity conditions. Borkar and Varaiya [6] removed this identifiability condition and showed that when Θ is finite, the maximum likelihood estimate $\widehat{\theta}_t$ converges a.s. to a random variable θ^* such that

$$p(x, y; g_{j(\theta^*)}(x), \theta^*) = p(x, y, g_{j(\theta^*)}(x), \theta) \tag{1.3}$$

for all $x, y \in S$ (finite). They also gave an example for which $\theta^* \neq \theta$ with positive probability, showing that the certainty-equivalence rule can prematurely converge to a wrong parameter value so that it eventually uses only the suboptimal stationary control law $g_j(\theta^*)$ to the exclusion of other control laws.

In view of this difficulty with the certainty-equivalence rule, various modifications of the rule have appeared in the literature. Kumar [11] and Kumar and Varaiya [12] have provided comprehensive surveys of the developments up to the mid-1980s, which include (i) forced choice schemes that reserve some prespecified sparse set of times for experimentation with all stationary control laws in G, (ii) randomization schemes for which every $g \in G$ has a positive probability, whose value is to be determined adaptively from the past data, of being applied at each time, and (iii) using penalized (cost-biased) maximum likelihood estimators $\widehat{\theta}_t$.

1.1. Ideas from bandit theory. The past decade has witnessed other developments in a classical example of adaptive choice from a finite set of control actions, namely, the multiarmed bandit problem. In its simplest form, the problem can be described as follows. There are L statistical populations $\Pi_1, \ldots, \Pi_L$ with univariate density functions $p(y; \theta_1), \ldots, p(y; \theta_L)$ with respect to some measure M. At each time t we can sample from one of these populations, and the reward is the sampled value X_t. Thus the control set U is $\{1, \ldots, L\}$, where control action j refers to sampling from Π_j. An adaptive sampling rule consists of a sequence of random variables $u_1, u_2, \ldots$ taking values in $\{1, \ldots, k\}$ such that the event $\{u_t = j\}$ ("X_{t+1} is sampled from Π_j") belongs to the σ-field generated by $u_0, X_1, u_1, \ldots, X_{t-1}, u_{t-1}, X_t$. Let $\theta = (\theta_1, \ldots, \theta_L)$. If θ were known, then we would sample from the population $\Pi_{j(\theta)}$ with the largest mean; i.e., $\mu_\theta^* := \max_{1 \le j \le L} \mu_\theta(j) = \mu_\theta(j(\theta))$, where $\mu_\theta(j) = \int y p(y; \theta_j) dM(y)$ is assumed to be finite. In ignorance of θ, the problem is to sample $X_1, X_2, \ldots$ sequentially from the k populations to maximize $E_\theta(\sum_{i=1}^n X_i)$, or equivalently to minimize the regret

$$R_n(\theta) = n\mu_\theta^* - E_\theta\left(\sum_{i=1}^n X_i\right) = \sum_{j:\mu_\theta(j)<\mu_\theta^*} (\mu_\theta^* - \mu_\theta(j)) E_\theta T_n(j) \tag{1.4}$$

as $n \to \infty$, where $T_n(j) = \sum_{t=1}^n I_{\{u_{t-1}=j\}}$ and $I_A = 1$ if A occurs, $I_A = 0$ otherwise. Lai and Robbins [16] showed how to construct sampling rules for which $R_n(\theta) = O(\log n)$ at every θ. These rules are called "uniformly good." They also developed asymptotic lower bounds for the regret $R_n(\theta)$ of uniformly good rules and showed that the rules constructed actually attain these asymptotic lower bounds and are therefore asymptotically efficient. Specifically, they showed that under certain

regularity conditions

$$\liminf_{n\to\infty} R_n(\theta)/\log n \geq c(\theta) \tag{1.5}$$

for uniformly good rules and gave an explicit formula for $c(\theta)$ in terms of $\mu_\theta^* - \mu_\theta(j)$ and certain Kullback–Leibler information numbers. A more general representation of the lower bound $c(\theta)$ is given in section 2, where we extend this result on the multiarmed bandit problem to the general setting of adaptive choice of stationary control laws in controlled Markov chains.

Anantharam, Varaiya, and Walrand [5] generalized the results of [16] to the multiarmed bandit problem in which each Π_j represents an aperiodic, irreducible Markov chain on a finite state space S so that successive observations from Π_j are no longer independent but are governed by the Markov transition density $p(x, y; \theta_j)$. Assuming the successive observations from Π_j to be independent with a common density function $p(y; \theta_j)$, Agrawal, Hedge, and Teneketzis [1] incorporated an additional switching cost and showed that the sampling rules of [16] can be modified by sampling in blocks so that the asymptotic lower bound in (1.5) is still attained and the cumulative switching cost up to time n is of the order $o(\log n)$ when no more than one population has the largest mean μ_θ^*.

For the problem of adaptive choice of stationary control laws in controlled Markov chains, switching costs are particularly relevant since it usually takes time to change from a new control strategy to another. We shall assume no switching cost for switching among the (typically equivalent) optimal stationary control laws that attain the maximum in (1.2) and a cost $a(\theta)$ for each switch from one $g \in G$ to another $g' \in G$ when g and g' are not both optimal. An *adaptive control rule* ϕ is a sequence of random variables $\phi_1, \phi_2, \ldots$ taking values in G such that $\{\phi_t = g\} \in \mathcal{F}_t$ for all $g \in G$ and $t \geq 0$, where

$$\mathcal{F}_t = \sigma\text{-field generated by } X_0, \phi_0, \ldots, X_{t-1}, \phi_{t-1}, X_t. \tag{1.6}$$

Defining $\mu_\theta(g)$ and μ_θ^* by (1.1) and (1.2), we generalize (1.4) to controlled Markov chains by letting

$$R_n(\theta) = \sum_{g\in G:\mu_\theta(g)<\mu_\theta^*} (\mu_\theta^* - \mu_\theta(g)) E_\theta T_n(g), \text{ with } T_n(g) = \sum_{i=0}^{n-1} I_{\{\phi_i = g\}}. \tag{1.7}$$

In view of the additional switching cost $a(\theta)$ for each switch between two control laws in G, not both optimal, we define the overall regret to be $R_n(\theta) + a(\theta)S_n(\theta)$, where

$$S_n(\theta) = E_\theta \left(\sum_{i=1}^{n} I_{\{\phi_i \neq \phi_{i-1}, \min(\mu_\theta(\phi_i), \mu_\theta(\phi_{i-1})) < \mu_\theta^*\}} \right). \tag{1.8}$$

An adaptive control rule ϕ is said to be *uniformly good* if

$$R_n(\theta) = O(\log n) \text{ and } S_n(\theta) = o(\log n) \text{ for every } \theta \in \Theta. \tag{1.9}$$

In section 2 we develop an asymptotic lower bound for $R_n(\theta)$ among all uniformly good rules, and in section 3 we construct adaptive control rules that attain this lower bound. These results therefore generalize those of [16] on the multiarmed bandit problem to the setting of adaptive choice of control laws in controlled Markov chains.

A major technical difficulty in this generalization is that unlike Markovian bandit processes in which the state of Π_j is "frozen" until a new observation is sampled from Π_j, for controlled Markov chains X_{t+1} is governed by the immediately preceding state X_t and control action $\phi_t(X_t)$ irrespective of whether $\phi_{t+1} = \phi_t$ or not. We resolve this difficulty by using certain change-of-measure arguments in section 2 and some limit theorems for controlled Markov chains developed in section 4.

This difficulty disappears in the special case where the controlled Markov chain $\{X_t, t \geq 1\}$ is a sequence of independent random variables so that the conditional density of X_{t+1} given $u_t = u$ is $p(y; u, \theta)$. Assuming U and Θ to be finite, Agrawal, Teneketzis, and Anantharam [3] studied this special case by regarding each control action $u \in U$ as an arm and $r(X_1, u_1), r(X_2, u_2), \ldots$ as a sequence of rewards obtained by choosing the arms $u_1, u_2, \ldots$. They noted, however, another difficulty in reducing this problem to the multiarmed bandit problem because of the differences in how the parameter space Θ is defined in the two problems. In the controlled independent sequence problem, θ parameterizes all the arms $u \in U$, whereas in the multiarmed bandit problem $\theta = (\theta_1, \ldots, \theta_k)$ with each θ_j parameterizing the individual arm Π_j. They circumvented this difficulty by making use of the finiteness of Θ and introducing a finite set $B(\theta)$ of "bad" parameter values associated with θ. They thereby obtained an asymptotic lower bound for the regret (1.7) of uniformly good control rules and developed a rule that attains this bound. In section 2, without assuming Θ to be finite, we define the bad set $B(\theta)$ in the setting of controlled Markov chains with general state and parameter spaces. When the state space S, the control set U, and the parameter space Θ are all finite, Agrawal, Teneketzis, and Ananthanam [4] developed a "translation scheme" which together with the construction of an "extended probability space" enabled them to solve the controlled Markov chain problem by converting it to a form similar to that for the controlled independent sequence problem in [3]. This ingenious idea of translation schemes, however, depends heavily on the finiteness of S. Our development of an asymptotic lower bound for (1.7) in section 2 uses a different approach which involves large deviation probabilities for controlled Markov chains on general state spaces S satisfying certain uniform recurrence assumptions.

As a consequence of the translation scheme under their finiteness assumptions, Agrawal, Teneketzis, and Ananthanam [4] obtained the approximation

$$E_\theta \left\{ \sum_{i=0}^{n-1} r(X_i, \phi_i(X_i)) \right\} = \sum_{g \in G} \mu_\theta(g) E_\theta T_n(g) + O(1) \text{ as } n \to \infty. \tag{1.10}$$

Hence in this case (1.7) can be expressed as

$$R_n(\theta) = \widetilde{R}_n(\theta) + O(1), \text{ where } \widetilde{R}_n(\theta) = n\mu_\theta^* - E_\theta \left\{ \sum_{i=0}^{n-1} r(X_i, \phi_i(X_i)) \right\}. \tag{1.11}$$

Note that $\widetilde{R}_n(\theta)$ is the shortfall between the long-run cumulative reward using the optimal stationary control law $g_{j(\theta)}$ and the cumulative reward of the adaptive control rule ϕ. Moreover, by making use of the translation scheme in the development of their asymptotic lower bound for $R_n(\theta)$, Agrawal, Teneketzis, and Ananthanam [4] did not need to impose the constraint on the expected number of switches in (1.9) for uniformly good rules. However, for general state spaces, (1.10) and (1.11) need no longer hold, and there may even exist adaptive control rules for which $\lim_{n\to\infty} \widetilde{R}_n(\theta) = -\infty$ at certain values of θ. This difficulty arises because in the absence of (1.10), the long-run average optimality property (1.2) of the stationary control law $g_{j(\theta)}$ no longer

ensures it to be asymptotically optimal among adaptive control rules that can switch freely among stationary control laws in G. Note that G does not contain such adaptive control rules which are not stationary. We therefore have to put some constraint on the expected number of switches in the adaptive control rules to compare them with the optimal stationary control law $g_{j(\theta)}$ (which makes no switch in G). This can be regarded as a "complexity constraint," consistent with our basic assumption of a *finite* set of *stationary* control policies to reduce the complexity of the Markov control problem. Under the switching constraint that $S_n(\theta) = o(\log n)$ and assuming the transition probability function $\{P_\theta^{g(x)}(x, A) : x \in S, A \in \mathcal{A}\}$ to be uniformly recurrent for every $g \in G$, it is shown in [14] that the "reward regret" $\widetilde{R}_n(\theta)$ is asymptotically equivalent to the more tractable weighted sum (1.7) of expected frequencies of using suboptimal stationary control laws; i.e.,

$$\widetilde{R}_n(\theta) = R_n(\theta) + o(\log n) \text{ as } n \to \infty. \tag{1.12}$$

The constraint on $S_n(\theta)$ in (1.9) relates only to switches between two stationary control laws which are not both optimal when θ is the true parameter. We do not impose the $o(\log n)$ constraint on the expected number of switches between two optimal stationary control laws. In fact, since one cannot infer from the past data which of these optimal stationary control laws is significantly inferior, one is expected to keep switching among them to learn their performance, as in [16] for the multiarmed bandit problem.

1.2. Uncertainty adjustments to the certainty-equivalence rule via sequential testing theory. Lai [13] pointed out the usefulness of sequential testing theory in making uncertainty adjustments of the certainty-equivalence rule, leading to asymptotically optimal rules when the control set is finite. To illustrate this, he considered the following bivariate bandit problem. Let Π_1, Π_2, Π_3 be three bivariate normal populations with respective mean vectors $(\mu_1, \xi), (\mu_2, \mu_3)$, and $(\mu_3, \mu_2 + \xi)$ and with a common known covariance matrix which is equal to the identity matrix. Here $\theta = (\mu_1, \mu_2, \mu_3, \xi)$ is the unknown parameter vector and the problem is to sample $X_1, X_2, \ldots$ sequentially from the three populations in order to maximize the expected value of the first component of $\sum_{i=1}^n X_i$ as $n \to \infty$. The relevant information we need for optimal control can be represented by the three hypotheses $H_j : \mu_j = \max(\mu_1, \mu_2, \mu_3),\ j = 1, 2, 3$. In other words, we do not need to know the actual values of μ_1, μ_2, μ_3, ξ but need only to determine which of μ_1, μ_2, μ_3 is the largest. While information about μ_1 can only be obtained by sampling from Π_1, information about μ_2 and μ_3 can be obtained by sampling from Π_2 alone or from Π_3 and Π_1. Using results from sequential testing theory, Lai [13] constructed an asymptotically optimal rule whose regret (1.7) satisfies $R_n(\theta) = O(1)$ if $\mu_1 = \max(\mu_2, \mu_3)$ and $R_n(\theta) \sim c(\theta) \log n$ otherwise, where

$$\begin{aligned} c(\theta) &= 2/\{\mu_1 - \max(\mu_2, \mu_3)\} \text{ if } \mu_1 > \max(\mu_2, \mu_3) \\ &= 2/(\mu_2 - \mu_1) \text{ if } \mu_2 > \max(\mu_1, \mu_3) \\ &= 2/(\mu_3 - \mu_1) \text{ if } \mu_3 = \mu_2 > \mu_1 \text{ or } \mu_3 > \mu_1 \geq \mu_2 \\ &= 2/(\mu_3 - \mu_1) + 2/(\mu_3 - \mu_2) - 2(\mu_3 - \mu_2)/(\mu_3 - \mu_1)^2 \text{ if } \mu_3 > \mu_2 > \mu_1. \end{aligned}$$

In section 3 we use sequential testing theory to construct asymptotically efficient adaptive control rules in controlled Markov chains. These rules are considerably

simpler than those in [4] which require finiteness of S and Θ for their implementation and for the analysis that shows their regret $R_n(\theta)$ to be of the order $O(\log n)$. The rules in section 3 are applicable to general state spaces S and compact metric spaces Θ, and we prove in section 4 that they attain the asymptotic lower bound $(c(\theta)+o(1))\log n$ for the regret established in section 2.

In summary, by making use of bandit theory and sequential testing methodology, we generalize herein previous work of Agrawal, Tenektetzis, and Ananthanam [4] from the case of finite Θ and S to compact Θ and general state spaces S while still assuming finiteness of G, which is crucial for both the asymptotic lower bound in section 2 and the rules proposed in section 3. This generalization requires certain constraints on the expected number of switches among the stationary control laws in G and uniform recurrence assumptions on the transition probability functions. We construct in section 3 adaptive control rules with regret $R_n(\theta)$ having the asymptotically minimal order $c(\theta)\log n$, where the constant $c(\theta)$ is given in section 2. Using nonparametric sequential testing theory instead of the parametric likelihood ratio approach here and assuming G to be countable instead of finite, Lai and Yakowitz [17] removed the parametric and related assumptions herein and developed adaptive control rules with regret $R_n(\theta)=O(\alpha_n\log n)$ for any given nondecreasing sequence of positive numbers $\alpha_n\to\infty$ and $\alpha_{2n}=O(\alpha_n)$. Earlier, Agrawal and Teneketzis [2] also used a nonparametric approach to construct adaptive control rules with regret $R_n(\theta)=O((\log n))^{1+\delta})$ for any given $\delta>0$ in the case of finite G, Θ, and S so that the translation scheme of Agrawal, Teneketzis, and Ananthanam [4] is applicable.

2. Decomposition of the parameter space and an asymptotic lower bound for the regret of uniformly good rules. Using the same notation as that introduced at the beginning of section 1, define for $g\in G$ the Kullback–Leibler information number

$$I^g(\theta,\lambda)=\int\int\left\{\log\frac{p(x,y;g(x),\theta)}{p(x,y;g(x),\lambda)}\right\}p(x,y;g(x),\theta)dM(y)d\pi_\theta^g(x), \tag{2.1}$$

which will be assumed to be finite for all $\theta,\lambda\in\Theta$. We shall decompose Θ as the union of L subsets: $\Theta=\Theta_1\cup\cdots\cup\Theta_L$, where

$$\Theta_j=\{\theta\in\Theta:\mu_\theta(g_j)=\max_{g\in G}\mu_\theta(g)\}; \tag{2.2}$$

i.e., g_j is an optimal stationary control law if $\theta\in\Theta_j$. For $\theta\in\Theta$, let

$$J(\theta)=\{1\le j\le L:\mu_\theta(g_j)=\max_{g\in G}\mu_\theta(g)(=\mu_\theta^*)\}, \tag{2.3}$$

$$B(\theta)=\left\{\lambda\in\Theta:\ \lambda\notin\bigcup_{j\in J(\theta)}\Theta_j \text{ and } I^{g_j}(\theta,\lambda)=0 \text{ for all } j\in J(\theta)\right\}, \tag{2.4}$$

(2.5)

$$c(\theta)=\inf\left\{\sum_{j\notin J(\theta)}c_j[\mu_\theta^*-\mu_\theta(g_j)]:c_j\in[0,\infty),\ \inf_{\lambda\in B(\theta)}\sum_{j\notin J(\theta)}c_jI^{g_j}(\theta,\lambda)\ge 1\right\}(\inf\emptyset=\infty).$$

Thus, $\{g_j, j\in J(\theta)\}$ is the set of all optimal stationary control laws when θ is the true parameter value, and $B(\theta)$ consists of all "bad" parameter values $\lambda\notin\cup_{j\in J(\theta)}\Theta_j$

which are statistically indistinguishable from θ if one only uses the optimal control laws $g_j, j \in J(\theta)$, because $I^{g_j}(\theta, \lambda) = 0$. Theorem 1 below shows that under certain regularity conditions $(c(\theta) + o(1)) \log n$ is an asymptotic lower bound for the regret (1.7) of uniformly good rules. Note that (2.5) can also be expressed as

(2.6)
$$c(\theta) = \inf \left\{ \frac{\sum_{j \notin J(\theta)} \alpha_j [\mu_\theta^* - \mu_\theta(g_j)]}{\inf_{\lambda \in B(\theta)} \sum_{j \notin J(\theta)} \alpha_j I^{g_j}(\theta, \lambda)} \ : \ \alpha_j \geq 0, \ \sum_{j \notin J(\theta)} \alpha_j = 1 \right\} \ (\inf \emptyset = \infty).$$

This alternative form of the asymptotic lower bound of $R_n(\theta)/\log n$ was obtained by Agrawal, Teneketzis, and Ananthanam [4] when the state space S and the parameter space Θ are finite. Theorem 1 uses a different argument which involves the equivalent form (2.5) of (2.6) to establish the result for general state spaces and compact parameter spaces. We first give some examples to illustrate the computation of $c(\theta)$.

Example 1. Consider the multiarmed bandit problem of section 1.1. Here $\theta = (\theta_1, \cdots, \theta_L)$, $g_j = j$ ("sample from Π_j") and $I^j(\theta, \lambda) = I(\theta_j, \lambda_j)$, where

$$I(a, b) = \int p(y; a) \log[p(y; a)/p(y; b)] dM(y), \ \mu(a) = \int y p(y; a) dM(y). \tag{2.7}$$

Assume that $I(a, b) < \infty$ and that $I(a, b) = 0$ iff $\mu(a) = \mu(b)$, analogous to the assumptions (1.6) and (1.7) of Lai and Robbins [16].

(i) Suppose $L = 2, \Theta = \{(\alpha, \beta), (\beta, \alpha)\}$, and $\mu(\alpha) \neq \mu(\beta)$. Thus, it is known that one population has a specified parameter value α and the other has parameter value β, but it is not known whether Π_1 or Π_2 is associated with α. This is the two-armed bandit problem studied by Feldman [7]. Here $I^1((\alpha, \beta), (\beta, \alpha)) = I(\alpha, \beta) > 0, I^2((\alpha, \beta), (\beta, \alpha)) = I(\beta, \alpha) > 0$, and therefore $B(\theta) = \emptyset, c(\theta) = 0$ for $\theta \in \Theta$. In fact, Feldman's procedure has regret $R_n(\theta) = O(1)$. Lai and Robbins [15] considered more general k-armed bandit problems in which $B(\theta) = \emptyset$ for all $\theta \in \Theta$ and developed sampling rules with $R_n(\theta) = O(1)$.

(ii) Suppose $\Theta = \Delta^L$, where Δ is a compact metric space. For $\theta = (\theta_1, \ldots, \theta_L)$, let $\theta^* \in \{\theta_1, \ldots, \theta_L\}$ be such that $\mu(\theta^*) = \max_{1 \leq i \leq L} \mu(\theta_i)$. Then $J(\theta) = \{1 \leq j \leq L : \mu(\theta_j) = \mu(\theta^*)\}$ and

(2.8)
$$B(\theta) = \left\{ (\lambda_1, \ldots, \lambda_L) \in \Theta : \mu(\lambda_j) = \mu(\theta^*) \text{ for all } j \in J(\theta), \max_{1 \leq i \leq L} \mu(\lambda_i) > \mu(\theta^*) \right\}$$

since $I(a, b) = 0$ iff $\mu(a) = \mu(b)$ by assumption. Assume as in (1.7) of [16] that

$$I(\theta_j, b) \to I(\theta_j, \theta^*) \quad \text{as} \quad \mu(b) \downarrow \mu(\theta^*). \tag{2.9}$$

Consider the minimization problem in (2.5) which reduces here to finding nonnegative numbers $c_j, j \notin J(\theta)$, to minimize $\sum_{j \notin J(\theta)} c_j(\mu(\theta^*) - \mu(\theta_j))$ subject to the constraints

$$\inf_{\lambda \in B_i(\theta)} \sum_{j \notin J(\theta)} c_j I(\theta_j, \lambda_j) \geq 1, \ \ i \notin J(\theta), \tag{2.10}$$

where $B_i(\theta) = \{\lambda \in B(\theta) : \mu(\lambda_i) = \max_{1 \leq s \leq L} \mu(\lambda_s)\}$. For fixed $i \notin J(\theta)$, since $(\theta_1, \ldots, \theta_{i-1}, b, \theta_{i+1}, \ldots, \theta_L) \in B_i(\theta)$ for any $b \in \Delta$ with $\mu(b) > \mu(\theta^*)$, (2.9) and

(2.10) imply that $c_i \geq 1/I(\theta_i, \theta^*)$. Hence $c(\theta) \geq \sum_{j \notin J(\theta)} (\mu(\theta^*) - \mu(\theta_j))/I(\theta_j, \theta^*)$, which is the asymptotic lower bound for $R_n(\theta)/\log n$ given in [16], where sampling rules that attain this lower bound are also constructed for certain parametric families having the monotonicity property

$$I(a, b) \geq I(a, \theta^*) \quad \text{whenever} \quad \mu(b) \geq \mu(\theta^*) \geq \mu(a). \tag{2.11}$$

Under (2.11), $I(\theta_i, \lambda_i)/I(\theta_i, \theta^*) \geq 1$ for all $\lambda \in B_i(\theta)$, and therefore the constraint (2.10) holds with $c_j = 1/I(\theta_j, \theta^*)$, $j \notin J(\theta)$. This choice of c_j therefore solves the minimization problem in (2.5) under the assumptions (2.9) and (2.11), yielding $c(\theta) = \sum_{j \notin J(\theta)} (\mu(\theta^*) - \mu(\theta_j))/I(\theta_j, \theta^*)$.

Example 2. Consider the following variant of Example 1. Let Π_1, Π_2, Π_3 be three univariate normal populations with respective means $\gamma, \xi + 1$, and ξ^2 and common variance 1, where γ and ξ are unknown parameters. Here $\Theta = \{\theta = (\gamma, \xi) : -\infty < \gamma < \infty, -\infty < \xi < \infty\}$, and the problem is to sample $X_1, X_2, \ldots$ sequentially from the three populations to maximize the expected value of $\sum_{i=1}^n X_i$ as $n \to \infty$. Therefore, as in Example 1, $g_j = j$ ("sample from Π_j"), $I^1((\gamma, \xi), (\widetilde{\gamma}, \widetilde{\xi})) = (\gamma - \widetilde{\gamma})^2/2, I^2((\gamma, \xi), (\widetilde{\gamma}, \widetilde{\xi})) = (\xi - \widetilde{\xi})^2/2, I^3((\gamma, \xi), (\widetilde{\gamma}, \widetilde{\xi})) = (\xi^2 - \widetilde{\xi}^2)^2/2$, and

$$\begin{aligned} B(\gamma, \xi) &= \{(\widetilde{\gamma}, \widetilde{\xi}) : \max(\widetilde{\xi} + 1, \widetilde{\xi}^2) > \gamma\} \text{ if } \gamma > \max(\xi + 1, \xi^2) \\ &= \{(\widetilde{\gamma}, \xi) : \widetilde{\gamma} > \xi + 1\} \text{ if } \xi + 1 > \max(\gamma, \xi^2) \\ &= \{(\widetilde{\gamma}, \widetilde{\xi}) : |\widetilde{\xi}| = |\xi|, \max(\widetilde{\gamma}, \widetilde{\xi} + 1) > \xi^2\} \text{ if } \xi^2 > \max(\gamma, \xi + 1). \end{aligned}$$

To compute $c(\theta)$, we can use arguments similar to those in Example 1 to show that

$$c(\gamma, \xi) = 2/(\xi + 1 - \gamma) \text{ if } \xi + 1 > \max(\gamma, \xi^2). \tag{2.12}$$

The case $\gamma > \max(\xi + 1, \xi^2)$ is considerably more complicated, and it is more convenient to use the representation (2.6), which reduces to

$$c(\gamma, \xi) = \inf_{0 \leq \pi \leq 1} \frac{\pi(\gamma - \xi - 1) + (1 - \pi)(\gamma - \xi^2)}{\inf_{\widetilde{\xi}: \widetilde{\xi} + 1 > \gamma \text{ or } \widetilde{\xi}^2 > \gamma} \pi(\xi - \widetilde{\xi})^2/2 + (1 - \pi)(\xi^2 - \widetilde{\xi}^2)^2/2}. \tag{2.13}$$

To solve the minimization problem in (2.13), first fix $\pi \in [0, 1]$ and find $\widetilde{\xi}_\pi$ to minimize $\psi_\pi(\widetilde{\xi}) := \pi(\xi - \widetilde{\xi})^2/2 + (1 - \pi)(\xi^2 - \widetilde{\xi}^2)^2/2$ subject to $\widetilde{\xi} \geq \gamma - 1$ or $|\widetilde{\xi}| \geq \sqrt{\gamma}$. Then find $\pi(\gamma, \xi) \in [0, 1]$ that minimizes $\{\pi(\gamma - \xi - 1) + (1 - \pi)(\gamma - \xi^2)\}/\psi_\pi(\widetilde{\xi}_\pi)$. Note that $d\psi_\pi/d\widetilde{\xi} = -(\xi - \widetilde{\xi})\{\pi + 2(1 - \pi)\widetilde{\xi}(\xi + \widetilde{\xi})\}$, which has zeroes at $\widetilde{\xi} = \xi$ and $\widetilde{\xi} = \frac{1}{2}\{-\gamma + [\gamma^2 - 2\pi/(1 - \pi)]^{1/2}\}$. For example, if $(\gamma, \xi) = (1.69, -1)$, then $\pi(\gamma, \xi) = 0.112$. This is in sharp contrast to (2.12) or Example 1, for which the optimizing π is always 0 or 1 if we use the representation (2.6) to evaluate $c(\theta)$. In section 3 we shall consider the case $\xi^2 > \max(\gamma, \xi + 1)$.

In the following theorem we use the same notation as that introduced at the beginning of section 1. We shall assume that the transition probability function $\{P_\theta^{g(x)}(x, A) : x \in S, A \in \mathcal{A}\}$ is uniformly recurrent for every $\theta \in \Theta$ and $g \in G$; i.e., there exist positive constants $a_\theta^g < b_\theta^g$ such that

$$a_\theta^g \leq p(x, y; g(x), \theta) \leq b_\theta^g \text{ for almost every (with respect to } M) \ x \text{ and } y \tag{2.14}$$

(cf. [9]). This implies that for every $g \in G, \theta \in \Theta$, and $\lambda \in B(\theta)$, there exist positive constants $\alpha_{\theta,\lambda}^g < \beta_{\theta,\lambda}^g$ such that

$$\alpha_{\theta,\lambda}^g \leq p(x, y; g(x), \theta)/p(x, y; g(x), \lambda) \leq \beta_{\theta,\lambda}^g \text{ for } (M\text{-})\text{almost every } x \text{ and } y. \tag{2.15}$$

We consider situations where there are switching costs, for which "uniformly good" rules are defined by (1.9). The following theorem gives an asymptotic lower bound for the regret (1.7) of uniformly good rules.

THEOREM 1. *Under* (2.14), *for any uniformly good rule* ϕ,

$$\liminf_{n\to\infty} \sum_{j\notin J(\theta)} I^{g_j}(\theta,\lambda) E_\theta T_n(g_j)/\log n \geq 1 \text{ for every } \lambda \in B(\theta) \tag{2.16}$$

and therefore

$$\liminf_{n\to\infty} R_n(\theta)/\log n \geq c(\theta) \tag{2.17}$$

for every $\theta \in \Theta$.

Proof. Since $R_n(\theta) = \sum_{j\notin J(\theta)} [\mu_\theta^* - \mu_\theta(g_j)] E_\theta T_n(g_j)$ by (1.7), (2.17) follows from (2.5) and (2.16) (writing $E_\theta T_n(g_j) = c_{j,n} \log n$ and noting that $\inf \emptyset = \infty$). To prove (2.16), it suffices to show that for every $\lambda \in B(\theta)$ and $\epsilon > 0$,

$$\lim_{n\to\infty} P_\theta \left\{ \sum_{j\notin J(\theta)} I^{g_j}(\theta,\lambda) T_n(g_j) \geq (1-\epsilon)\log n \right\} = 1. \tag{2.18}$$

The proof of (2.18) uses a change-of-measure argument similar to that in the proof of Theorem 2 of Lai and Robbins [16] on the multiarmed bandit problem. Since $\lambda \in B(\theta)$, $\lambda \notin \cup_{j\in J(\theta)} \Theta_j$ and therefore $J(\lambda) \cap J(\theta) = \emptyset$. Since ϕ is uniformly good, $E_\lambda\{n - \sum_{i\in J(\lambda)} T_n(g_i)\} = O(\log n)$ by (1.9). For $g \in G$, if

$$0 = I^g(\theta,\lambda) = \int\int p(x,y;g(x),\theta) \log[p(x,y;g(x),\theta)/p(x,y;g(x),\lambda)] dM(y) d\pi_\theta^g(x), \tag{2.19}$$

then $p(x,y;g(x),\theta) = p(x,y;g(x),\lambda)$ for M-almost everywhere (a.e.) x and y (noting that $d\pi_\theta^g/dM > 0$ a.e.$[M]$ by (2.14)), and therefore $\mu_\theta(g) = \mu_\lambda(g)$. Since $\lambda \in B(\theta), \mu_\lambda^* > \mu_\lambda(g_i)$ and $I^{g_i}(\theta,\lambda) = 0$ for all $i \in J(\theta)$. Hence $\mu_\theta(g_i) = \mu_\lambda(g_i) < \mu_\lambda^*$ for all $i \in J(\theta)$, implying that $\mu_\lambda^* > \mu_\theta^*$. For $j \in J(\lambda), \mu_\lambda(g_j) = \mu_\lambda^* > \mu_\theta^* \geq \mu_\theta(g_j)$ and therefore $I^{g_j}(\theta,\lambda) > 0$. It then follows that for all large n,

$$\begin{aligned}
&P_\lambda \left\{ \sum_{j\notin J(\theta)} I^{g_j}(\theta,\lambda) T_n(g_j) < (1-\epsilon)\log n \right\} \\
&\leq P_\lambda \left\{ \sum_{j\in J(\lambda)} I^{g_j}(\theta,\lambda) T_n(g_j) < (1-\epsilon)\log n \right\} \\
&\leq P_\lambda \left\{ \sum_{j\in J(\lambda)} T_n(g_j) \leq n/2 \right\} = P_\lambda \left\{ n - \sum_{j\in J(\lambda)} T_n(g_j) \geq n/2 \right\} \\
&\leq 2n^{-1} E_\lambda \left\{ n - \sum_{j\in J(\lambda)} T_n(g_j) \right\} = O(n^{-1}\log n).
\end{aligned} \tag{2.20}$$

Let $L_n = \sum_{i=0}^{n-1} \log[p(X_i, X_{i+1}, \phi_i(X_i), \theta)/p(X_i, X_{i+1}; \phi_i(X_i), \lambda)]$ and let $0 < \delta < \epsilon/2$. Note that

(2.21)
$$\begin{aligned} &P_\lambda \left\{ \sum_{j \notin J(\theta)} I^{g_j}(\theta, \lambda) T_n(g_j) < (1-\epsilon)\log n, L_n \le (1-\delta)\log n \right\} \\ &= \int_{\left\{\sum_{j \notin J(\theta)} I^{g_j}(\theta,\lambda) T_n(g_j) < (1-\epsilon)\log n, L_n \le (1-\delta)\log n\right\}} e^{-L_n} dP_\theta \\ &\ge e^{-(1-\delta)\log n} P_\theta \left\{ \sum_{j \notin J(\theta)} I^{g_j}(\theta, \lambda) T_n(g_j) < (1-\epsilon)\log n, L_n \le (1-\delta)\log n \right\}. \end{aligned}$$

Combining (2.20) and (2.21) yields

$$\lim_{n\to\infty} P_\theta \left\{ \sum_{j \notin J(\theta)} I^{g_j}(\theta, \lambda) T_n(g_j) < (1-\epsilon)\log n, L_n \le (1-\delta)\log n \right\} = 0. \tag{2.22}$$

Let $h_g(x, y) = \log[p(x, y; g(x), \theta)/p(x, y; g(x), \lambda)]$. For $-\infty < \alpha < \infty$, define the measure

$$\widehat{P}_{x,\alpha,g}(A) = \int_A e^{\alpha h_g(x,y)} p(x, y; g(x), \theta) M(dy), \quad A \in \mathcal{A},$$

and define the linear operator $\widehat{P}_g(\alpha)$ on the space of bounded measurable functions $f : S \to \mathbf{R}$ by $\widehat{P}_g(\alpha) f(x) = \int f(y) \widehat{P}_{x,\alpha,g}(dy)$. In view of (2.14) and (2.15), $\widehat{P}_g(\alpha)$ has a maximal simple real eigenvalue $\rho_g(\alpha)$, with associated right eigenfunction $r_g(\cdot; \alpha) : S \to (0, \infty)$ and left eigenmeasure $\ell_g(\cdot; \alpha) : \mathcal{A} \to [0, \infty)$ normalized so that $\int r_g(x; \alpha) \ell_g(dx; \alpha) = 1$; moreover, $r_g(\cdot; \alpha)$ is bounded and uniformly positive for every fixed α (cf. [10]). For $j \in J(\theta)$, since $\lambda \in B(\theta)$, it follows that $I^{g_j}(\theta, \lambda) = 0$, and therefore by (2.19), $p(x, y; g_j(x), \theta) = p(x, y; g_j(x), \lambda)$; i.e., $h_g(x, y) = 0$, for M-a.e. x and y. Hence

$$L_n = \sum_{i=0}^{n-1} I_{\{\phi_i \notin G_J\}} \log[p(X_i, X_{i+1}; \phi_i(X_i), \theta)/p(X_i, X_{i+1}; \phi_i(X_i), \lambda)] \quad \text{a.s. } [P_\theta],$$

where $G_J = \{g_j : j \in J(\theta)\}$, recalling that the initial distribution of X_0 under P_θ is assumed to be absolutely continuous with respect to M.

Let $\Lambda_g(\alpha) = \log \rho_g(\alpha)$ and define a new probability measure Q_α on the controlled Markov chain by the "twisting" transformation (cf. [9], [10]):

$$Q_\alpha(B) = E_\theta \left\{ I_B \prod_{0 \le i < n : \phi_i \notin G_J} \frac{r_{\phi_i}(X_{i+1}; \alpha)}{r_{\phi_i}(X_i; \alpha)} e^{-\Lambda_{\phi_i}(\alpha) + \alpha h_{\phi_i}(X_i, X_{i+1})} \right\}, \quad B \in \mathcal{F}_n,$$

where $\Pi_{i \in \emptyset} = 1$ and $\mathcal{F}_n$ is the σ-field defined in (1.6). Letting

$$\begin{aligned} B = \Bigg\{ &\sum_{j \notin J(\theta)} I^{g_j}(\theta, \lambda) T_n(g_j) < (1-\epsilon)\log n, \quad L_n > (1-\delta)\log n, \text{ and} \\ &\sum_{i=1}^{n} I_{\{\phi_i \ne \phi_{i-1}, \phi_i \notin G_J\}} \le \delta \log n \Bigg\}, \end{aligned} \tag{2.23}$$

and noting that $L_n = \sum_{0\le i<n:\phi_i\notin G_J} h_{\phi_i}(X_i, X_{i+1})$, it then follows that for $\alpha > 0$,

(2.24)
$$\begin{aligned} &P_\theta(B) \\ &= \int_B \left\{ \prod_{0\le i<n:\phi_i\notin G_J} \frac{r_{\phi_i}(X_i;\alpha)}{r_{\phi_i}(X_{i+1};\alpha)} \right\} \exp\left\{ -\alpha L_n + \sum_{0\le i<n:\phi_i\notin G_J} \Lambda_{\phi_i}(\alpha) \right\} dQ_\alpha \\ &\le e^{-\alpha(1-\delta)\log n} \int_B \left\{ \prod_{0\le i<n:\phi_i\notin G_J} \frac{r_{\phi_i}(X_i;\alpha)}{r_{\phi_i}(X_{i+1};\alpha)} \right\} \exp\left\{ \sum_{j\notin J(\theta)} \Lambda_{g_j}(\alpha) T_n(g_j) \right\} dQ_\alpha \end{aligned}$$

since $L_n > (1-\delta)\log n$ on B. For $j \notin J(\theta)$,

$$\Lambda_{g_j}(0) = 0,\ (d\Lambda_{g_j}/d\alpha)(0) = \int\int h_{g_j}(x,y) p(x,y;g_j(x),\theta) M(dy) \pi_\theta^{g_j}(dx) = I^{g_j}(\theta,\lambda).$$

Therefore, we can choose $\alpha > 0$ sufficiently small so that $\Lambda_{g_j}(\alpha)/\alpha \le (1+\epsilon/2) I^{g_j}(\theta,\lambda)$. Since $\sum_{j\notin J(\theta)} I^{g_j}(\theta,\lambda) T_n(g_j) < (1-\epsilon)\log n$ on B, it then follows that

$$\sum_{j\notin J(\theta)} \Lambda_{g_j}(\alpha) T_n(g_j) < \alpha(1+\epsilon/2)(1-\epsilon)\log n < \alpha(1-\epsilon/2)\log n \text{ on } B. \tag{2.25}$$

Noting that $C := \max_{g\in G} \sup_{x\in S} r_g(x;\alpha) < \infty, D := \min_{g\in G} \inf_{x\in S} r_g(x;\alpha) > 0$ and that $\sum_{i=1}^n I_{\{\phi_i\ne\phi_{i-1},\phi_i\notin G_J\}} \le \delta \log n$ on B, we obtain that

$$\prod_{0\le i<n:\phi_i\notin G_J} \{r_{\phi_i}(X_i,\alpha)/r_{\phi_i}(X_{i+1},\alpha)\} \le (C/D)^{\delta\log n+1} \quad \text{on} \quad B. \tag{2.26}$$

Indeed, letting $i_1 < \cdots < i_m$ denote the elements of $\{1 \le i \le n : \phi_i \ne \phi_{i-1}, \phi_i \notin G_J\}$, we have $\phi_0 = \phi_1 = \cdots = \phi_{i_1-1}, \phi_{i_1} = \phi_{i_1+1} = \cdots = \phi_{i_2-1}, \ldots, \phi_{i_m} = \phi_{i_m+1} = \cdots = \phi_n$, and therefore the left-hand side of (2.26) can be expressed as

$$\{r_{\phi_0}(X_0,\alpha)/r_{\phi_{n-1}}(X_n,\alpha)\} \prod_{t=1}^m \{r_{\phi_{i_t}}(X_{i_t},\alpha)/r_{\phi_{i_t-1}}(X_{i_t},\alpha)\},$$

from which (2.26) follows since $m \le \delta \log n$ on B.

From (2.24)–(2.26), it follows that by choosing δ sufficiently small,

(2.27)
$$P_\theta(B) \le CD^{-1} \exp(\{-\alpha(1-\delta) + \alpha(1-\epsilon/2) + \delta\log(C/D)\}\log n) \to 0 \text{ as } n\to\infty.$$

Since ϕ is uniformly good, (1.9) yields that

(2.28)
$$\begin{aligned} &P_\theta\left\{ \sum_{i=1}^n I_{\{\phi_i\ne\phi_{i-1},\phi_i\notin G_J\}} > \delta\log n \right\} \\ &\quad\le E_\theta\left(\sum_{i=1}^n I_{\{\phi_i\ne\phi_{i-1},\phi_i\notin G_J \text{ or } \phi_{i-1}\notin G_J\}} \right) / (\delta\log n) \to 0 \end{aligned}$$

as $n\to\infty$. From (2.22), (2.23), (2.27), and (2.28), the desired conclusion (2.18) follows.

3. Construction of asymptotically efficient rules. The main idea behind the adaptive control rule ϕ^* presented in this section is to introduce suitable "uncertainty adjustments" into the certainty-equivalence rule that uses the control law $g_{j(\widehat{\theta}_t)}$ at time t, where $j(\theta)$ is defined in (1.2) and $\widehat{\theta}_n$ is the following weighted maximum likelihood estimate of θ at time n:

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} L_n(\theta),$$
$$L_n(\theta) = \sum_{g\in G}(T_n(g))^{-1} \sum_{1\le t\le n, \phi^*_{t-1}=g} \log p(X_{t-1}, X_t; g(X_{t-1}), \theta), \tag{3.1}$$

if the maximizer in (3.1) exists, as is the case when p is a continuous function of θ, since Θ is assumed to be compact. If the maximizer in (3.1) does not exist, then we define $\widehat{\theta}_n$ as an ϵ_n-maximizer of $L_n(\theta)$ in the sense that $L_n(\widehat{\theta}_n) \ge \sup_{\theta\in\Theta} L_n(\theta) - \epsilon_n$, where ϵ_n are positive numbers such that $\lim_{n\to\infty}\epsilon_n = 0$. The asymptotic lower bounds in section 2 provide valuable insights into how the uncertainty adjustments should be made and quantify the need for experimentation with the inferior control laws. In particular, they suggest that the total amount of experimentation with an inferior control law g_j up to time n should be at least of the order $\{c_j(\theta)+o(1)\}\log n$, where the $c_j(\theta)$ solve the minimization problem that defines $c(\theta)$ in (2.5); i.e.,

$$c(\theta) = \sum_{j\notin J(\theta)} c_j(\theta)[\mu^*_\theta - \mu_\theta(g_j)] \text{ and } \inf_{\lambda\in B(\theta)} \sum_{j\notin J(\theta)} c_j(\theta) I^{g_j}(\theta,\lambda) = 1 \text{ if } B(\theta)\neq\emptyset, \tag{3.2}$$
$$c_j(\theta) = 0 \text{ for all } j\notin J(\theta) \text{ if } B(\theta)=\emptyset.$$

Example 1 (continuation). We have shown in the multiarmed bandit problem of Example 1(ii) that $c_j(\theta) = 1/I(\theta_j,\theta^*)$ if $j\notin J(\theta)$. This suggests that to achieve the asymptotic lower bound $(c(\theta)+o(1))\log n$ for the regret, the sampling rule should take $(1/I(\theta_j,\theta^*)+o(1))\log n$ observations, up to stage n, from an inferior population Π_j to determine whether it is indeed inferior. If the sampling rule should indeed attain the asymptotic lower bound, then it would take $n - O(\log n)$ observations, up to stage n, from the population with mean $\mu(\theta^*)$, so we can regard the value of $\mu(\theta^*)$ as known with relatively negligible uncertainty in this case. The problem of determining whether Π_j has a larger mean than $\mu(\theta^*)$ then becomes that of testing the null hypothesis $H_j : \mu(\theta_j) > \mu(\theta^*)$. The theory of optimal stopping and sequential analysis shows that subject to the constraint that the probability of rejecting H_j when it is true be $\le\alpha$, the expected number of observations from Π_j of a sequential test under the alternative hypothesis is at least $\{1/I(\theta_j,\theta^*)+o(1)\}|\log\alpha|$ as $\alpha\to 0$, and there are sequential tests based on generalized likelihood ratio statistics or mixture likelihood ratio statistics that attain this asymptotic lower bound for the expected sample size. The construction of asymptotically efficient sampling rules in section 4 of [16] uses this sequential testing theory, with $|\log\alpha| \sim |\log n|$, although the procedure is described there in terms of certain "upper confidence bounds."

Example 2 (continuation). Here the $c_j(\theta)$ are considerably more complicated than those in Example 1. The means of the normal populations Π_1,Π_2,Π_3 are $\gamma, \xi+1$, and ξ^2, involving only two unknown parameters γ (which has to be learned from Π_1) and ξ (which can be learned from Π_2 or Π_3). If $\xi+1>\gamma$ and $\xi+1\ge\xi^2$, sampling from the superior population Π_2 would give information about the means of both Π_2 and Π_3, and therefore the same argument as that in Example 1 yields $c_1(\gamma,\xi) = 2/(\xi+1-\gamma)^2$, $c_3(\gamma,\xi)=0$, which in turn gives (2.12) as the solution of the minimization problem (2.5).

In the case $\xi^2 > \gamma$ and $\xi^2 \geq \xi+1$, sampling from the superior population Π_3 would give information about $|\xi|$ but not about the sign of ξ, which has to be learned from Π_2. If $\xi^2 \geq \xi + 1$ and $\xi^2 \geq -\xi + 1$ or, equivalently, if $\xi \notin ((-1-\sqrt{5})/2, (1-\sqrt{5})/2)$, then $B(\gamma,\xi) = \{(\widetilde{\gamma},\widetilde{\xi}) : |\widetilde{\xi}| = |\xi|, \max(\widetilde{\gamma},\widetilde{\xi}+1) > \xi^2\} = \{(\widetilde{\gamma},\xi) : \widetilde{\gamma} > \xi^2\}$, and the same argument as in Example 1 yields $c(\gamma,\xi) = 2/(\xi^2-\gamma)$, $c_1(\gamma,\xi) = 2/(\xi^2-\gamma)^2$, $c_2(\gamma,\xi) = 0$. On the other hand, if $-\xi+1 > \xi^2$, then $B(\gamma,\xi) = \{(\widetilde{\gamma},\widetilde{\xi}) : |\widetilde{\xi}| = |\xi|, \widetilde{\gamma} > \xi^2$ or $\widetilde{\xi}+1 > \xi^2\}$, and putting this in (2.5) yields

$$c_1(\gamma,\xi) = \frac{1}{(\xi^2-\gamma)^2/2}, \quad c_2(\gamma,\xi) = \frac{1}{(2\xi)^2/2}, \quad c(\gamma,\xi) = \frac{2}{\xi^2-\gamma} + \frac{\xi^2-\xi-1}{2\xi^2}. \tag{3.3}$$

In the case $\gamma > \max(\xi+1, \xi^2)$, (2.13) yields

$$c_2(\gamma,\xi) = \pi(\gamma,\xi)/\psi_{\pi(\gamma,\xi)}(\widetilde{\xi}_{\pi(\gamma,\xi)}), \quad c_3(\gamma,\xi) = (1-\pi(\gamma,\xi))/\psi_{\pi(\gamma,\xi)}(\widetilde{\xi}_{\pi(\gamma,\xi)}), \tag{3.4}$$

where $\psi_\pi(\widetilde{\xi})$, $\widetilde{\xi}_\pi$, and $\pi(\gamma,\xi)$ are defined in the two sentences following (2.13). For the case $\gamma = \xi + 1 \geq \xi^2$, sampling from Π_2 will give information about ξ, from which one can learn that Π_3 has mean ξ^2, and $B(\gamma,\xi) = \emptyset$ in this case. If $\gamma = \xi^2 \geq \xi + 1$ with $\xi \notin (-(1-\sqrt{5})/2, (1-\sqrt{5})/2)$, then $\xi^2 \geq \xi + 1$ and $\xi^2 \geq -\xi + 1$, and knowledge of ξ^2 will show that Π_2 does not have a larger mean than Π_3, so $B(\gamma,\xi) = \emptyset$ in this case. If $\gamma = \xi^2 > \xi + 1$ and $(-1-\sqrt{5})/2 < \xi < (1-\sqrt{5})/2$, then $-\xi + 1 > \xi^2$ and $B(\gamma,\xi) = \{(\gamma,-\xi)\}$, so putting this in (2.5) yields

$$J(\gamma,\xi) = \{1,3\}, \quad c_2(\gamma,\xi) = 2/(2\xi)^2, \quad c(\gamma,\xi) = (\xi^2-\xi-1)/(2\xi^2). \tag{3.5}$$

The main idea behind the uncertainty adjustments, presented below, to certainty-equivalence rules in controlled Markov chains is to apply sequential testing theory to assess whether an inferior-looking control law is indeed inferior on the basis of all the current and past observations. We shall use sequential likelihood ratio tests of composite hypotheses in general stochastic systems to test the null hypothesis that θ belongs to Θ_i, with prescribed error probability of wrongly rejecting the null hypothesis when it is true and with asymptotically minimal expected waiting time to reject the null hypothesis when it is false. In the present context, the "waiting time" has to be interpreted broadly as a weighted sum of the number of times that an inferior control law g_j is used. Because of switching costs and because of the technical difficulties in controlled Markov chains due to the change of the transition probability function P^g whenever the control law is changed, we shall designate blocks of successive times to use control law g_j for an entire block if the sequential likelihood ratio test performed at the beginning of the block does not reject the hypothesis that θ belongs to Θ_j.

Since θ is unknown, it is natural to replace $c_j(\theta)$ by $c_j(\widehat{\theta}_t)$, where $\widehat{\theta}_t$ is the weighted maximum likelihood estimate defined in (3.1), as in the "certainty-equivalent" testing phase of the control scheme described below. This certainty equivalence approach raises the question concerning how well $c_j(\widehat{\theta}_t)$ approximates $c_j(\theta)$. When one does not have enough information to estimate θ well, an alternative approach is to ignore the constants $c_j(\theta)$ and to experiment equally with each stationary control law, as is done in the following control scheme during its "evenly allocated" testing phase. The control scheme takes an integer $a \geq 2$ and initializes with a common control law for times $1, \ldots, a$.

Outline of control scheme between times a^i and a^{i+1}. Let n_i be positive integers such that

$$n_i \sim i/\log i \quad \text{as} \quad i \to \infty. \tag{3.6}$$

For fixed i, we now describe our control scheme at times $n \in \{a^i+1, \ldots, a^{i+1}\}$, which we partition into $m(i) := [(a^{i+1}-a^i)/n_i]$ blocks of consecutive integers, each block of length n_i except possibly the last one whose length may range from n_i to $2n_i - 1$. Label these blocks as $B_1^i, \ldots, B_{m(i)}^i$ so that the mth block begins at time $\nu_i(m) := a^i + 1 + (m-1)n_i$ for $1 \le m \le m(i)$. To begin with, at time a^i compute the weighted maximum likelihood estimate $\widehat{\theta}_{a^i}$ of θ. To this estimate corresponds a set $\{g_j : j \in J(\widehat{\theta}_{a^i})\}$ of apparently optimal stationary control laws, where $J(\theta)$ is defined in (2.3). We use $c_j(\widehat{\theta}_{a^i})$ to define below the "certainty-equivalent" testing phase (during the period from time a^i+1 to a^{i+1}), whose objective is to test sequentially whether $\theta \notin \cup_{j \in J(\widehat{\theta}_{a^i})} \Theta_j$. The certainty-equivalent testing phase is continued until we either (i) switch to the "evenly allocated" testing phase or (ii) terminate testing and use the same (apparently optimal) stationary control law up to time a^{i+1}. This adaptive control rule will be denoted by ϕ^*. Let $\mathcal{C}_i$ denote the set of times belonging to all those blocks B_m^i that begin with certainty-equivalent testing (at $\nu_i(m)$). Let

$$\mathcal{C} = \bigcup_{s=1}^{\infty} \mathcal{C}_s, \quad \tau_n(g) = \sum_{t=0}^{n-1} I_{\{t \in \mathcal{C}, \phi_t^* = g\}} \quad \text{for} \quad g \in G. \tag{3.7}$$

Thus, $\tau_n(g)$ is the total number of times $t < n$, within these certainty-equivalent-tested blocks, that use the control law $g \in G$. For $t \in \mathcal{C}$, let

$$\widehat{G}_{J,t} = \{g_j : j \in J(\widehat{\theta}_{a^s})\} \quad \text{if} \quad t \in \mathcal{C}_s, \tag{3.8}$$

which is the set of apparently optimal stationary control laws used for the certainty-equivalent test (3.11) below.

The certainty-equivalent testing phase. During the first L or fewer blocks of the certainty-equivalent testing phase, we use in succession the stationary control laws $g_j (j \in L)$ with $\tau_{a^i}(g_j) < n_i$. Suppose $\{1 \le j \le L : j \notin J(\widehat{\theta}_{a^i})$ and $\tau_{a^i}(g_j) < (\log a^i)[c_j(\widehat{\theta}_{a^i}) \wedge \log i]\} = \{j_1, \ldots, j_N\}$. The next blocks of stages use g_{j_1} until time $\nu_i(m_1) - 1$ and then use g_{j_2} until time $\nu_i(m_2) - 1$, etc., where

$$m_k = \inf\{m > m_{k-1} : \tau_{\nu_i(m)}(g_{j_k}) \ge (\log a^i)[c_{j_k}(\widehat{\theta}_{a^i}) \wedge \log i]\}, \quad k = 1, \ldots, N. \tag{3.9}$$

For $m \ge m_N$, alternate using the stationary control laws g_j (one for each block of consecutive times) that satisfy either (i) $j \notin J(\widehat{\theta}_{a^i})$ and

$$\tau_{\nu_i(m)}(g_j) \le (2\log a^i)\{c_j(\widehat{\theta}_{a^i}) \wedge \log i\} + n_i \ \text{ and } \ g_j \ \text{ has not been eliminated} \tag{3.10a}$$

or (ii) $j \in J(\widehat{\theta}_{a^i})$ and

$$\tau_{\nu_i}(m)(g_j) \le (2\log a^i)\log i + n_i. \tag{3.10b}$$

Sequential testing of the hypotheses $H_j : \theta \in \Theta_j$, $j \notin J(\widehat{\theta}_{a^i})$, is performed at times $\nu_i(m)$ with $m \ge m_N$, and we eliminate g_j from further use through time a^{i+1} once the hypothesis H_j is rejected. Rejection of H_j occurs at the first time $n = \nu_i(m)$ with $m \ge m_N$ when

$$\inf_{\lambda \in \Theta_j} \max \left\{ \frac{\int \Pi_{1 \le t \le n, t-1 \in \mathcal{C}, \phi_{t-1}^* \notin \widehat{G}_{J,t}} p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \theta) dF(\theta)}{\Pi_{1 \le t \le n, t-1 \in \mathcal{C}, \phi_{t-1}^* \notin \widehat{G}_{J,t}} p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \lambda)}, \right.$$
$$\left. \frac{\int \Pi_{1 \le t \le n, T_{t-1}(\phi_{t-1}^*) \ge a^{i-1}/L} \, p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \theta) dF(\theta)}{\Pi_{1 \le t \le n, T_{t-1}(\phi_{t-1}^*) \ge a^{i-1}/L} \, p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \lambda)} \right\} \ge ia^i, \tag{3.11}$$

where $\inf \emptyset = \infty$, $\Pi_{t=\emptyset} = 1$, F is a probability measure on Θ such that $F(A) > 0$ for all open subsets A of Θ, and $\mathcal{C}$ and $\widehat{G}_{J,t}$ are defined in (3.7) and (3.8). Certainty-equivalent testing is terminated when (3.10a) fails for all $j \notin J(\widehat{\theta}_{a^i})$. If only one stationary control law g_{j^*} is not eliminated at the termination of certainty-equivalent testing, we use g_{j^*} up to time a^{i+1}. Otherwise we switch to the evenly allocated testing phase.

The evenly allocated testing phase. This testing phase does not use the maximum likelihood estimate $\widehat{\theta}_{a^i}$, its associated set $J(\widehat{\theta}_{a^i})$, and the estimates $c_j(\widehat{\theta}_{a^i})$ that have been used in (3.9)–(3.11). Sequential testing of the hypotheses $H_j : \theta \in \Theta_j$ is performed at the times $\nu_i(m)$ for those g_j not yet eliminated (during the times $\nu_i(m')$ between a^i and a^{i+1} with $m' < m$, which include the times of certainty-equivalent testing) in succession in ascending order of j, and we eliminate g_j from further use through time a^{i+1} once the hypothesis H_j is rejected. We reject H_j at the test time $n = \nu_i(m)$ if

$$\inf_{\lambda \in \Theta_j} \max \left\{ \frac{\int \Pi_{1 \le t \le n, \phi^*_{t-1} = g_j} p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \theta) dF(\theta)}{\Pi_{1 \le t \le n, \phi^*_{t-1} = g_j} p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \lambda)} , \right. \tag{3.12}$$

$$\left. \frac{\int \Pi_{1 \le t \le n, T_{t-1}(\phi^*_{t-1}) \ge a^{i-1}/L} \, p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \theta) dF(\theta)}{\Pi_{1 \le t \le n, T_{t-1}(\phi^*_{t-1}) \ge a^{i-1}/L} \, p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \lambda)} \right\} \ge i a^i.$$

The "even" allocation rule alternates using the stationary control laws that have not been eliminated, one for each block B^i_m of consecutive times. The testing phase terminates as soon as all except one stationary control law have been eliminated, and we use the remaining stationary control law up to time a^{i+1}. For example, a typical pattern of the evenly allocated phase, sampling from four control laws labeled $1, 2, 3, 4$, is

$$1 \cdots 1 \; 2 \cdots 2 \; 3 \cdots 3 \; 4 \cdots 4 \; 1 \cdots 1 \; 2 \cdots 2_{\uparrow} 3 \cdots 3 \; 4 \cdots 4 \; 1 \cdots 1 \; 3 \cdots 3 \; 4 \cdots 4_{\downarrow} 1 \cdots 1 \; 3 \cdots 3 \; 1 \cdots 1_*,$$

where $\uparrow$ denotes the time at which 2 is eliminated, $\downarrow$ denotes the time at which 4 is eliminated, and $*$ denotes the time at which the testing phase terminates with the elimination of 1, leaving behind only the control law 3.

In the certainty-equivalent testing phase, we consider the maximum of two mixture likelihood ratio statistics instead of combining them into a single mixture likelihood ratio because we have different roles in mind for the two statistics. One of them has the form

$$\frac{\int \prod_{1 \le t \le n, T_{t-1}(\phi^*_{t-1}) \ge a^{i-1}/L} p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \theta) dF(\theta)}{\prod_{1 \le t \le n, T_{t-1}(\phi^*_{t-1}) \ge a^{i-1}/L} p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \lambda)}. \tag{3.13}$$

Controls from G that have been used most often as in (3.13) would thus provide accurate information about many of the characteristics of the unknown parameter but are incapable of distinguishing the true parameter θ_0 from the candidate values in $B(\theta_0)$, and therefore may not be able to settle whether $H_j : \theta \in \Theta_j$ is true. The controls in the complement of $\widehat{G}_{J,t}$, which make up the other mixture likelihood ratio statistic, are therefore needed to distinguish θ_0 from $B(\theta_0)$, but they would be used relatively infrequently. Similar reasoning has led us to replace the usual maximum likelihood estimate by a weighted version with weights $(T_n(g))^{-1}$ in (3.1). Since $\inf_{\lambda \in \Theta_j} = 1/\sup_{\lambda \in \Theta_j}$, the $\inf_{\lambda \in \Theta_j}$ in (3.11) is essentially tantamount to taking the

supremum of the denominator of (3.13) over $\lambda \in \Theta_j$, which is typically done in generalized likelihood ratio tests of the composite null hypothesis $H_j : \theta \in \Theta_j$. Our modification of the usual generalized likelihood ratio statistics consists of replacing $\sup_{\theta\in\Theta}$ by an integral with respect to a probability measure on Θ in the numerator of (3.13) and replacing a single likelihood ratio statistic by the maximum of two likelihood ratio statistics. The evenly allocated testing phase does not make use of $\widehat{\theta}_{a^i}$ and $\widehat{G}_{J,t}$. To test $H_j : \theta \in \Theta_j$, it uses the maximum of (3.13) and another mixture likelihood ratio statistic, which has the form (3.13) but with $T_{t-1}(\phi^*_{t-1}) \geq a^{i-1}/L$ replaced by $\phi^*_{t-1} = g_j$ and which is therefore based on data generated by the stationary control law g_j. For the special case of controlled independently and identically distributed (i.i.d.) processes, further details of the statistical ideas behind the modifications (3.11) and (3.12) of the classical generalized likelihood ratio statistics, together with illustrative examples and some variants of the adaptive control rule ϕ^*, are given in [8].

Throughout what follows we shall let θ_0 denote the true value of the unknown parameter. We shall also let $\mathbf{E}^\phi_x$ denote expectation under the probability measure $\mathbf{P}^\phi_x$ of the controlled Markov chain starting at x and using control rule ϕ, assuming the true value θ_0 of the parameter. Theorem 2 below shows that the regret $R_n(\theta_0)$ of ϕ^* satisfies

$$R_n(\theta_0) \sim \sum_{j\notin J(\theta_0)} c_j(\theta_0)[\mu^*_{\theta_0} - \mu_{\theta_0}(g_j)] \log n \tag{3.14}$$

under regularity conditions (C1)–(C5) in addition to those assumed at the beginning of section 1. Although we still use the notations (2.1)–(2.4) and define the $c_j(\theta)$ by (3.2) in Theorem 2, we do not assume condition (2.14) of Theorem 1. Our objective here is to establish (3.14), irrespective of whether it is an asymptotic lower bound for the regret of uniformly good rules. For $\delta > 0$, let $B_\delta(\theta_0)$ denote the open δ-neighborhood of $B(\theta_0)$; i.e., $B_\delta(\theta_0) = \{\theta \in \Theta : \inf_{\lambda\in B(\theta_0)} \rho(\theta, \lambda) < \delta\}$ ($B_\delta(\theta_0) = \emptyset$ if $B(\theta_0) = \emptyset$), where ρ denotes the metric of the compact metric space Θ.

(C1) For every $\epsilon > 0$, there exists $\delta > 0$ such that if $\rho(\theta_0, \theta) \leq \delta$ then $J(\theta) \subset J(\theta_0)$ and $\max_{j\notin J(\theta_0)} |c_j(\theta) - c_j(\theta_0)| < \epsilon$. Moreover, there exist ξ and $\delta^* > 0$ such that $\max_{j\notin J(\theta)} |c_j(\theta)| \leq \xi$ if $\rho(\theta_0, \theta) \leq \delta^*$.

(C2) $I^g(\theta_0, \theta)$ is a continuous function of θ for every $g \in G$.

(C3) $\max_{g\in G} I^g(\theta_0, \lambda) > 0$ for all $\lambda \in \cup_{j\in J(\theta_0)}\Theta_j - \{\theta_0\}$, $\inf_{\lambda\in\Theta_i\cap B(\theta_0)} I^{g_i}(\theta_0, \lambda) > 0$ and $\inf_{\lambda\in\Theta_i\setminus B_\delta(\theta_0)} \max_{j\in J(\theta_0)} I^{g_j}(\theta_0, \lambda) > 0$ for all $i \notin J(\theta_0)$ and $\delta > 0$ ($\inf \emptyset = \infty$).

(C4) For every $\theta \in \Theta$ and $g \in G$, there exist $\delta_\theta > 0$ and $r_\theta > 2$ such that

$$\sup_{x\in S} \mathbf{E}^g_x \left\{ \sup_{\lambda:\rho(\theta,\lambda)\leq\delta_\theta} \left| \log \frac{p(x, X_1; g(x), \theta_0)}{p(x, X_1; g(x), \lambda)} \right|^{r_\theta} \right\} < \infty$$

and, as $\delta \to 0$,

$$\sup_{x\in S} \mathbf{P}^g_x \left\{ \sup_{\lambda:\rho(\theta,\lambda)\leq\delta} \left| \frac{p(x, X_1; g(x), \theta)}{p(x, X_1; g(x), \lambda)} - 1 \right| \geq \epsilon \right\} \to 0 \text{ for all } \epsilon > 0.$$

(C5) For every $\theta \in \Theta$ and $g \in G$, as $n \to \infty$,

$$\sup_{x\in S} \left| \mathbf{E}^g_x \left\{ \frac{1}{n} \sum_{i=1}^n \log \frac{p(X_{t-1}, X_t; g(X_{t-1}), \theta_0)}{p(X_{t-1}, X_t; g(X_{t-1}), \theta)} \right\} - I^g(\theta_0, \theta) \right| \to 0.$$

Conditions (C1) and (C2) are continuity assumptions on $c_j(\theta)$ at $\theta = \theta_0$ and on $I^g(\theta_0, \theta)$. (C3) ensures that under ϕ^* one can estimate θ_0 consistently by the method of maximum likelihood, as will be shown in the proof of Theorem 2. From the definition (2.4) of $B(\theta_0)$, it follows that $\max_{j\in J(\theta_0)} I^{g_j}(\theta_0, \lambda) > 0$ if $\lambda \notin B(\theta_0) \cup (\cup_{j\in J(\theta_0)}\Theta_j)$, and the last inequality of (C3) requires this to be uniformly bounded away from zero for $\lambda \in \Theta_i \backslash B_\delta(\theta_0)$ with $i \notin J(\theta_0)$. Conditions (C4) and (C5) are natural moment and ergodicity assumptions on the log-likelihood ratio statistics. The uniformity over $x \in S$ in these assumptions enables us to get around difficulties with controlled Markov chains whose transition probability function $P_\theta^{g(x)}(x, A)$ is changed when the control law is changed. In section 4 we make use of martingale theory and uniform integrability to analyze the adaptive control rule ϕ^* in the proof of the following.

THEOREM 2. *Under* (C1)–(C5), *for every* $x \in S$ *and* $j \notin J(\theta_0)$,

$$\lim_{n\to\infty} \mathbf{E}_x^{\phi^*}(T_n(g_j))/\log n = c_j(\theta_0),$$

and therefore the regret $R_n(\theta_0)$ *of the rule* ϕ^* *satisfies* (3.14). *Moreover,* $S_n(\theta_0) = o(\log n)$, *where* $S_n(\theta_0)$ *is the expected number* (1.8) *of times that* ϕ^* *switches from one control law in* G *to another, not both optimal, up to stage* n.

The $c_j(\theta)$ are obtained by solving a constrained optimization problem in (3.2), which may be quite difficult in certain cases. Although the certainty-equivalent testing phase involves $J(\theta)$ and $c_j(\theta)$, these quantities are not used in the evenly allocated testing phase. In cases where the $c_j(\theta)$ are difficult to determine or fail to satisfy the continuity assumption (C1), an obvious modification of the adaptive control rule ϕ^* is to abandon the certainty-equivalent testing phase. Thus, partitioning $\{a^i + 1, \ldots, a^{i+1}\}$ into $m(i)$ blocks of consecutive integers so that the mth block B_m^i begins at time $\nu_i(m) := a^i + 1 + (m-1)n_i$, this modified rule $\widetilde{\phi}$ performs sequential testing of the hypotheses $H_j : \theta \in \Theta_j$ at the times $\nu_i(m)$ for those g_j not yet eliminated (during the times $\nu_i(m')$ with $m' < m$), in succession in ascending order of j, and eliminates g_j from further use through time a^{i+1} once the hypothesis H_j is rejected. Rejection of H_j occurs at the test time $n = \nu_i(m)$ if (3.12) holds. The rule $\widetilde{\phi}$ alternates using the stationary control laws that have not been eliminated, one for each block B_m^i of consecutive times. If all except one stationary control law have been eliminated, then $\widetilde{\phi}$ uses the remaining stationary control law up to time a^{i+1}. The following theorem shows that although this simpler rule $\widetilde{\phi}$ may be less efficient than ϕ^*, which attains the asymptotic lower bound $(c(\theta_0) + o(1))\log n$ for the regret, $\widetilde{\phi}$ still has a regret of the order $O(\log n)$.

THEOREM 3. *Under* (C2)–(C5), *the rule* $\widetilde{\phi}$ *satisfies* $R_n(\theta_0) = O(\log n)$ *and* $S_n(\theta_0) = o(\log n)$.

4. Martingale inequalities, uniform integrability, and proof of Theorems 2 and 3. We first consider some simple implications of conditions (C1)–(C5). Let $\epsilon > 0$ and take $2 < r'_\theta < r_\theta$. By (C2) together with (C4) for every $\theta \in \Theta$ we can choose $0 < \delta'_\theta \le \delta_\theta$ such that

$$\sup_{\lambda:\rho(\theta,\lambda)\le\delta'_\theta} |I^g(\theta_0, \theta) - I^g(\theta_0, \lambda)| < \epsilon \text{ for all } g \in G \text{ and} \tag{4.1}$$

$$\sup_{x\in S, g\in G} \mathbf{E}_x^g \left\{ \sup_{\lambda:\rho(\theta,\lambda)\le\delta'_\theta} \left| \log \frac{p(x, X_1; g(x), \theta_0)}{p(x, X_1; g(x), \lambda)} - \log \frac{p(x, X_1; g(x), \theta_0)}{p(x, X_1; g(x), \theta)} \right|^{r'_\theta} \right\} \le \epsilon^{r'_\theta}, \tag{4.2}$$

as will be explained in the next paragraph. Since Θ is compact, there exist finitely many points $\theta_1, \ldots, \theta_K$ such that

$$\Theta = \cup_{k=0}^{K} \{\lambda : \rho(\theta_k, \lambda) < \delta'_{\theta_k}\}. \tag{4.3}$$

By (C5) we can choose a positive integer D large enough so that

(4.4)

$$\sup_{x \in S, g \in G, k \le K} \left| \mathbf{E}_x^g \left\{ \frac{1}{n} \sum_{t=1}^{n} \log \frac{p(X_{t-1}, X_t; g(X_{t-1}), \theta_0)}{p(X_{t-1}, X_t; g(X_{t-1}), \theta_k)} \right\} - I^g(\theta_0, \theta_k) \right| \le \epsilon \text{ for all } n \ge D.$$

Concerning (4.2), first note that by (C4), $\sup_{x \in S} \mathbf{P}_x^g(\Omega(\delta, \eta; x)) \to 0$ as $\delta \to 0$ for every fixed $\eta > 0$, where $\Omega(\delta, \eta; x) = \{\sup_{\lambda : \rho(\theta, \lambda) \le \delta} |p(x, X_1; g(x), \theta)/p(x, X_1; g(x), \lambda) - 1| \ge \eta\}$. For sufficiently small η, $|\log y| < 2\eta$ if $|y - 1| < \eta$, and therefore, for $0 < \delta < \delta_\theta$,

$$\begin{aligned}
&\sup_{x \in S} \mathbf{E}_x^g \left\{ \sup_{\lambda : \rho(\theta, \lambda) \le \delta} \left| \log \frac{p(x, X_1; g(x), \theta)}{p(x, X_1; g(x), \lambda)} \right|^{r'_\theta} \right\} \\
&\le (2\eta)^{r'_\theta} + \sup_{x \in S} \mathbf{E}_x^g \left\{ \sup_{\lambda : \rho(\theta, \lambda) \le \delta} \left| \log \frac{p(x, X_1; g(x), \theta)}{p(x, X_1; g(x), \lambda)} \right|^{r'_\theta} I_{\Omega(\delta, \eta; x)} \right\} \\
&\le (2\eta)^{r'_\theta} \\
&\quad + \sup_{x \in S} \left\{ \mathbf{E}_x^g \left[\sup_{\lambda : \rho(\theta, \lambda) \le \delta_\theta} \left| \log \frac{p(x, X_1; g(x), \theta_0)}{p(x, X_1; g(x), \lambda)} - \log \frac{p(x, X_1; g(x), \theta_0)}{p(x, X_1; g(x), \theta)} \right|^{r_\theta} \right] \right\}^{r'_\theta / r_\theta} \\
&\quad \times \{\mathbf{P}_x^g(\Omega(\delta, \eta; x))\}^{1 - r'_\theta / r_\theta} \le (2\eta)^{r'_\theta} + o(1) \text{ as } \delta \to 0
\end{aligned}$$

by (C4), where we have used Hölder's inequality to obtain the second inequality.

We next make use of martingale theory to analyze the log-likelihood ratio statistics from the controlled Markov chain using the adaptive control rule ϕ^*. For $t \ge 0$, let $\mathcal{F}_t$ be the σ-field defined by (1.6). The control rule ϕ^* uses the same stationary control law for an entire block of stages $\nu_i(m), \ldots, \nu_i(m+1) - 1$, with the choice of the control law determined at the beginning of the block. Define a sequence of positive integers h_s such that $D \le h_s - h_{s-1} \le 2D - 1$ for $s > 1$ and all the $\nu_i(m)$ with $i \ge D$ belong to the sequence, where D is given by (4.4). For example, take $h_0 = 0, h_1 = a^D$, and for $s > 1$ let $h_s = h_{s-1} + D$ except when $h_s = \nu_i(m)$, for which we may change the above recursive definition of h_s to $h_s = h_{s-1} + D + r$, with $0 \le r < D$ being the remainder obtained when $\nu_i(m) - \nu_i(m-1)$ is divided by D. Since ϕ^* uses the same stationary control law for $h_{s-1}, \ldots, h_s - 1$ on the basis of observations prior to h_{s-1}, it follows that for $h_{s-1} < t \le h_s$ and $g \in G$, $\{\phi^*_{t-1} = g\} \in \mathcal{F}_{h_{s-1}}$. Therefore, by (4.4),

(4.5)

$$\begin{aligned}
\sup_{n \ge 1, g \in G, k \le K} \Bigg| \frac{1}{T_{h_n}(g)} \sum_{s=1}^{n} \sum_{h_{s-1} < t \le h_s} \mathbf{E}_x^{\phi^*} \left\{ \log \frac{p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \theta_0)}{p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \theta_k)} \middle| \mathcal{F}_{h_{s-1}} \right\} \\
\cdot\ I_{\{\phi^*_{t-1} = g\}} - I^g(\theta_0, \theta_k) \Bigg| I_{\{T_{h_n}(g) \ge D\}} \le \epsilon,
\end{aligned}$$

noting that $T_{h_n}(g) = \sum_{s=1}^{n} \sum_{h_{s-1} \le i < h_s} I_{\{\phi_i^* = g\}}$ and

$$\sum_{t=h_{s-1}+1}^{h_s} \mathbf{E}_x^{\phi^*} \left\{ \log \frac{p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \theta_0)}{p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \theta_k)} \middle| \mathcal{F}_{h_{s-1}} \right\} I_{\{\phi_{t-1}^* = g\}} \tag{4.6}$$

$$= \mathbf{E}_y^g \left\{ \sum_{t=1}^{h_s - h_{s-1}} \log \frac{p(X_{t-1}, X_t; g(X_{t-1}), \theta_0)}{p(X_{t-1}, X_t; g(X_{t-1}), \theta_k)} \right\} \text{ on } \{\phi_{h_{s-1}}^* = g, X_{h_{s-1}} = y\}.$$

Let

$$\ell_t(\theta) = \log[p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \theta_0) / p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \theta)]. \tag{4.7}$$

Lemma 1 states a result of [14], and Lemmas 2, 3, and 4 use it to approximate $\sum_{t=1}^{h_n} \ell_t(\theta)\ \chi_t$, where the χ_t are indicator variables (taking values in $\{0, 1\}$).

LEMMA 1. *Let $\{Z_n\}$ be a martingale difference sequence with respect to an increasing sequence of σ-fields $\{\mathcal{B}_n\}$ such that $\sup_n E(|Z_n|^\beta | \mathcal{B}_{n-1}) \le C$ a.s. for some nonrandom constants $\beta > 2$ and $C < \infty$. Let χ_n be $\mathcal{B}_{n-1}$-measurable variables taking values in $\{0, 1\}$ and let $\#_n = \sum_{t=1}^{n} \chi_t$. Then there exists a universal constant A depending only on C and β (and not on the distribution of $\{(\chi_n, Z_n) : n \ge 1\}$) such that for any $\eta > 0$ and $m \ge 1$,*

$$P\left\{ \sup_{n: \#_n \ge m} \left| \sum_{t=1}^{n} Z_t \chi_t \right| / \#_n > \eta \right\} \le A\{m^{-(\beta-1)} \eta^{-\beta} + (\eta^2 m)^{-(\beta-1)}\},$$

$$P\left\{ \max_{1 \le n \le m} \left| \sum_{t=1}^{n} Z_t \chi_t \right| > \eta m \right\} \le A\{m^{-(\beta-1)} \eta^{-\beta} + (\eta^2 m)^{-(\beta-1)}\}.$$

LEMMA 2. *Let $\beta = \min_{0 \le k \le K} r'_{\theta_k} (> 2)$, where the r'_θ are given by (4.2). Define $\ell_t(\theta)$ by (4.7) and $\mathcal{C}, \tau_n(g)$ by (3.7). Then for any $g \in G$ and $0 \le k \le K$,*

$$\mathbf{P}_x^{\phi^*} \left\{ \sup_{n: T_{h_n}(g) \ge m} \left| (T_{h_n}(g))^{-1} \sum_{t=1}^{h_n} \ell_t(\theta_k) I_{\{\phi_{t-1}^* = g\}} - I^g(\theta_0, \theta_k) \right| > 2\epsilon \right\} = O(m^{-(\beta-1)}),$$

$$\mathbf{P}_x^{\phi^*} \left\{ \sup_{n: \tau_{h_n}(g) \ge m} \left| (\tau_{h_n}(g))^{-1} \sum_{t=1}^{h_n} \ell_t(\theta_k) I_{\{t-1 \in \mathcal{C}, \phi_{t-1}^* = g\}} - I^g(\theta_0, \theta_k) \right| > 2\epsilon \right\}$$
$$= O(m^{-(\beta-1)}).$$

Proof. Let $Z_s = \sum_{h_{s-1} < t \le h_s} \ell_t(\theta_k) - \mathbf{E}_x^{\phi^*} \{\sum_{h_{s-1} < t \le h_s} \ell_t(\theta_k) | \mathcal{F}_{h_{s-1}}\}$. Then $\{Z_s, \mathcal{B}_s, s \ge 1\}$ is a martingale difference sequence, where $\mathcal{B}_s = \mathcal{F}_{h_s}$. Moreover, since $h_s - h_{s-1} \le 2D - 1$,

$$\sup_{s \ge 1} \mathbf{E}_x^{\phi^*} (|Z_s|^\beta | \mathcal{B}_{s-1}) \le A_{\beta, D} \sup_{y \in S, g \in G} \sum_{t=1}^{2D-1} \mathbf{E}_y^g |\ell_t(\theta_k)|^\beta$$

for some positive constant $A_{\beta,D}$ that depends only on β and D. For fixed $g \in G$, let $\chi_s = I_{\{\phi_{h_{s-1}}^* = g\}}$, which is $\mathcal{B}_{s-1}$ measurable, and note that

$$\sum_{h_{s-1} < t \le h_s} \ell_t(\theta_k) I_{\{\phi_{t-1}^* = g\}} = \chi_s \sum_{h_{s-1} < t \le h_s} \ell_t(\theta_k), \tag{4.8}$$

since ϕ^* uses the same stationary control law at the times $h_{s-1}, \ldots, h_s - 1$. Moreover,

$$D\sum_{s=1}^{n}\chi_s \le \sum_{s=1}^{n}(h_s - h_{s-1})\chi_s = T_{h_n}(g) \le (2D-1)\sum_{s=1}^{n}\chi_s.$$

Therefore it follows from Lemma 1 that

$$\mathbf{P}_x^{\phi^*}\left\{\sup_{n:T_{h_n}(g)\ge m}\left|\sum_{s=1}^{n} Z_s\chi_s\right| / T_{h_n}(g) > \epsilon\right\} = O(m^{-(\beta-1)}). \tag{4.9}$$

Since $\sum_{t=1}^{h_n}\ell_t(\theta_k)I_{\{\phi^*_{t-1}=g\}} = \sum_{s=1}^{n} Z_s\chi_s + \sum_{s=1}^{n}\sum_{h_{s-1}<t\le h_s}\mathbf{E}_x^{\phi^*}\{\ell_t(\theta_k)|\mathcal{F}_{h_{s-1}}\}I_{\{\phi^*_{t-1}=g\}}$ by (4.8), the desired conclusion for $T_{h_n}(g)$ follows from (4.5) and (4.9). The conclusion for $\tau_{h_n}(g)$ can be proved similarly, noting that for $h_{s-1} < t \le h_s$, $\{t-1 \in \mathcal{C}, \phi^*_{t-1} = g\} \in \mathcal{F}_{h_{s-1}}$.

LEMMA 3. *With β defined in Lemma 2 and δ'_θ given by (4.1) and (4.2), let χ_n be $\mathcal{F}_{n-1}$-measurable random variables taking values in $\{0,1\}$ and let $\#_n = \sum_{t=1}^{n}\chi_t$. Then for any $0 \le k \le K$,*

$$\mathbf{P}_x^{\phi^*}\left\{\sup_{n:\#_n\ge m}\left[\sup_{\theta:\rho(\theta_k,\theta)\le\delta'_{\theta_k}}\sum_{t=1}^{n}|\ell_t(\theta)-\ell_t(\theta_k)|\chi_t\right]/\#_n > 2\epsilon\right\} = O(m^{-(\beta-1)}).$$

Proof. Let $\Gamma_k = \{\theta : \rho(\theta_k,\theta) \le \delta'_{\theta_k}\}$. In view of (C4), applying Lemma 1 to $Z_t = \sup_{\theta\in\Gamma_k}|\ell_t(\theta)-\ell_t(\theta_k)| - \mathbf{E}_x^{\phi^*}\{\sup_{\theta\in\Gamma_k}|\ell_t(\theta)-\ell_t(\theta_k)||\mathcal{F}_{t-1}\}$ yields

$$\mathbf{P}_x^{\phi^*}\left\{\sup_{n:\#_n\ge m}\left(\sum_{t=1}^{n} Z_t\chi_t\right)/\#_n > \epsilon\right\} = O(m^{-(\beta-1)}). \tag{4.10}$$

Moreover, by (4.2),

$$\begin{aligned}\sum_{t=1}^{n}\chi_t\mathbf{E}_x^{\phi^*}&\left\{\sup_{\theta\in\Gamma_k}|\ell_t(\theta)-\ell_t(\theta_k)|\,\middle|\,\mathcal{F}_{t-1}\right\}\\ &\le\left(\sum_{t=1}^{n}\chi_t\right)\sup_{y\in S,g\in G}\mathbf{E}_y^g\left(\sup_{\lambda\in\Gamma_k}|\ell_t(\lambda)-\ell_t(\theta_k)|\right)\\ &\le\epsilon\#_n.\end{aligned} \tag{4.11}$$

Since $\sup_{\theta\in\Gamma_k}\sum_{t=1}^{n}|\ell_t(\theta)-\ell_t(\theta_k)|\chi_t \le \sum_{t=1}^{n} Z_t\chi_t + \sum_{t=1}^{n}\chi_t\mathbf{E}_x^{\phi^*}\{\sup_{\theta\in\Gamma_k}|\ell_t(\theta)-\ell_t(\theta_k)||\mathcal{F}_{t-1}\}$, the desired conclusion follows from (4.10) and (4.11).

LEMMA 4. *With the same notation as in Lemma 3, for any $0 \le k \le K$,*

$$\mathbf{P}_x^{\phi^*}\left\{\max_{1\le n\le m}\sup_{\theta:\rho(\theta_k,\theta)\le\delta'_{\theta_k}}\sum_{t=1}^{n}|\ell_t(\theta)-\ell_t(\theta_k)|\chi_t > 2\epsilon m\right\} = O(m^{-(\beta-1)}).$$

Moreover, for any $g \in G$ and $0 \le k \le K$,

$$\sup_{\tau\in\mathcal{T}}\mathbf{P}_x^{\phi^*}\left\{\max_{1\le n\le m}\left|\sum_{t=1}^{h_n}(\ell_t(\theta_k)-I^g(\theta_0,\theta_k))I_{\{t>\tau,\phi^*_{t-1}=g\}}\right| > 3\epsilon h_m\right\} = O(m^{-(\beta-1)}),$$

where $\mathcal{T}$ denotes the class of all stopping times (with respect to $\{\mathcal{F}_n\}$).

Proof. By using the second instead of the first inequality of Lemma 1, we can proceed as in the proof of Lemma 3 to obtain the first conclusion. To prove the second conclusion, let τ be a stopping time and let $\sigma = \inf\{s : h_s \geq \tau\}$. Define Z_s and $\mathcal{B}_s$ as in the proof of Lemma 2 but change the definition of χ_s there to $\chi_s = I_{\{\phi^*_{h_{s-1}}=g, s>\sigma\}}$. Since

$$\{s > \sigma\} = \{s - 1 \geq \sigma\} = \{h_{s-1} \geq \tau\} \in \mathcal{F}_{h_{s-1}} = \mathcal{B}_{s-1},$$

χ_s is $\mathcal{B}_{s-1}$ measurable. Therefore, by Lemma 1 there exists a constant A (which does not depend on σ) such that, for all $m \geq 1$,

$$\mathbf{P}_x^{\phi^*}\left\{\max_{1\leq n\leq m}\left|\sum_{s=1}^{n} Z_s\chi_s\right| > \epsilon m\right\} \leq A\{m^{-(\beta-1)}\epsilon^{-\beta} + (\epsilon^2 m)^{-(\beta-1)}\}.$$

Note that for $n \geq \sigma$, $\sum_{t=1}^{h_n} \ell_t(\theta_k) I_{\{\phi^*_{t-1}=g, t>\tau\}}$ can be written as

$$\sum_{\tau+1\leq t\leq h_\sigma} \ell_t(\theta_k) I_{\{\phi^*_{t-1}=g\}} + \sum_{s=1}^{n} \chi_s \mathbf{E}_x^{\phi^*}\left\{\sum_{h_{s-1}<t\leq h_s} \ell_t(\theta_k)\,\middle|\, \mathcal{F}_{h_{s-1}}\right\} + \sum_{s=1}^{n} Z_s\chi_s.$$

Since $h_s - h_{s-1} \geq D$, it follows from (4.4) that

$$\max_{1\leq s\leq m}\left|\sum_{s=1}^{n} \chi_s \sum_{h_{s-1}<t\leq h_s} \{\mathbf{E}_x^{\phi^*}[\ell_t(\theta_k)|\mathcal{F}_{h_{s-1}}] - I^g(\theta_0,\theta_k)\}\right| \leq \epsilon h_m.$$

Since $h_{\sigma-1} < \tau \leq h_\sigma$ and $h_\sigma - h_{\sigma-1} \leq 2D - 1$, the strong Markov property implies that

$$\mathbf{E}_x^{\phi^*}\left\{\left(\sum_{\tau+1\leq t\leq h_\sigma} |\ell_t(\theta_k)| I_{\{\phi^*_{t-1}=g\}}\right)^{\beta}\,\middle|\, \mathcal{F}_\tau\right\} \leq A_{\beta,D} \sup_{y\in S} \sum_{t=1}^{2D-1} \mathbf{E}_y^g|\ell_t(\theta_k)|^\beta$$

for some positive constant $A_{\beta,D}$ that depends only on β and D. Hence by the Markov inequality

$$\mathbf{P}_x^{\phi^*}\left\{\sum_{\tau+1\leq t\leq h_\sigma} |\ell_t(\theta_k)| I_{\{\phi^*_{t-1}=g\}} > \epsilon m\right\} \leq C\epsilon^{-\beta} m^{-\beta},$$

where $C = A_{\beta,D} \sup_{y\in S} \sum_{t=1}^{2D-1} \mathbf{E}_y^g|\ell_t(\theta_k)|^\beta < \infty$. Since the same constants A and C in the above probability bounds hold for all stopping times τ, these bounds yield the second conclusion of the lemma.

We shall make use of Lemmas 2–4 to prove the following two lemmas from which Theorems 2 and 3 follow easily. Recall that $G_J = \{g_j : j \in J(\theta_0)\}$.

LEMMA 5. *With $\beta > 2$ defined in Lemma 2, for every $\eta > 0$,*

$$\mathbf{P}_x^{\phi^*}\{\rho(\theta_0, \widehat{\theta}_{h_n}) \geq \eta \text{ for some } h_n \geq a^i\} = O((i/\log i)^{-(\beta-1)}). \tag{4.12}$$

Moreover, for any $j \notin J(\theta_0)$, as $n \to \infty$,

$$T_n(g_j)/\log n \to c_j(\theta_0),\ \sum_{i=1}^{n} I_{\{\phi_i^* \neq \phi_{i-1}^*, \phi_i^* \notin G_J \text{ or } \phi_{i-1}^* \notin G_J\}}/\log n \to 0 \text{ a.s. } [\mathbf{P}_x^{\phi^*}]. \tag{4.13}$$

LEMMA 6. *For any $j \notin J(\theta_0)$, $\{T_n(g_j)/\log n, n \geq 2\}$ is uniformly integrable under $\mathbf{P}_x^{\phi^*}$ or $\mathbf{P}_x^{\hat{\phi}}$.*

Proof of Theorem 2. From (4.13) and Lemma 6, it follows that $\mathbf{E}_x^{\phi^*} T_n(g_j)/\log n \to c_j(\theta_0)$ for any $j \notin J(\theta_0)$. This and (3.2) imply (3.14). The uniform integrability of

$$\sum_{i=1}^{n} I_{\{\phi_i^* \neq \phi_{i-1}^*, \phi_i^* \notin G_J \text{ or } \phi_{i-1}^* \notin G_J\}}/\log n,$$

which is $\leq 2\{1 + \sum_{j \notin J(\theta_0)} T_n(g_j)\}/\log n$, follows from Lemma 6. Therefore $S_n(\theta_0) = o(\log n)$ by (4.13).

Proof of Theorem 3. The desired conclusion on $R_n(\theta_0)$ follows from Lemma 6, and that on $S_n(\theta_0)$ can be proved by an argument similar to the proof of the second convergence in (4.13) and the associated uniform integrability in Theorem 2.

The proof of Lemmas 5 and 6 makes use of the following lemma, which applies martingale inequalities to analyze boundary crossing probabilities associated with (3.11) and (3.12).

LEMMA 7. *As in* (3.11) *and* (3.12), *let F be a probability measure on Θ such that $F(A) > 0$ for all open subsets A of Θ. For $a^i < n \leq a^{i+1}$, let*

$$U_n(\lambda) = \frac{\int \prod_{1 \leq t \leq n, t-1 \in \mathcal{C}, \phi_{t-1}^* \notin \hat{G}_{J,t}} p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \theta) dF(\theta)}{\prod_{1 \leq t \leq n, t-1 \in \mathcal{C}, \phi_{t-1}^* \notin \hat{G}_{J,t}} p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \lambda)}, \tag{4.14}$$

$$\widetilde{W}_{n,j}(\lambda) = \frac{\int \prod_{1 \leq t \leq n, \phi_{t-1}^* = g_j} p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \theta) dF(\theta)}{\prod_{1 \leq t \leq n, \phi_{t-1}^* = g_j} p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \lambda)}, \tag{4.15}$$

$$W_n(\lambda) = \frac{\int \prod_{1 \leq t \leq n, T_{t-1}(\phi_{t-1}^*) \geq a^{i-1}/L} p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \theta) dF(\theta)}{\prod_{1 \leq t \leq n, T_{t-1}(\phi_{t-1}^*) \geq a^{i-1}/L} p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \lambda)}. \tag{4.16}$$

If $\theta_0 \in \Theta_j$, then

$$\mathbf{P}_x^{\phi^*}\left\{\inf_{\lambda \in \Theta_j} \max(U_n(\lambda), W_n(\lambda)) \geq ia^i \text{ for some } n > a^i\right\} \leq 2(ia^i)^{-1},$$

$$\mathbf{P}_x^{\phi^*}\left\{\inf_{\lambda \in \Theta_j} \max(\widetilde{W}_{n,j}(\lambda), W_n(\lambda)) \geq ia^i \text{ for some } n > a^i\right\} \leq 2(ia^i)^{-1}.$$

Proof. Note that $\{t-1 \in \mathcal{C}, \phi_{t-1}^* \notin \widehat{G}_{J,t}\} \in \mathcal{F}_{t-1}$ by (3.7) and (3.8). Hence $U_n(\theta_0), W_n(\theta_0)$ and $\widetilde{W}_{n,j}(\theta_0), n > a^i$, are nonnegative martingales with common mean 1. Therefore, if $\theta_0 \in \Theta_j$, then

$$\begin{aligned}
&\mathbf{P}_x^{\phi^*}\left\{\inf_{\lambda \in \Theta_j} \max(U_n(\lambda), W_n(\lambda)) \geq ia^i \text{ for some } n > a^i\right\} \\
&\leq \mathbf{P}_x^{\phi^*}\{U_n(\theta_0) \geq ia^i \text{ for some } n > a^i\} + \mathbf{P}_x^{\phi^*}\{W_n(\theta_0) \geq ia^i \text{ for some } n > a^i\} \\
&\leq \{EU_{a^i+1}(\theta_0) + EW_{a^i+1}(\theta_0)\}/(ia^i).
\end{aligned}$$

Replacing $U_n(\lambda)$ by $\widetilde{W}_{n,j}(\lambda)$ in the above argument proves the second inequality.

Proof of Lemma 5. We first prove (4.12). By Lemma 3 (with $\chi_t = I_{\{\phi_{t-1}^* = g\}}$), for every $g \in G$ and $0 \leq k \leq K$,

(4.17)
$$\mathbf{P}_x^{\phi^*}\left\{\sup_{\theta\in\Gamma_k}\sum_{t=1}^{n}|\ell_t(\theta)-\ell_t(\theta_k)|I_{\{\phi^*_{t-1}=g\}}\geq 2\epsilon T_n(g)\text{ for some }T_n(g)\geq m\right\}=O(m^{-(\beta-1)}).$$

From (4.17) and Lemma 2, it follows that

(4.18)
$$\mathbf{P}_x^{\phi^*}\left\{\inf_{\theta\in\Gamma_k}(T_{h_n}(g))^{-1}\sum_{t=1}^{h_n}\ell_t(\theta)I_{\{\phi^*_{t-1}=g\}}\geq I^g(\theta_0,\theta_k)-4\epsilon\text{ for all }T_{h_n}(g)\geq m\right.$$
$$\left.\text{and every } g\in G \text{ and } 0\leq k\leq K\right\}\geq 1-O(m^{-(\beta-1)}).$$

By (4.1) and the compactness of Θ, for every $\epsilon>0$, there exists $\delta>0$ such that

(4.19)
$$\sup_{g\in G}|I^g(\theta_0,\lambda)-I^g(\theta_0,\lambda')|<\epsilon \text{ if } \rho(\lambda,\lambda')\leq\delta.$$

For $i\notin J(\theta_0)$, since $\inf_{\lambda\in\Theta_i\cap B(\theta_0)}I^{g_i}(\theta_0,\lambda)>0$ by (C3), it follows from (4.19) (with ϵ sufficiently small) that $\inf_{\lambda\in\Theta_i\cap B_\delta(\theta_0)}I^{g_i}(\theta_0,\lambda)>0$ for some $\delta>0$. This and (C3) imply that $\max_{g\in G}I^g(\theta_0,\lambda)>0$ for all $\lambda\neq\theta_0$. Hence given $\eta>0$, we can choose ϵ sufficiently small so that

(4.20)
$$\max_{g\in G}I^g(\theta_0,\theta)\geq 5L\epsilon \quad\text{if}\quad \rho(\theta_0,\theta)\geq\eta,$$

in view of (C2) and the compactness of Θ. Since

$$L_n(\theta_0)-L_n(\theta)=\sum_{g\in G}(T_n(g))^{-1}\sum_{t=1}^{n}\ell_t(\theta)I_{\{\phi^*_{t-1}=g\}}$$

by (3.1), and since $\ell_t(\theta_0)=0$ and $I^g(\theta_0,\lambda)\geq 0$ for all $g\in G$ and $\lambda\in\Theta$, it follows from (4.18) and (4.20) that

(4.21)
$$\mathbf{P}_x^{\phi^*}\left\{\sup_{\theta:\rho(\theta_0,\theta)\geq\eta}L_{h_n}(\theta)<L_{h_n}(\theta_0)-\epsilon\text{ for all }h_n\geq a^i\right\}\geq 1-O((i^{-1}\log i)^{\beta-1}),$$

noting that $T_{a^i}(g)\geq\tau_{a^i}(g)\geq n_{i-1}(\sim i/\log i)$ because at least n_{i-1} stages in the certainty-equivalent testing phase between the times a^{i-1} and a^i use g if $\tau_{a^{i-1}}(g)<n_{i-1}$ for every $g\in G$. From (4.21), (4.12) follows.

Combining (4.12) with (C1) yields $\mathbf{P}_x^{\phi^*}(\cap_{i\geq t}A_i)\geq 1-O((t^{-1}\log t)^{\beta-1})$, where

(4.22)
$$A_i=\left\{J(\widehat{\theta}_{a^i})\subset J(\theta_0),\ \max_{j\notin J(\theta_0)}|c_j(\widehat{\theta}_{a^i})-c_j(\theta_0)|<\epsilon,\ \max_{j\in J(\theta_0)-J(\widehat{\theta}_{a^i})}|c_j(\widehat{\theta}_{a^i})|\leq\xi\right\}.$$

Let $\Gamma_k=\{\theta:\rho(\theta_k,\theta)\leq\delta'_{\theta_k}\}$ and $w_k=F(\Gamma_k)(>0)$ for $0\leq k\leq K$. By Lemma 3, $\mathbf{P}_x^{\phi^*}(\cap_{i\geq t}C_i)\geq 1-O((t^{-1}\log t)^{\beta-1})$, where

(4.23)
$$C_i=\left\{\max_{0\leq k\leq K}\sup_{\theta\in\Gamma_k}\sum_{t=1}^{n}|\ell_t(\theta)-\ell_t(\theta_k)|I_{\{t-1\in\mathcal{C},\phi^*_{t-1}=g\}}\leq 2\epsilon\tau_n(g)\right.$$
$$\left.\text{for all }\ a^i<n\leq a^{i+1}\text{ and }g\in G\right\},$$

since $\tau_{a^i}(g) \geq n_{i-1} \sim i/\log i$ for every $g \in G$. Note that $\ell_t(\theta_0) = 0$ and that

$$\begin{aligned}
&\int \prod_{1\leq t\leq n, t-1\in\mathcal{C}, \phi^*_{t-1}\notin \widehat{G}_{J,t}} p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \theta) dF(\theta) \\
&\geq \left\{ \prod_{1\leq t\leq n, t-1\in\mathcal{C}, \phi^*_{t-1}\notin \widehat{G}_{J,t}} p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \theta_0) \right\} \\
&\quad \cdot w_0 \inf_{\theta\in\Gamma_0} \exp\left(- \sum_{1\leq t\leq n, t-1\in\mathcal{C}, \phi^*_{t-1}\notin \widehat{G}_{J,t}} \ell_t(\theta) \right).
\end{aligned} \tag{4.24}$$

Suppose $\Theta_j \cap B_\delta(\theta_0) \neq \emptyset$. Let $\chi_t = I_{\{t-1\in\mathcal{C}, \phi^*_{t-1}\notin \widehat{G}_{J,t}\}}$. On $A_i \cap C_i$, if $a^i < n \leq a^{i+1}$, then

$$\inf_{\lambda\in\Gamma_k} \exp\left(\sum_1^n \chi_t \ell_t(\lambda) \right) \geq \exp\left\{ \sum_1^n \chi_t \ell_t(\theta_k) - 2\epsilon \sum_1^n \chi_t \right\},$$

$$\inf_{\lambda\in\Gamma_0} \exp\left(-\sum_1^n \chi_t \ell_t(\lambda) \right) \geq \exp\left(-2\epsilon \sum_1^n \chi_t \right),$$

so it follows from (4.7), (4.14), and (4.24) that

$$\inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} U_n(\lambda) \geq w_0 \exp\left\{ \inf_{0\leq k\leq K:\Gamma_k\cap\Theta_j\cap B_\delta(\theta_0)\neq\emptyset} \sum_{t=1}^n (\ell_t(\theta_k) - 4\epsilon)\chi_t \right\}. \tag{4.25}$$

Since $\tau_{a^i}(g) \geq n_{i-1}$, it follows from Lemma 2 that $\mathbf{P}_x^{\phi^*}(\cap_{i\geq t} D_i) \geq 1 - O((t^{-1}\log t)^{\beta-1})$, where

$$\begin{aligned}
D_i = \Bigg\{ &\max_{g\in G} \max_{0\leq k\leq K} \left| (\tau_{h_n}(g))^{-1} \sum_{t=1}^{h_n} \ell_t(\theta_k) I_{\{t-1\in\mathcal{C}, \phi^*_{t-1}=g\}} - I^g(\theta_0, \theta_k) \right| \leq 2\epsilon \\
&\text{for all } a^i < h_n \leq a^{i+1} \Bigg\}.
\end{aligned} \tag{4.26}$$

Let $\Omega = \cup_{t=1}^\infty \cap_{i\geq t} (A_i \cap C_i \cap D_i)$. Then $\mathbf{P}_x^{\phi^*}(\Omega) = \lim_{t\to\infty} \mathbf{P}_x^{\phi^*}\{\cap_{i\geq t}(A_i\cap C_i\cap D_i)\} = 1$. On Ω, for all large i and at the times h_n during the certainty-equivalent testing phase of ϕ^* between a^i and a^{i+1}, it follows from (3.9), (3.10a), and (4.22) that $J(\widehat{\theta}_{a^i}) \subset J(\theta_0)$ and

$$(\log a^i)(c_\ell(\theta_0) - \epsilon) \leq \tau_{h_n}(g_\ell) \leq (2\log a^i)(c_\ell(\theta_0) + \epsilon) + 3n_i \text{ if } \ell \notin J(\theta_0), \tag{4.27}$$

$$\tau_{h_n}(g_\ell) \leq 3\xi \log a^i \text{ if } \ell \in J(\theta_0) - J(\widehat{\theta}_{a^i}), \tag{4.28}$$

and from (4.25), (4.26), and (4.28) that

(4.29)
$$\inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} \log U_{h_n}(\lambda)$$
$$\geq \inf_{0\leq k\leq K:\Gamma_k\cap\Theta_j\cap B_\delta(\theta_0)\neq\emptyset} \sum_{\ell\notin J(\theta_0)} \{I^{g_\ell}(\theta_0,\theta_k)-6\epsilon\}\tau_{h_n}(g_\ell) - 6L\epsilon(3\xi\log a^i)+O(1),$$

noting that $I^g(\theta_0,\lambda)\geq 0$ for all $g\in G$ and $\lambda\in\Theta$. From (4.1) and (4.29), it follows that on Ω

(4.30)
$$\begin{aligned}&\inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} \log U_{h_n}(\lambda)\\ &\geq \inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} \sum_{\ell\notin J(\theta_0)} \{I^{g_\ell}(\theta_0,\lambda)-7\epsilon\}\tau_{h_n}(g_\ell) - 6L\epsilon(3\xi\log a^i)+O(1)\end{aligned}$$

at the times h_n during the certainty-equivalent phase of ϕ^* between a^i and a^{i+1} for all large i.

In view of (3.11), (3.12), and Lemma 7, for every $\ell\in J(\theta_0)$,

(4.31) $\mathbf{P}_x^{\phi^*}\{g_\ell \text{ is eliminated at some testing time between } a^i \text{ and } a^{i+1}\}\leq 4(ia^i)^{-1}$.

Hence by the Borel–Cantelli lemma, $\mathbf{P}_x^{\phi^*}\{\Omega_i \text{ for all large } i\}=1$, where

(4.32)
$$\begin{aligned}&\Omega_i\\ &=\{g_\ell \text{ is not eliminated during all test times between } a^{i-2} \text{ and } a^{i+1},\ \text{for all } \ell\in J(\theta_0)\}.\end{aligned}$$

In the event Ω_i, since $1-a^{-2}\geq 3/4$ and $3/5>a^{-1}$, we have the following for all sufficiently large i:

(4.33) $$T_{a^i-1}(g_\ell)\geq (4/5)\{(a^i-1-a^{i-2})/L\}>a^{i-1}/L \text{ for all } \ell\in J(\theta_0).$$

Note that $\{T_{t-1}(g)\geq a^{i-1}/L\}=\{t-1\geq\tau^{(i)}\}=\{t>\tau^{(i)}\}$, where $\tau^{(i)}=\inf\{s: T_s(g)\geq a^{i-1}/L\}$ is a stopping time. Let $N_n(g)=\sum_{t=1}^n I_{\{T_{t-1}(g)\geq a^{i-1}/L,\phi^*_{t-1}=g\}}$, and define

(4.34)
$$\begin{aligned}\Lambda_i=\Bigg\{&\max_{0\leq k\leq K}\sup_{\theta\in\Gamma_k}\sum_{t=1}^{h_n}|\ell_t(\theta)-\ell_t(\theta_k)|I_{\{T_{t-1}(g)\geq a^{i-1}/L,\phi^*_{t-1}=g\}}\leq 2\epsilon a^{i+1} \text{ and}\\ &\max_{0\leq k\leq K}\left|\sum_{t=1}^{h_n}\ell_t(\theta_k)I_{\{T_{t-1}(g)\geq a^{i-1}/L,\phi^*_{t-1}=g\}}-I^g(\theta_0,\theta_k)N_{h_n}(g)\right|\leq 2\epsilon a^{i+1}\\ &\text{for all } a^i<h_n\leq a^{i+1} \text{ and all } g\in G\Bigg\}.\end{aligned}$$

Letting Λ^c denote the complement of an event Λ, it follows from Lemma 4 that

(4.35) $$\mathbf{P}_x^{\phi^*}(\Lambda_i^c)=O(a^{-i(\beta-1)}).$$

Therefore by the Borel–Cantelli lemma, $\mathbf{P}_x^{\phi^*}\{\Lambda_i \text{ for all large } i\}=1$.

Suppose $\Theta_j \backslash B_\delta(\theta_0) \neq \emptyset$. Then we can use (4.33) and an argument similar to that leading to (4.25) and (4.30) to show that, for all large i, on $\Omega_i \cap \Lambda_i$,

$$\begin{aligned} &\min_{1\le m\le m(i)} \inf_{\lambda\in\Theta_j\backslash B_\delta(\theta)} \log W_{\nu_i(m)}(\lambda) \\ &\ge \inf_{\lambda\in\Theta_j\backslash B_\delta(\theta_0)} \sum_{\ell\in J(\theta_0)} I^{g_\ell}(\theta_0,\lambda) a^{i-1}/L - 7\epsilon a^{i+1} + \log w_0, \end{aligned} \tag{4.36}$$

noting that $I^g(\theta_0,\lambda) \ge 0$. Let $\Omega_* = \cup_{t=1}^\infty \cap_{i\ge t} (\Omega_i \cap \Lambda_i)$. Since

$$K := \inf_{\lambda\in\Theta_j\backslash B_\delta(\theta_0)} \sum_{\ell\in J(\theta_0)} I^{g_\ell}(\theta_0,\lambda) > 0$$

by (C3), (4.36) with $\epsilon > 0$ sufficiently small implies that on Ω_*, for all large i and $a^i < h_n \le a^{i+1}$,

$$\inf_{\lambda\in\Theta_j\backslash B_\delta(\theta_0)} \log W_{h_n}(\lambda) > K a^{i-1}/(2L). \tag{4.37}$$

Note that

$$\inf_{\lambda\in\Theta_j} \max\{U_{h_n}(\lambda), W_{h_n}(\lambda)\} \ge \min\{\inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} U_{h_n}(\lambda), \inf_{\lambda\in\Theta_j\backslash B_\delta(\theta_0)} W_{h_n}(\lambda)\}.$$

By (4.19), for sufficiently small ϵ,

$$0 \le \inf_{\lambda\in\Theta_j\cap B(\theta_0)} \sum_{\ell\notin J(\theta_0)} c_\ell(\theta_0) I^{g_\ell}(\theta_0,\lambda) - \inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} \sum_{\ell\notin J(\theta_0)} c_\ell(\theta_0) I^{g_\ell}(\theta_0,\lambda) \le \epsilon^{3/4}. \tag{4.38}$$

From (4.27), (4.30), (4.37), and (4.38) with ϵ sufficiently small, it follows that on $\Omega\cap\Omega_*$, for all large i, the certainty-equivalent testing phase between times a^i and a^{i+1} rejects H_j at time $\nu_i(m)$ with $(c_j(\theta_0) - \epsilon)\log a^i \le \tau_{\nu_i(m)}(g_j) \le (c_j(\theta_0) + \sqrt{\epsilon})\log a^i$ for every $j \notin J(\theta_0)$ (or equivalently $\theta_0 \notin \Theta_j$). In particular, the upper bound for $\tau_{\nu_i(m)}(g_j)$ follows from the lower bound in (4.27) together with (3.11), (4.37), and (4.30), noting that the $\epsilon^{3/4}$ in (4.38) is much smaller than $\sqrt{\epsilon}$ if ϵ is sufficiently small and that $\inf_{\lambda\in\Theta_j\cap B(\theta_0)} \sum_{\ell\notin J(\theta_0)} c_\ell(\theta_0) I^{g_\ell}(\theta_0,\lambda) \ge 1$ if $B(\theta_0) \neq \emptyset$ by (3.2). Hence on $\Omega\cap\Omega_*$, for all large i, the evenly allocated testing phase between times a^i and a^{i+1} is applied only to controls g_ℓ with $\ell \in J(\theta_0)$. Since ϵ can be arbitrarily small, this implies that $T_n(g_j) = \tau_n(g_j) + O(1)$ and that $T_n(g_j)/\log n \to c_j(\theta_0)$ a.s. $[\mathbf{P}_x^{\phi^*}]$ for every $j \notin J(\theta_0)$. This also implies that with probability 1, for all large i, ϕ^* only uses rules from G_J after certainty-equivalent testing between times a^i and a^{i+1}. Hence in view of (4.27) and the use of the same stationary control law for an entire block (of stages) B_m^i, of size $\ge n_i \sim i/\log i$, the desired conclusion on $\sum_1^n I_{\{\phi_i^* \neq \phi_{i-1}^*, \phi_i^* \notin G_J \text{ or } \phi_{i-1}^* \notin G_J\}}$ follows.

Proof of Lemma 6. Fix $j \notin J(\theta_0)$. The evenly allocated testing phase of ϕ^*, which was shown to use eventually only controls from G_J in the proof of the a.s. convergence of $T_n(g_j)/\log n$ in Lemma 5, will play a crucial role here in establishing uniform integrability of $T_n(g_j)/\log n$. Also the assumption $\beta > 2$ will be important here. Let $\tau_t = \tau_{a^{t+1}}(g_j)$ and $\widetilde{\tau}_t = T_{a^{t+1}}(g_j) - \tau_t$. It suffices to show that

$$\{\tau_t/t, t \ge 1\} \text{ and } \{\widetilde{\tau}_t/t, t \ge 1\} \text{ are uniformly integrable under } \mathbf{P}_x^{\phi^*}. \tag{4.39}$$

Take any $\epsilon > 0$ and define A_i as in (4.22). In the line above (4.22) we have shown that

$$\mathbf{P}_x^{\phi^*}(\Delta_t) \geq 1 - O(t^{-(\beta-1)}(\log t)^{2\beta}), \text{ where } \Delta_t = \cap_{i \geq [t/\log t]} A_i. \tag{4.40}$$

From the constraint (3.10a) or (3.10b) in the certainty-equivalent testing phase, it follows that

$$\tau_t \leq (2 \log a^t) \log t + 3n_t. \tag{4.41}$$

Moreover, in view of (4.41) together with (3.10a,b) and (4.22) we have, for all large t,

$$\tau_{[t/\log t]} < 3t \log a \text{ and } \tau_i < 3t(c_j(\theta_0) + \epsilon) \log a \text{ for all } [t/\log t] < i \leq t \text{ on } \Delta_t.$$

Hence for all large t,

$$\tau_t/t \leq (3 \log a)(\log t) I_{\Delta_t^c} + (3 \log a)(c_j(\theta_0) + \epsilon) I_{\Delta_t}. \tag{4.42}$$

Since $\mathbf{P}_x^{\phi^*}(\Delta_t^c) = O(t^{-(\beta-1)}(\log t)^{2\beta})$ by (4.40), the uniform integrability of $\{\tau_t/t, t \geq 1\}$ follows from (4.42).

Labeling the elements of $\{\nu_i(m) : i \geq 1, 1 \leq m \leq m(i)\}$ as $s_1 < s_2 < \cdots$ (test times), define

$$\sigma_t = \sup \left\{ s_n \leq a^{t+1} : \inf_{\lambda \in \Theta_j \cap B_\delta(\theta_0)} \widetilde{W}_{s_n,j}(\lambda) < t a^t \right\}, \quad \#_{t,1} = T_{\sigma_t}(g_j), \tag{4.43}$$

$$\#_{t,2} = \sum_{i=1}^{t} I_{\Omega_i \cap \Lambda_i} \sum_{m=1}^{m(i)} 2n_i I_{\{\inf_{\lambda \in \Theta_j \setminus B_\delta(\theta_0)} W_{\nu_i(m)}(\lambda) < i a^i\}}, \tag{4.44}$$

$$\#_{t,3} = \sum_{i=1}^{t} (a^{i+1} - a^i)(I_{\Omega_i^c} + I_{\Lambda_i^c}), \tag{4.45}$$

where Ω_i and Λ_i are defined in (4.32) and (4.34). Since $\nu_i(m) - \nu_i(m-1) \leq 2n_i$ and since

$$\inf_{\lambda \in \Theta_j} \max(\widetilde{W}_{n,j}(\lambda), W_n(\lambda)) \geq \min \left\{ \inf_{\lambda \in \Theta_j \cap B_\delta(\theta_0)} \widetilde{W}_{n,j}(\lambda), \inf_{\lambda \in \Theta_j \setminus B_\delta(\theta_0)} W_n(\lambda) \right\},$$

it follows from (3.12), (4.15), and (4.16) that

$$\widetilde{\tau}_t (= T_{a^{t+1}}(g_j) - \tau_t) \leq \#_{t,1} + \#_{t,2} + \#_{t,3}. \tag{4.46}$$

By (4.31) and (4.35) with $\beta > 2$, $\mathbf{E}_x^{\phi^*}(\#_{t,3}) = \sum_{i=1}^{t} O(i^{-1}) = O(\log t)$. Since $\#_{t,3} \geq 0$ and $\mathbf{E}_x^{\phi^*}(t^{-1}\#_{t,3}) \to 0$ as $t \to \infty$, it follows that $\{\#_{t,3}/t, t \geq 1\}$ is uniformly integrable under $\mathbf{P}_x^{\phi^*}$.

To prove that $\{\#_{t,1}/t, t \geq 1\}$ is uniformly integrable under $\mathbf{P}_x^{\phi^*}$, take $\epsilon > 0$, choose δ by (4.19) and define $\theta_1, \ldots, \theta_K$ as in (4.3). Letting $\Gamma_k = \{\theta : \rho(\theta_k, \theta) \leq \delta'_{\theta_k}\}$, we shall modify the use of (4.23) and (4.26) in the proof of Lemma 5 by introducing

$$\begin{aligned} \sigma^* = \sup \left\{ T_{s_n}(g_j) : \max_{0 \leq k \leq K} \sup_{\theta \in \Gamma_k} \sum_{s=1}^{s_n} |\ell_s(\theta) - \ell_s(\theta_k)| I_{\{\phi^*_{s-1} = g_j\}} > 2\epsilon T_{s_n}(g_j) \right\} \\ \vee \sup \left\{ T_{s_n}(g_j) : \max_{0 \leq k \leq K} \left| (T_{s_n}(g_j))^{-1} \sum_{s=1}^{s_n} \ell_s(\theta_k) I_{\{\phi^*_{s-1} = g_j\}} - I^{g_j}(\theta_0, \theta_k) \right| > 2\epsilon \right\}. \end{aligned} \tag{4.47}$$

For $T_{s_n}(g_j) > \sigma^*$, we have

$$\max_{0\le k\le K} \sup_{\theta\in\Gamma_k} \sum_{s=1}^{s_n} |\ell_s(\theta) - \ell_s(\theta_k)| I_{\{\phi^*_{s-1}=g_j\}} \le 2\epsilon T_{s_n}(g_j)$$

and

$$\max_{0\le k\le K} \left| T_{s_n}(g_j))^{-1} \sum_{s=1}^{s_n} \ell_s(\theta_k) I_{\{\phi^*_{s-1}=g_j\}} - I^{g_j}(\theta_0,\theta_k) \right| \le 2\epsilon.$$

Hence an argument similar to that used to derive (4.25) and (4.30) can be used to show that if $T_{s_n}(g_j) > \sigma^*$ then

$$\inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} \log \widetilde{W}_{s_n,j}(\lambda) \ge \left\{ \inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} I^{g_j}(\theta_0,\lambda) - 7\epsilon \right\} T_{s_n}(g_j) + \log w_0, \tag{4.48}$$

where $w_0 = F(\Gamma_0)$. From (4.48) and (4.43) it follows that

$$\#_{t,1}(= T_{\sigma_t}(g_j)) \le \max\{\sigma^*, 1 + \log(ta^t/w_0)/[\inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} I^{g_j}(\theta_0,\lambda) - 7\epsilon]\}, \tag{4.49}$$

noting that $\inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} I^{g_j}(\theta_0,\lambda) - 7\epsilon > 0$ by (C3) and (4.19), provided that ϵ is chosen sufficiently small. By (4.47) and Lemmas 2 and 3, $\sum_{m=1}^\infty \mathbf{P}_x^{\phi^*}\{\sigma^* \ge m\} = \sum_{m=1}^\infty O(m^{-(\beta-1)}) < \infty$ since $\beta > 2$. Therefore $\mathbf{E}_x^{\phi^*}(\sigma^*) < \infty$ and the uniform integrability of $\{\#_{t,1}/t, t \ge 1\}$ follows from (4.49).

To prove the uniform integrability of $\{\#_{t,2}/t, t \ge 1\}$ under $\mathbf{P}_x^{\phi^*}$, recall that (4.36) holds on $\Omega_i \cap \Lambda_i$ for all large i. By (C3) and choosing ϵ sufficiently small, this implies that $\{\#_{t,2}, t \ge 1\}$ is uniformly bounded by some constant.

The case where ϕ^* is replaced by $\widetilde{\phi}$ is even simpler and is similar to the preceding proof of the uniform integrability of $\{\widetilde{\tau}_t/t,\ t \ge 1\}$.

REFERENCES

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

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

[3] R. Agrawal, D. Teneketzis, and V. Anantharam, *Asymptotically efficient adaptive allocation schemes for controlled I.I.D. process: Finite parameter space*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 258–267.

[4] R. Agrawal, D. Teneketzis, and V. Anantharam, *Asymptotically efficient adaptive allocation schemes for controlled Markov chains: Finite parameter space*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 1249–1259.

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

[6] V. Borkar and P. Varaiya, *Adaptive control of Markov chains,* I: *Finite parameter set*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 953–958.

[7] D. Feldman, *Contributions to the "two-armed bandit" problem*, Ann. Math. Statist., 33 (1962), pp. 847–856.

[8] T. L. Graves, *Comparison of Treatments Under Adaptive Treatment Allocation in Clinical Trials and Stochastic Adaptive Control*, Ph.D. dissertation, Department of Statistics, Stanford University, Stanford, CA, 1995.

[9] I. Iscoe, P. Ney, and E. Nummelin, *Large deviations of uniformly recurrent Markov additive processes*, Adv. Appl. Math., 6 (1985), pp. 373–412.

[10] J. L. Jensen, *Saddlepoint expansions for sums of Markov dependent variables on a continuous state space*, Probab. Theory Related Fields, 89 (1991), pp. 181–199.
[11] P. R. Kumar, *A survey of some results in stochastic adaptive control*, SIAM J. Control Optim., 23 (1985), pp. 329–380.
[12] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall, Englewood Cliffs, NJ, 1986.
[13] T. L. Lai, *Certainty equivalence with uncertainty adjustments in stochastic adaptive control*, in Stochastic Theory and Adaptive Control, T. Duncan and B. Pasik-Duncan, eds., Springer-Verlag, New York, 1992, pp. 270–284.
[14] T. L. Lai, *Tail Probability Bounds for Martingales and Markov Random Walks with Applications to Sequential Analysis and Stochastic Control*, Tech. Report, Department of Statistics, Stanford University, Stanford, CA, 1994.
[15] T. L. Lai and H. Robbins, *Asymptotically optimal allocation of treatments in sequential experiments*, in Design of Experiments, T. J. Santner and A. C. Tamhane, eds., Marcel Dekker, New York, 1984, pp. 127–142.
[16] T. L. Lai and H. Robbins, *Asymptotically efficient adaptive allocation rules*, Adv. Appl. Math., 6 (1985), pp. 4–22.
[17] T. L. Lai and S. Yakowitz, *Machine learning and nonparametric bandit theory*, IEEE Trans. Automat. Control, AC-40 (1995), pp. 1199–1209.
[18] P. Mandl, *Estimation and control of Markov chains*, Adv. in Appl. Probab., 6 (1974), pp. 40–60.