APPLICATIONES MATHEMATICAE
25,4 (1999), pp. 339–358

**V. S. BORKAR (Bangalore)**
**S. M. MUNDRA (Suwon)**

# BAYESIAN PARAMETER ESTIMATION AND ADAPTIVE CONTROL OF MARKOV PROCESSES WITH TIME-AVERAGED COST

*Abstract.* This paper considers Bayesian parameter estimation and an associated adaptive control scheme for controlled Markov chains and diffusions with time-averaged cost. Asymptotic behaviour of the posterior law of the parameter given the observed trajectory is analyzed. This analysis suggests a "cost-biased" estimation scheme and associated self-tuning adaptive control. This is shown to be asymptotically optimal in the almost sure sense.

**I. Introduction.** A popular scheme for adaptive control is the so-called "self-tuning" control wherein a parameterized family of system models is presupposed and the parameter is estimated "on-line". One then uses at each time instant that control which would have been the optimal choice for the current value of the system state if the current parameter estimate were the true parameter. Also known as "certainty equivalence", this artificial separation between estimation and control is expected to lead to an asymptotically optimal behaviour in an appropriate sense.

In the context of controlled Markov chains, such a scheme was first introduced in [22] and was shown to be asymptotically optimal for the time-averaged (or "ergodic") cost under a certain "identifiability condition". (See [10] for some extensions.) The latter condition essentially ensures complete model discrimination under arbitrary control policies. It is an extremely strong condition and the scheme may not be optimal in its absence, as was clearly brought out in [12]. This led to various modifications of the basic scheme, such as randomization of the control or estimate [13], [17] and in-

1991 *Mathematics Subject Classification*: 93E20, 62F15.
*Key words and phrases*: Bayesian estimation, cost-biased estimate, adaptive control, time-averaged cost, asymptotic optimality.

troduction of an explicit cost bias in estimation [19], [20], [23]. The latter methodology, which artificially biases the estimation scheme in favour of parameters leading to lower optimal cost, was introduced in [19], [20] and extended in [6] to a very general class of Markov chains. A variant appeared in [23], [8]. Extensions of the results to controlled diffusion processes appear in [11], [7], [9]—the extensions of Mandl's scheme (assuming the "identifiability condition") in [11], extensions of the scheme of [19], [20], [6] in [7] and that of [23], [8] in [9].

Yet another important development in this direction is the work on "asymptotically efficient" control policies [1] wherein one seeks to meet in an asymptotic sense certain precomputed lower bounds on the difference between the actual cost and the true optimum (the "loss"), whatever be the value of the time parameter. This analysis, however, is confined to the finite parameter space and does not seem to extend easily to more general situations.

All these works consider a non-Bayesian framework. However, Bayesian set-up may be more attractive in some circumstances. One reason is the possibility of incorporating to advantage any prior knowledge through one's choice of the prior probability measure on the parameter space. For ergodic cost, this certainly will not affect the ergodic or "long-run average" behaviour of the estimation scheme and hence the cost. But it should improve the transient behaviour of the algorithm.

Secondly, Bayesian schemes offer a naturally recursive structure since the conditional law of the parameter after an additional observation can be computed from that at the preceding instance and the new data via Bayes rule, the very reason why Bayesian formalism is standard in nonlinear filtering. Conventional wisdom suggests that one should convert the problem into a problem with "complete observations" simply by appending the conditional law of the parameter given the observed trajectory as an extra state variable. This works fine for the "expected integral / sum of running cost" kind of problems (finite horizon, infinite horizon discounted cost etc.—see, e.g., [25] or [21], Ch. 11).

For the ergodic cost, however, this is not appealing. The reason is that at least one component of the extended state, i.e. the posterior law of the parameter given the observed trajectory, does not exhibit suitable "recurrence" properties on which the conventional analysis of the ergodic cost problem crucially depends. In fact, it exhibits the opposite kind of asymptotic behaviour, viz., it gets absorbed into a random limit state (i.e., converges). The conventional dynamic programming-based analysis of the ergodic cost problem, if possible at all, prescribes at best the optimal behaviour on the positive recurrent (here, absorbing) part of the state space. Thus it does not tell one what to do in transient states, i.e., in this special set-up, at any

time except in the limit! Add to this the difficulty of pushing through such analysis to continuum state spaces in absence of any Doeblin-type strong recurrence conditions. This suggests that one should try instead an ad hoc self-tuning scheme as in the case of the non-Bayesian framework. The aim of this paper is to propose one such scheme and to prove its almost sure asymptotic optimality.

In a related work, Di Masi and Stettner [16] consider Bayesian adaptive control where they work around the lack of "identifiability" by means other than the use of a cost bias. Specifically, they consider two classes of controls, controls with forcing and controls with randomization, that ensure adequate model discrimination without affecting optimality or near-optimality.

The paper is organized as follows: The next section describes the hypotheses and the adaptive control scheme for discrete Markov chains. Section III provides an analysis of the asymptotic behaviour of the posterior law, which is of independent interest (and is, in fact, the major component of this work). Section IV proves the a.s. asymptotic optimality of the adaptive control scheme. Both these sections and Section II depend on [6] for considerable detail. This, unfortunately, cannot be avoided since the inclusion in toto thereof would make the present paper extremely unwieldy, requiring essentially the reproduction of [6] here almost in its entirety.

Section V gives a brief account of the corresponding results for controlled diffusions. This discussion relies heavily on [7] for details, for the same reason as above.

We conclude this section with some remarks concerning the implementation aspects of this work. It shares with other cost-biased schemes [18], [19], [23] one basic difficulty, viz., its requirement that the optimal cost as a function of parameter be precomputed and stored. Although this computation is "off-line" in principle, it still can be a considerable overhead. A more realistic approach would be to have an "on-line" approximation scheme for the same. One promising possibility is to merge this adaptive control scheme with a stochastic approximation-based "perturbation analysis" as in [14]. We propose this as a promising direction for future research, the present work then just becomes a key step in this larger programme. It should be remarked in this context that stochastic approximation has been effectively used in adaptive control in the recent work on Q-learning [24], where it approximates a variant of the value function rather than the optimal cost. This work, however, is confined to finite action sets.

**II. Control scheme in the discrete case.** We follow the notation of [5], [6]. Let $X_n$, $n = 1, 2, \ldots$, be a controlled Markov chain on the state space $S = \{1, 2, \ldots\}$ with transition matrix

$$P_u^\theta = [[p(i, j, u_i, \theta)]], \qquad i, j \in S,$$

indexed by the control vector $u = [u_1, u_2, \ldots]$ and the unknown parameter $\theta$. Here $u_i \in D(i)$ for some prescribed compact metric space $D(i)$, $i \in S$. By replacing each $D(i)$ by $\prod_k D(k)$ and $p(i,j,\cdot,\theta)$ by its composition with the projection $\prod_k D(k) \to D(i)$ for each $i$, $j$, $\theta$, we may (and do) assume that all $D(i)$'s are replicas of a fixed compact metric space $D$. The parameter $\theta$ takes values in a compact subset $A$ of $\mathbb{R}^m$, $m \geq 1$, containing a distinguished element $\theta_0$, the true parameter. The actual system is assumed to correspond to $\theta_0$ which is unknown. Denote by $P^\theta(\cdot)$, $E_\theta(\cdot)$ the probabilities and expectations under $\theta \in A$, dropping the subscript $\theta$ when $\theta = \theta_0$. The functions $p(i,j,\cdot,\cdot)$ are assumed to be continuous and Lipschitz in the last argument uniformly with respect to the rest. Fix $\theta \in A$ for the time being. We now introduce the key terminology to be followed throughout.

(1) $P(Y)$: For any Polish (i.e., separable and metrizable with a complete metric) space $Y$, $P(Y)$ will denote the Polish space of probability measures on $Y$ with the topology of weak convergence.

(2) CS: A *control strategy* (CS for short) is a sequence $\{\xi_n\}$, $\xi_n = [\xi_n(1), \xi_n(2), \ldots]$, of $D^\infty$-valued random variables such that for $i \in S$ and $n \geq 0$,

$$P_\theta(X_{n+1} = i \,|\, X_m, \xi_m, m \leq n) = p(X_n, i, \xi_n(X_n), \theta). \tag{2.1}$$

We say that $\{X_n\}$ is *governed by the control strategy* $\{\xi_n\}$ whenever (2.1) holds.

(3) SRS $\gamma[\Phi]$: If $\xi_n$ is independent of $X_m$, $m \leq n$, and of $\xi_m, m < n$, for each $n$, and $\{\xi_n\}$ are identically distributed, call the CS a *stationary randomized strategy* (SRS). If the common law of each $\xi_n$ therein is $\Phi \in P(D^\infty)$, we denote the SRS as $\gamma[\Phi]$. As argued in [6] we may take $\Phi$ to be a product measure $\prod_i \widehat{\phi}_i$ with $\widehat{\phi}_i \in P(D)$ for each $i$. Conversely, each such measure can be identified with an SRS.

(4) SS $\gamma\{\xi\}$: If $\Phi$ is a Dirac measure at $\xi \in D^\infty$, call the corresponding SRS a *stationary strategy* (SS), denoted by $\gamma\{\xi\}$.

(5) $P^\theta[\Phi]$, $P^\theta\{\xi\}$: Under an SRS (resp. SS), $\{X_n\}$ is a Markov chain with stationary transitions, the transition matrix being given by

$$P^\theta[\Phi] = [[p^\theta_\Phi(i,j)]] = \left[\left[\int p(i,j,\xi,\theta)\ \widehat{\phi}_i(d\xi)\right]\right], \qquad i,j \in S$$
$$[\text{resp. } P^\theta\{\xi\} = P^\theta_\xi].$$

We assume throughout that $S$ is a single communicating class under each $\gamma[\Phi]$.

(6) SSRS, SSS: If the resulting chain is positive recurrent, we call the SRS a *stable SRS* (*SSRS*) or if it is a SS, we call it a *stable SS* (*SSS*).

(7) $\Pi[\Phi], \Pi\{\xi\}$: Under an SSRS (resp. SSS), the chain will have a unique invariant probability measure denoted by

$$\Pi^\theta[\Phi] = [\Pi^\theta[\Phi](1), \Pi^\theta[\Phi](2), \ldots]$$
$$[\text{resp. } \Pi^\theta\{\xi\} = [\Pi^\theta\{\xi\}(1), \Pi^\theta\{\xi\}(2), \ldots]].$$

(8) $\widehat{\Pi}[\Phi], \widehat{\Pi}\{\xi\}$: Define $\widehat{\Pi}^\theta[\Phi] \in P(S \times D)$ by

$$\int f \, d\widehat{\Pi}^\theta[\Phi] = \sum_{i \in S} \int f(i,\xi) \, \widehat{\phi}_i(d\xi) \, \Pi^\theta[\Phi](i), \qquad f \in C_{\mathrm{b}}(S \times D).$$

$\widehat{\Pi}^\theta\{\xi\}$ is defined analogously.

In the foregoing and in what follows, we may drop the subscript $\theta$ when $\theta = \theta_0$.

Let $k : S \times D \to \mathbb{R}^+$ be a continuous "cost" function. The *ergodic* or *long run average cost control problem* is to a.s. minimize over all the CS the quantity

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} k(X_m, \xi_m(X_m)). \tag{2.2}$$

Under an SSRS $\gamma[\Phi]$ or an SSS $\gamma\{\xi\}$ and with $\theta$ as the operative parameter, (2.2) a.s. equals

$$\int k \, d\widehat{\Pi}^\theta[\Phi] \tag{2.3}$$

in the former case and the same with $\widehat{\Pi}^\theta\{\xi\}$ replacing $\widehat{\Pi}^\theta[\Phi]$ in the latter. If $\theta$ were known, this would be the classical ergodic control problem. Since it is not, one has to take recourse to some adaptive control scheme such as the self-tuning. Our variant in the Bayesian set-up is as follows.

Under the hypotheses we shall be making later on in this section, it is possible to find a measurable $\nu : A \times S \to D$ such that the SS $\gamma\{\xi\}$ given by $\xi(\cdot) = \nu(\theta, \cdot)$ is an optimal SSS when $\theta$ is the operative parameter (Lemma 2.1 below). Let $\beta(\theta)$ be the corresponding, i.e. the optimal, cost (2.3). Then $\theta \to \beta(\theta)$ is continuous (Lemma 2.1 below). Let $\mu_0(d\theta)$ be the prior probability on $A$, with $\theta_0 \in \mathrm{supp}(\mu_0)$. In other words, we view $\theta_0$ as the actual realization of an $A$-valued random variable $\eta$ with law $\mu_0$, such that the regular conditional law of $\{X_n, n \geq 0\}$ given $\eta = \theta$ is the law of the controlled Markov chain described above with $\theta$ as the operative parameter. (This is precisely the Bayesian paradigm.) Let $X^n = [X_0, X_1, \ldots, X_n]$ for $0 \leq n < \infty$, and $X^\infty = [X_0, X_1, \ldots]$.

Let $\mu_n(d\theta \,|\, X^n)$, $0 \leq n \leq \infty$, be the posterior law of $\eta$ given the observed trajectory $X^n$. We shall see below (equation (3.2)) that $\mu_n \ll \mu_0$ a.s. with Radon–Nikodym derivative, say, $\alpha_n(\cdot)$ for $n = 0, 1, \ldots$ Let

$$A_n = \{\theta \in A \mid \alpha_n(\theta) \geq 1/n\}, \qquad n = 1, 2, \ldots,$$

$$\widetilde{\theta}_n = \text{argmin}_{A_n}\, \beta(\cdot)$$

with any tie for the argmin resolved according to some fixed priority rule. Let $\{y(n)\}$ be a prescribed increasing sequence of positive integers such that $\sum_n y(n)^{-l} < \infty$ for some $l \geq 1$. Define the stopping times $\tau_n$, $n \geq 1$, by

$$\tau_1 = 0,$$
$$\tau_n = (\min\{m > \tau_{n-1} \mid X_m = 1\}) \wedge (\tau_{n-1} + y(n)).$$

Define $\widehat{\theta}_n = \widetilde{\theta}_{[n]}$ where $[n]$ = the largest $\tau_i$ not exceeding $n$, for $n \geq 0$. Our adaptive control strategy $\{\xi_n\}$ will be

$$\xi_n(i) = \nu(\widehat{\theta}_n, i), \qquad i \in S, \tag{2.4}$$

where $\nu(\cdot,\cdot)$ is as described earlier. We shall prove its a.s. asymptotic optimality under suitable assumptions.

Our first assumption will be the following.

Assumption A1. There exist $\Delta_{ij} > 0$, $i, j \in S$, such that for all $\xi, \theta$ either $p(i,j,\xi,\theta) = 0$ or $p(i,j,\xi,\theta) > \Delta_{ij}$. Assume that $I\{p(i,j,\xi,\theta_0) > 0\} \times \ln[p(i,j,\xi,\theta)/p(i,j,\xi,\theta_0)]$ for $\theta \in A$, $\xi \in D$, $i, j \in S$ is bounded uniformly in $i, j, \xi, \theta$ and Lipschitz continuous in $\theta$ uniformly with respect to $i, j, \xi$.

As remarked in [6], this assumption is rather restrictive as it stands, but could be relaxed to a good extent at the expense of a lot more technicalities in the proofs of [6] and here. Consider the following two conditions.

Condition C1. For each $i \in S$, there exists a finite $R_i \subset S$ such that $p(i,j,\cdot,\cdot) \equiv 0$ for $j \notin R_i$.

Condition C2. For any finite $S_1 \subset S$ and $M \geq 1$, there exists an integer $N \geq 1$ such that for $i \geq N$ the length of the minimum path from $i$ to any state in $S_1$ exceeds $M$ under any SRS.

Our second assumption is the following.

Assumption A2. At least one of the following two sets of alternative hypotheses holds.

(A2a) *Lyapunov condition*: Condition (C1) holds and there is an $\omega : S \to \mathbb{R}^+$ such that

(i) $\omega(i) \to \infty$ as $i \to \infty$.

(ii) There exist $a, \varepsilon > 0$ such that under any CS and any $\theta$,

$$E_\theta[(\omega(X_{n+1}) - \omega(X_n) + \varepsilon) I\{\omega(X_n) > a\} \mid F_n] \leq 0 \tag{2.5}$$

for $n \geq 1$ where $F_n = \sigma(X_i, \xi_i, i \leq 1)$.

(iii) There exists a random variable $Z$ and a scalar $\lambda > 0$ such that $E[\exp(\lambda Z)] < \infty$ and for any $c \in \mathbb{R}$ and CS and any $\theta \in A$,

$$P^\theta(|\omega(X_{n+1}) - \omega(X_n)| > c) \leq P(Z > c), \qquad n \geq 1.$$

(A2b) *Near-Monotonicity Condition*: Conditions (C1), (C2) hold and $k$ is near-monotone; i.e.,

$$\liminf_{i\to\infty} \inf_{\xi} k(i,\xi) > \sup_{\theta} \beta(\theta).$$

In addition there exist $\omega_1 : S \to \mathbb{R}^+$, $a_1, \varepsilon_1, \lambda_1 > 0$ and a random variable $Z_1$ such that $\omega_1, a_1, \varepsilon_1, \lambda_1, Z_1$ satisfy the analog of (i)–(iii) above except that (2.5) is now required to hold only when the CS is an SS $\gamma\{\xi\}$ of the type $\xi(\cdot) = \nu(\theta, \cdot)$ for some $\theta \in A$.

Conditions (i)–(iii) above are fashioned after [18]. See [18], [6], [10] for a further discussion.

The conditions above are essentially motivated by queuing applications. Consider, e.g., the simple example of a routing problem wherein packets (customers) arrive in discrete time slots, independently, at most one at a time, with the probability of a packet being present in a given slot being $p > 0$. These are to be routed each to one of two servers. The $i$th server, $i = 1, 2$, when busy, completes service in a given time slot with probability $q_i > 0$. Assume $q_1, q_2 > p$, ensuring stability. The problem is to find the optimal routing scheme for ergodic control with running cost = the sum of queue lengths at the two servers. The adaptive element enters if we suppose that $q_i$'s are unknown except for the information that $p < a < q_1$, $q_2 < b < 1$ for some prescribed $a, b$. This problem satisfies both (A2a), (A2b), the latter with the Lyapunov function $w$ being the sum of queue lengths.

We list below without proof some of the consequences of our assumptions.

LEMMA 2.1. *An optimal SSS exists under any $\theta$. Furthermore, there exists a measurable map $\nu : A \times S \to D$ such that $\xi(\cdot) = \nu(\theta, \cdot)$ is an optimal SSS under $\theta$ for each $\theta \in A$. Also the map $\theta \to \beta(\theta)$ is continuous.*

See [6], p. 296 and p. 306 for details. Let $\theta_0$ be the operative parameter from now on. Define the $P(S \times D \times A)$-valued random sequence $\{\overline{\mu}_n\}$ by

$$\overline{\mu}_n(A_1 \times A_2 \times A_3) = \frac{1}{n} \sum_{m=1}^{n} I\{X_m \in A_1,\ \xi_m(X_m) \in A_2,\ \widehat{\theta}_m \in A_3\}$$

for $A_1, A_2, A_3$ Borel in $S, D, A$ respectively. Let $\nu_n \in P(S \times D)$ be the image of $\overline{\mu}_n$ under the projection $S \times D \times A \to S \times D$.

LEMMA 2.2. *Almost surely, $\{\overline{\mu}_n\}$, $\{\nu_n\}$ are tight sequences and any limit point of $\{\nu_n\}$ is of the type $\widehat{\Pi}[\Phi]$ for some SRS $\gamma[\Phi]$.*

The first claim for $\{\mu_n\}$ (which implies that for $\{\nu_n\}$) is Lemma 4.1 of [6] and the second claim is Lemma 10.3 of [5].

**III. Asymptotic behaviour of Bayes estimates.** Recall the definition of $\mu_n(d\theta \mid X^n)$, $n = 1, 2, \ldots, \infty$. Elementary martingale convergence

arguments show that

$$\mu_n(d\theta \,|\, X^n) \to \mu_\infty(d\theta \,|\, X^\infty) \quad \text{a.s. in } P(A). \tag{3.1}$$

In this section we characterize the support of $\mu_\infty(d\theta \,|\, X^\infty)$, almost surely. Define the following random subsets of $A$:

$$B_1(X^\infty) = \left\{\theta \in A \,\middle|\, \sum_{j\in S} p(X_k, j, \xi_k, \theta_0) \ln(p(X_k, j, \xi_k, \theta)/p(X_k, j, \xi_k, \theta_0)) \to 0\right\},$$

$$B_2(X^\infty) = \left\{\theta \in A \,\middle|\, \frac{1}{n}\sum_{k=0}^{n-1}\sum_{j\in S} p(X_k, j, \xi_k, \theta_0) \ln(p(X_k, j, \xi_k, \theta)/p(X_k, j, \xi_k, \theta_0)) \to 0\right\}.$$

A simple application of the Bayes rule gives

$$\mu_n(d\theta \,|\, X^n) = \alpha_n(\theta)\mu_0(d\theta), \quad n \geq 1, \tag{3.2}$$

where

$$\alpha_n(\theta) = \Lambda_n(\theta)/\textstyle\int \Lambda_n(\theta')\,\mu_0(d\theta') \tag{3.3}$$

with

$$\Lambda_n(\theta) = \prod_{k=0}^{n-1} p(X_k, X_{k+1}, \xi_k(X_k), \theta)/p(X_k, X_{k+1}, \xi_k(X_k), \theta_0)$$

being the likelihood ratio. Let

$$M_k(\theta) = \ln(p(X_k, X_{k+1}, \xi_k(X_k), \theta)/p(X_k, X_{k+1}, \xi_k(X_k), \theta_0))$$

for $k \geq 0$. Then

$$B_1(X^\infty) = \{\theta \in A \mid E[M_k(\theta) \,|\, X^k] \to 0\},$$

$$B_2(X^\infty) = \left\{\theta \in A \,\middle|\, \frac{1}{n}\sum_{k=0}^{n-1} E[M_k(\theta) \,|\, X^k] \to 0\right\}.$$

Lemma 3.1.

$$\sup_\theta \left| \frac{1}{n}\sum_{k=0}^{n-1}(M_k(\theta) - E[M_k(\theta) \,|\, X^k]) \right| \to 0 \quad \textit{a.s.} \tag{3.4}$$

Proof. (A1) implies that $\sup_{k,\theta} E[M_k(\theta)^2 \,|\, X^k] < \infty$ a.s., leading to

$$\sup_{k,\theta} E[(M_k(\theta) - E[M_k(\theta) \,|\, X^k])^2 \,|\, X^k] < \infty \quad \text{a.s.}$$

Thus the strong law for large number of martingales ([15], p. 244) can be

used to deduce

$$\frac{1}{n}\sum_{k=0}^{n-1}(M_k(\theta) - E[M_k(\theta)\,|\,X^k]) \to 0 \qquad \text{a.s.}$$

for each $\theta \in A$. The claim now follows from the uniform Lipschitz continuity part of (A1). ∎

LEMMA 3.2. *$E[M_k(\theta)\,|\,X^k] \le 0$ a.s., $\theta \in A$, $k \ge 0$.*

Proof. From conditional Jensen's inequality applied to the convex function $x \to x\ln(x)$, one has

$$\begin{aligned}
E[M_k(\theta)\,|\,X^k] &= \sum_{j\in S} p(X_k, j, \xi_k, \theta_0)\ln(p(X_k, j, \xi_k, \theta)/p(X_k, j, \xi_k, \theta_0)) \\
&= -\sum_{j\in S} p(X_k, j, \xi_k, \theta)[p(X_k, j, \xi_k, \theta_0)/p(X_k, j, \xi_k, \theta)] \\
&\quad \times \ln(p(X_k, j, \xi_k, \theta_0)/p(X_k, j, \xi_k, \theta)) \\
&\le -\Big[\sum_{j\in S} p(X_k, j, \xi_k, \theta)[p(X_k, j, \xi_k, \theta_0)/(p(X_k, j, \xi_k, \theta)]\Big] \\
&\quad \times \ln\Big[\sum_{j\in S} p(X_k, j, \xi_k, \theta)[p(X_k, j, \xi_k, \theta_0)/p(X_k, j, \xi_k, \theta)]\Big] \\
&= -1\ln(1) = 0. \ \blacksquare
\end{aligned}$$

In particular, we have

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} E[M_k(\theta)\,|\,X^k] \le 0 \qquad \text{a.s.} \tag{3.5}$$

THEOREM 3.1. *$\mu_\infty(B_2(X^\infty)\,|\,X^\infty) = 1$ a.s.*

Proof. It suffices to prove that a.s.

$$\bar\theta \notin B_2(X^\infty) \Rightarrow \bar\theta \notin \operatorname{supp}(\mu_\infty).$$

Consider a sample point outside the zero probability set where (3.1), (3.4), (3.5) fail. Suppose $\bar\theta \notin B_2(X^\infty)$. Then by (3.5) there exist $\varepsilon > 0$ and a subsequence $\{n(m)\}$ of $\{n\}$ such that

$$\frac{1}{n(m)}\sum_{k=0}^{n(m)-1} E[M_k(\bar\theta)\,|\,X^k] < -3\varepsilon, \qquad m \ge 1,$$

By (A1) there exists an open neighbourhood $O_1$ of $\bar\theta$ such that

$$\frac{1}{n(m)}\sum_{k=0}^{n(m)-1} E[M_k(\theta)\,|\,X^k] < -2\varepsilon, \qquad \theta \in O_1. \tag{3.6}$$

Since $M_n(\theta_0)$ is identically zero for all $n$, there exists an open neighbourhood $O_2$ of $\theta_0$ such that

$$\frac{1}{n}\sum_{k=0}^{n-1} E[M_k(\theta)\,|\,X^k] > -\varepsilon, \qquad \theta \in O_2. \tag{3.7}$$

Since $\theta_0 \in \mathrm{supp}(\mu_0)$, $\mu_0(O_2) = \delta > 0$. Hence for $\theta \in O_1$,

$$\begin{aligned}
&\Lambda_{n(m)}(\theta) \Big/ \int \Lambda_{n(m)}(\theta')\,\mu_0(d\theta') \\
&= \exp\Big[\sum_{k=0}^{n(m)-1} M_k(\theta)\Big] \Big/ \int \exp\Big[\sum_{k=0}^{n(m)-1} M_k(\theta')\Big]\,\mu_0(d\theta') \\
&\le \exp\left(\sup_\theta \left|\frac{1}{n(m)}\sum_{k=0}^{n(m)-1}(M_k(\theta) - E[M_k(\theta)\,|\,X^k])\right| n(m)\right)\exp(-2\varepsilon n(m)) \\
&\quad \times \left[\exp\left(-\sup_\theta \left|\frac{1}{n(m)}\sum_{k=0}^{n(m)-1}(M_k(\theta) - E[M_k(\theta)\,|\,X^k])\right| n(m)\right)\right. \\
&\qquad\qquad \left.\times \exp(-\varepsilon n(m))\delta\right]^{-1}.
\end{aligned}$$

From (3.4), (3.6), (3.7) it is clear that RHS decreases to zero exponentially, uniformly on $O_1$. In view of (3.1) it follows that $\bar{\theta} \notin \mathrm{supp}(\mu_\infty)$.

Thus $\mu_\infty(B_2(X^\infty)\,|\,X^\infty) = 1$ a.s. ∎

THEOREM 3.2. *For finite $A$, $\mu_\infty(B_1(X^\infty)\,|\,X^\infty) = 1$ a.s.*

Proof. Since $A$ is finite and $\theta_0 \in \mathrm{supp}(\mu_0)$, we have $\mu_0(\{\theta_0\}) = a > 0$. Since $\Lambda_n(\theta_0)$ is identically one,

$$\int \Lambda_n(\theta')\,\mu_0(d\theta') \ge a > 0. \tag{3.8}$$

For each $\theta$, $\{\Lambda_n(\theta),\ n \ge 0\}$ is a nonnegative martingale with respect to $\sigma(X^n)$, $n \ge 0$. Thus it converges a.s. to some $\Lambda_\infty(\theta) \ge 0$. In view of (3.2), (3.3) and (3.8), $\mathrm{supp}(\mu_\infty) = \{\theta \in A \mid \Lambda_\infty(\theta) > 0\}$ a.s. But when $\Lambda_\infty(\theta) > 0$,

$$M_n(\theta) = \ln(\Lambda_{n+1}(\theta)/\Lambda_n(\theta)) \to \ln(\Lambda_\infty(\theta)/\Lambda_\infty(\theta)) = 0.$$

Thus

$$M_n(\theta)\Lambda_n(\theta) \to 0 \quad \text{a.s. for } \theta \in A.$$

Consider

$$\begin{aligned}
E[M_n(\theta)\Lambda_n(\theta)\,|\,X^n] &= E[M_n(\theta)\Lambda_n(\theta) I\{M_n(\theta)\Lambda_n(\theta) \le N\}\,|\,X^n] \\
&\quad + E[M_n(\theta)\Lambda_n(\theta) I\{M_n(\theta)\Lambda_n(\theta) > N\}\,|\,X^n]
\end{aligned}$$

for $n \geq 1$ and $N \geq 1$. The first term on the right goes to zero a.s. as $n \to \infty$ by Theorem 2, p. 883 of [9]. The second term on the right equals

$$\begin{aligned} E[M_n(\theta) I\{M_n(\theta)\Lambda_n(\theta) > N\} \,|\, X^n]\Lambda_n(\theta) \\ &\leq E[M_n^2(\theta) \,|\, X^n]^{1/2} P(M_n(\theta)\Lambda_n(\theta) > N \,|\, X^n)^{1/2} \Lambda_n(\theta) \\ &\leq K P(M_n(\theta)\Lambda_n(\theta) > N \,|\, X^n)^{1/2} \Lambda_n(\theta) \end{aligned}$$

for some $K < \infty$. Now

$$\begin{aligned} P(M_n(\theta)\Lambda_n(\theta) > N \,|\, X^n) &\leq E[|M_n(\theta)|\Lambda_n(\theta) \,|\, X^n]/N \\ &\leq E[M_n^2(\theta) \,|\, X^n]^{1/2} \Lambda_n(\theta)/N \\ &\leq K\Lambda_n(\theta)/N. \end{aligned} \tag{3.9}$$

Since $\Lambda_n(\theta) \to \Lambda_\infty(\theta)$ a.s., one has from (3.9),

$$\limsup_{N\to\infty} \sup_{n\geq 0} E[M_n(\theta)\Lambda_n(\theta) I\{M_n(\theta)\Lambda_n(\theta) > N\} \,|\, X^n] = 0 \quad \text{a.s.}$$

Thus

$$E[M_n(\theta)\Lambda_n(\theta) \,|\, X^n] = E[M_n(\theta) \,|\, X^n]\Lambda_n(\theta) \to 0 \quad \text{a.s.},$$

implying

$$E[M_n(\theta) \,|\, X^n] \to 0 \quad \text{a.s. on } \{\Lambda_\infty(\theta) > 0\}.$$

Equivalently, $\mu_\infty(B_1(X^\infty) \,|\, X^\infty) = 1$ a.s. ∎

We conclude this section with some relevant remarks.

Remark 1. Consider the "identifiability condition": for each $\xi \in D$ and $\theta \neq \theta_0$ in $A$, $p(i,j,\xi,\theta) \neq p(i,j,\xi,\theta_0)$ for some $i,j \in S$. Under this condition, if $X_n = i$ i.o. for all $i \in S$, a.s. (which incidently can be shown to be true under our hypotheses), then $B_1(X^\infty) = B_2(X^\infty) = \{\theta_0\}$ a.s. Thus the Bayes estimation scheme is consistent in the strong sense. This follows easily from the fact that under the above conditions,

$$\sum_{j\in S} p(i,j,\xi_k,\theta_0) \ln(p(i,j,\xi_k,\theta)/p(i,j,\xi_k,\theta_0)) = 0$$

if and only if $\theta = \theta_0$. (Compare with [22], [10].) One may then mimick the arguments of [10] to deduce that the "raw" self-tuning rule $\xi_n(i) = \overline{\nu}(\mu_n(\cdot \,|\, X^n), i)$, $i \in S$, where $\overline{\nu}(\mu,\cdot)$ is the optimal SSS under the transition matrix $\overline{P}^\mu = [[\int p(i,j,\xi,\theta)\ \mu(d\theta)]]$, is optimal. We shall not go into the details of this as they are routine and we are more concerned with the situation where the identifiability condition fails.

Remark 2. The above scheme extends to more general situations as well. Consider, for example, an $\mathbb{R}^d$-valued sequence $\{X_n\}$ of random variables. With $\{X^n\}$ defined as before, let the law of $X^\infty$ belong to a parametrized family $\{P_\theta, \theta \in A\} \subset P(\mathbb{R}^d)$. Let $\theta_0 \in \text{supp}(\mu_0)$. Define $\mu_n(d\theta \,|\, X^n)$, $n =$

$1, 2, \ldots, \infty$ as before. Let $q_\theta(dx \,|\, X^n)$ be the regular conditional law of $X_{n+1}$ given $X^n$ under $P_\theta$. We assume this to have a density $p(n, \theta, x \,|\, x^n) > 0$ for $x \in \mathbb{R}^d$, $x^n \in (\mathbb{R}^d)^n$, for each $n, \theta$. Furthermore, the functions

$$\theta \to \ln(p(n, \theta, x \,|\, x^n)/p(n, \theta_0, x \,|\, x^n))$$

are assumed to be continuous uniformly with respect to $n, x, x^n$ and the following bound is assumed to hold:

$$\sup_{n, x^n, \theta} \int p(n, \theta_0, x \,|\, x^n)[\ln(p(n, \theta, x \,|\, x^n)/p(n, \theta_0, x \,|\, x^n))]^2 \, dx < \infty. \tag{3.10}$$

Let

$$B_1(X^\infty) = \left\{ \theta \in A \;\middle|\; \int p(n, \theta_0, x \,|\, X^n) \ln(p(n, \theta, x \,|\, X^n)/p(n, \theta_0, x \,|\, X^n)) \, dx \to 0 \right\},$$

$$B_2(X^\infty) = \left\{ \theta \in A \;\middle|\; \frac{1}{n} \sum_{m=1}^{n} \int p(n, \theta_0, x \,|\, X^n) \ln(p(n, \theta, x \,|\, X^n)/p(n, \theta_0, x \,|\, X^n)) \, dx \to 0 \right\}.$$

One may then mimick the foregoing to deduce that $\mu_\infty(B_2(X^\infty) \,|\, X^\infty) = 1$ a.s. and if $A$ is finite, this improves to $\mu_\infty(B_1(X^\infty) \,|\, X^\infty) = 1$ a.s. Condition (3.10) here facilitates the application of the martingale strong law of large numbers at the appropriate juncture.

These results have interesting interpretations. From the definition of $B_2(X^\infty)$, what they do imply is that even when the estimation scheme is not consistent, it asymptotically correctly predicts the one step future (in the sense that the Kullback–Leibler mutual information between the estimated one step regular conditional law and the true one approaches zero) along a sequence of time instants that exclude at most a "rare" set thereof in the sense of [13]. The intuitive content of this statement should be clear. We omit a precise statement to avoid a major digression. Suffice it to say that this is reminiscent of "merging of opinions" à la [4] (also, the "consistency in information" of [2]).

Remark 3. It is also interesting to compare these results with the corresponding results for maximum likelihood estimates given in [12], [10]. The latter are defined as $\theta'_n = \operatorname{argmin} \Lambda_n(\theta)$ with any tie for the argmin being settled according to some fixed priority rule. As shown in [13],

$$\theta'_n \to \left\{ \theta \in A \;\middle|\; \liminf_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} E[M_{k+1}(\theta) \,|\, X^k] = 0 \right\} \quad \text{a.s.}$$

Compare this with Theorem 3.1.

**IV. Asymptotic optimality of the adaptive control scheme.** In this section we prove the a.s. asymptotic optimality of the scheme proposed in Section II. The treatment here closely imitates that of [6]. In fact, we shall rely on [6] for some nontrivial details. We shall proceed through a sequence of lemmas.

LEMMA 4.1. *Almost surely* $\beta(\widehat{\theta}_n) \le \beta(\theta_0)$ *from some* $n$ *onwards.*

Proof. From the definition of $\{\widetilde{\theta}_n\}$, it suffices to prove that $\theta_0 \in A_n$ from some $n$ on, a.s. Let

$$\Gamma_n = \int \Lambda_n(\theta')\,\mu_0(d\theta').$$

Thus $\alpha_n(\theta) = \Lambda_n(\theta)/\Gamma_n$, $n \ge 0$. It is easily checked that $(\Gamma_n, \sigma(X^n))$ is a nonnegative martingale. Thus $\Gamma_n \to \Gamma_\infty$ a.s. for some $\Gamma_\infty \ge 0$. Since $\Lambda_n(\theta_0)$ is identically equal to 1, $\alpha_n(\theta_0) \to \infty$ a.s. on $\{\Gamma_\infty = 0\}$ and thus $\alpha_n(\theta_0) \ge 1/n$ from some $n$ on. On $\{\Gamma_\infty > 0\}$, $\alpha_n(\theta_0) \to 1/\Gamma_\infty$ a.s. Since $\Gamma_\infty < n$ for large $n$, $\alpha_n(\theta_0) > 1/n$ and therefore $\theta_0 \in A_n$ from some $n$ on. ∎

LEMMA 4.2. *Almost surely*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \sum_{j\in S} p(X_k, j, \xi_k(X_k), \theta_0) \times \ln(p(X_k, j, \xi_k(X_k), \widetilde{\theta}_n)/p(X_k, j, \xi_k(X_k), \theta_0)) = 0.$$

Proof. Consider a sample point outside the set of zero probability where the conclusions of Lemmas 3.1 and 3.2 fail. If the claim were false for this sample point, there exist $\varepsilon > 0$ and a subsequence $\{n(m)\}$ of $\{n\}$ such that

$$\frac{1}{n(m)} \sum_{k=0}^{n(m)-1} \sum_{j\in S} p(X_k, j, \xi_k(X_k), \theta_0) \times \ln(p(X_k, j, \xi_k(X_k), \widetilde{\theta}_{n(m)})/p(X_k, j, \xi_k(X_k), \theta_0)) < -2\varepsilon.$$

As in the proof of Theorem 3.1 we have

$$\begin{aligned}
&\alpha_{n(m)}(\widetilde{\theta}_{n(m)}) \\
&\le \exp\left( \sup_\theta \left| \frac{1}{n(m)} \sum_{k=0}^{n(m)-1} (M_k(\theta) - E[M_k(\theta) \mid X^k]) \right| n(m) \right) \exp(-2\varepsilon n(m)) \\
&\quad \times \left[ \exp\left( - \sup_\theta \left| \frac{1}{n(m)} \sum_{k=0}^{n(m)-1} (M_k(\theta) - E[M_k(\theta) \mid X^k]) \right| \right) \right. \\
&\qquad \left. \times \exp(-\varepsilon n(m))\delta \right]^{-1}
\end{aligned}$$

where $\delta > 0$ is as in the proof of Theorem 3.1. Thus $\alpha_{n(m)}(\widetilde{\theta}_{n(m)}) \le k_1 \exp(-k_2 n(m))$, $m \ge 1$, for some $k_1, k_2 > 0$ depending on the sample

path. Hence $\alpha_{n(m)}(\widetilde{\theta}_{n(m)}) < 1/n(m)$ from some $m$ on, which contradicts the definition of $\{\widetilde{\theta}_n\}$. This proves the claim. ∎

From now on we closely imitate the arguments of [6]. Call $\bar{\theta} \in A$ a *frequent limit point* of $\{\widehat{\theta}_n\}$ along a given sample path if for any open neighbourhood $B$ of $\bar{\theta}$,

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} I\{\widehat{\theta}_k \in B\} > 0.$$

Consider a sample path outside the set $Q$ of zero probability on which the conclusion of any of the lemmas above fails. Let $\bar{\theta}$ be a frequent limit point of $\{\widehat{\theta}_n\}$ along this sample path. Pick $\{n(k)\} \subset \{n\}$ such that $\widehat{\theta}_{n(k)} = \widetilde{\theta}_{n(k)} \in B_N$ and (see [6], p. 303)

$$\liminf_{k\to\infty} \frac{1}{n(k)} \sum_{m=0}^{n(k)} I\{\widetilde{\theta}_m \in B_N\} > 0,$$

$B_N$ being a ball of radius $1/N$ containing $\bar{\theta}$. Let $\overline{\mu}$ be a limit point of $\{\overline{\mu}_{n(k)}\}$.

LEMMA 4.3. *$\overline{\mu}(B_N \times \{i\} \times D) > 0$ for $i \in S$.*

This is proved exactly along the lines of (20), p. 303 of [6]. (It should be remarked that strictly speaking, one may need to replace $Q$ by a larger set of zero probability. We assume that $Q$ is suitably enlarged so that while still having zero probability it also satisfies: Lemma 4.3 holds on $Q^c$ for all $B_N$ in the collection of open balls with rational radii and rational centres in $A$.) Let $G(i,\theta) = \{u \in D \mid u$ is an optimal choice at state $i$ under parameter $\theta\}$, and

$$G = \bigcup_{i,\theta} \{\theta\} \times \{i\} \times G(i,\theta) \subset A \times S \times D$$

with the relative topology. Note that $G(i,\theta)$, $i \in S$, $\theta \in A$, is well defined due to the fact that the optimal choices at any given state do not depend on the choices elsewhere by virtue of dynamic programming-based characterization thereof—see pp. 295–296 of [6]. It is proved in Lemma 3.1 of [6] that $G$ is closed in $A \times S \times D$.

LEMMA 4.4. *$\overline{\mu}(G) = 1$.*

This is an easy consequence of the facts that $\mu_n(G) = 1$, $n \geq 1$, by our choice of $\{\xi_n\}$ and $G$ is closed. (See Lemma 4.8 of [6].)

LEMMA 4.5. *There exist $\theta_i(N), \theta'(N) \in \overline{B}_N$ and $\overline{\xi}_N \in D^\infty$ such that for $i \in S$,*

$$\overline{\xi}_{N(i)} \in G(i, \theta_i(N)),$$
$$p(i,j,\overline{\xi}_N(i),\theta'(N)) = p(i,j,\overline{\xi}_N(i),\theta_0), \qquad j \in S.$$

P r o o f. Let $\theta'(N)$ be any limit point of $\{\widehat{\theta}_{n(k)}\}$. By Lemma 4.2, it follows that

$$\int\limits_{A\times S\times D} d\overline{\mu}\Big[\sum_{j\in S} p(\cdot,j,\cdot,\theta_0)\ln(p(\cdot,j,\cdot,\theta'(N))/p(\cdot,j,\cdot,\theta_0))\Big] = 0. \tag{4.1}$$

From the strict convexity of the map $x \to x\ln x$, one easily checks that for any $i \in S$, $\theta \in A$, $\xi \in D$,

$$\sum_{j\in S} p(i,j,\xi,\theta_0)\ln(p(i,j,\xi,\theta'(N))/p(i,j,\xi,\theta_0)) \le 0 \tag{4.2}$$

with equality if and only if

$$p(i,j,\xi,\theta') = p(i,j,\xi,\theta_0), \qquad j \in S.$$

(Compare with the proof of Lemma 3.2.) From (4.1), it follows that (4.2) holds with equality $\overline{\mu}$-a.s. The claim now follows from Lemmas 4.3 and 4.4. ∎

Corollary 4.1. $\beta(\bar\theta) = \beta(\theta_0)$.

P r o o f. As $N \to \infty$ in the above, $\theta_i(N), \theta'(N) \to \bar\theta$. Let $\overline{\xi}$ be a limit point of $\{\overline{\xi}_N\}$ in $D^\infty$. From the preceding lemma, the continuity of $p(\cdot,j,\cdot,\cdot)$ and the fact that $G$ is closed, it then follows that

$$\overline{\xi}(i) \in G(i,\bar\theta), \qquad i \in S,$$
$$p(i,j,\overline{\xi}(i),\bar\theta) = p(i,j,\overline{\xi}(i),\theta_0), \qquad i,j \in S.$$

These together imply that the cost of $\nu\{\overline{\xi}\}$ under $\bar\theta$ is $\beta(\bar\theta)$, which in turn equals its cost under $\theta_0$. As the latter must be greater than or equal to $\beta(\theta_0)$, we have $\beta(\bar\theta) \ge \beta(\theta_0)$. Lemma 4.1 completes the proof. ∎

Corollary 4.2. $G(i,\bar\theta) = G(i,\theta_0),\ i \in S.$

This is precisely Corollary 5.1 of [6] and follows as there from Corollary 4.1 above.

Theorem 4.1. *The control strategy $\{\xi_n\}$ above is a.s. optimal.*

P r o o f. Consider a sample path outside $Q$. Let $\overline{\mu}$ be a limit point of $\{\overline{\mu}_n\}$. It is clear that any $\bar\theta$ in the support of the image of $\overline{\mu}$ under the projection $A \times S \times D \to A$ will be a frequent limit point of $\{\widehat{\theta}_n\}$. By Corollary 4.2, $G(i,\bar\theta) = G(i,\theta_0)$ for all $i \in S$. Since $\{\overline{\mu}_n\}$ are supported on $G$, so will be $\overline{\mu}$. Also, the image $\overline{\nu}$ of $\overline{\mu}$ under the projection $A\times S\times D \to S\times D$ is of the form $\widehat{\Pi}[\Phi]$ for some SSRS $\gamma[\Phi]$, $\Phi = \prod \widehat{\phi}_i$, by Lemma 2.2. It follows that $\widehat{\phi}_i$ is supported on $G(i,\theta_0)$ for $i \in S$. The dynamic programming characterization of an optimal SSRS (see [6], p. 295) then implies that $\nu[\Phi]$ is an optimal SSRS. Thus

$$\int k\, d\overline{\mu} = \int k\, d\widehat{\Pi}[\Phi] = \beta(\theta_0).$$

Since $\overline{\mu}$ was an arbitrary limit point of $\{\overline{\mu}_n\}$, the claim follows. ∎

**V. Extensions to continuous time.** In this section we present results analogous to the foregoing for the adaptive control of a diffusion process. Since the details are rather straightforward given the foregoing and [7], we shall only sketch the arguments.

Let $D, A$ be as before. Our control system will be the controlled diffusion $X(\cdot) = [X_1(\cdot), \ldots, X_d(\cdot)]^T$, $d \geq 1$, satisfying the stochastic differential equation

$$X(t) = X_0 + \int m(X(s), u(s), \theta)\, ds + \int \sigma(X(s))\, dW(s). \tag{5.1}$$

Here it is assumed that

(i) $m(\cdot,\cdot,\cdot) = [m_1(\cdot,\cdot,\cdot), \ldots, m_d(\cdot,\cdot,\cdot)]^T : \mathbb{R}^d \times D \times A \to \mathbb{R}^d$ is bounded, continuous and Lipschitz in its first and third arguments, uniformly with respect to the second,

(ii) $\sigma(\cdot) = [[\sigma_{i,j}(\cdot)]]_{1 \leq i,j \leq d} : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ is bounded Lipschitz and satisfies $\|\sigma^T(x)z\|^2 \geq \lambda_0 \|z\|^2$, $\lambda_0 > 0$,

(iii) $X_0$ is a random variable with a prescribed law,

(iv) $W(\cdot) = [W_1(\cdot), \ldots, W_d(\cdot)]^T$ is a $d$-dimensional standard Wiener process independent of $X_0$,

(v) $\theta$ is the parameter whose true value is $\theta_0 \in A$,

(vi) $u(\cdot)$ is a $D$-valued control process with measurable paths satisfying the following nonanticipativity condition: for $t \geq s \geq y$, $W(t) - W(s)$ is independent of $u([0,s])$ and $W([0,s])$. (Here, $f([0,t])$ denotes the entire trajectory $f(y)$, $0 \leq y \leq t$.)

Call such a $u(\cdot)$ an *admissible control*. If there exists a measurable map $v : \mathbb{R} \to D$ such that $u(\cdot) = v(X(\cdot))$ call $u(\cdot)$ (or, by abuse of notation, $v$ itself) a *Markov control*. Markov controls are admissible [7]. A Markov control $v$ is said to be *stable* if the resulting Markov process $X(\cdot)$ is positive recurrent and thus has a unique invariant probability measure, denoted by $\eta_v^\theta$ (see [3]). Let $k \in C_{\mathrm{b}}(\mathbb{R}^d \times D)$ be the "running cost" function. The *ergodic control problem* is to a.s. minimize over all admissible $u(\cdot)$ the cost

$$\limsup_{t \to \infty} \frac{1}{t} \int_0^t k(X(s), u(s))\, ds.$$

Under a stable Markov control $v$, this a.s. equals

$$\int k(x, v(x))\, d\eta_v^\theta \tag{5.2}$$

when $\theta$ is the operative parameter. Let $\beta(\theta)$ denote the infimum of (5.2) over all stable $v$.

We shall assume that one of the following two sets of conditions hold:

(A1′) There exist $w \in C^2(\mathbb{R}^d)$ and $a, \varepsilon > 0$ such that

(i) $0 \leq w(x) \to \infty$ as $\|x\| \to \infty$, uniformly in $\|x\|$,

(ii) $w(\cdot)$ and $\|\nabla w(\cdot)\|$ have polynomial growth,

(iii) for $\|x\| > a$,

$$\|\nabla w(x)\|^2 > \lambda_0^{-1} \quad \text{and} \quad \psi^\theta w(x,u) < -\varepsilon, \quad u \in D,\ \theta \in A, \tag{5.3}$$

where for $x = [x_1, \dots, x_d] \in \mathbb{R}^d$ and $f \in C^2(\mathbb{R}^d)$,

$$\psi^\theta f(x,u) = \frac{1}{2} \sum_{i,j,k} \sigma_{ik}\sigma_{jk} \frac{\partial^2 f}{\partial x_i \partial x_j} + \langle \nabla f(x), m(x,u,\theta) \rangle.$$

(A2′) $k$ is monotone, i.e.,

$$\liminf_{\|x\| \to \infty} \inf_u k(x,u) > \sup_\theta \beta(\theta). \tag{5.4}$$

Also there exist $w_1 \in C^2(\mathbb{R}^d)$ and $a_1, \varepsilon_1 > 0$ such that (i)–(iii) above hold with $w_1, a_1, \varepsilon_1$ in place of $w, a, \varepsilon$ except for (5.3) being replaced by

$$\psi^\theta f(x, v(\theta', x)) < -\varepsilon, \qquad \|x\| > a_1,\ \theta, \theta' \in A,$$

where $v : A \times \mathbb{R}^d \to D$ is a measurable map such that $v(\theta, \cdot)$ is an optimal stable Markov control under $\theta$.

Such a map is known to exist either under (A2′) or (5.4) ([7], p. 124).

Let $\mu_0 \in P(A)$ be as before, thus viewing $\theta_0$ as the actual realization of an $A$-valued random variable $\zeta$ with law $\mu_0$ and independent of $(X_0, W(\cdot))$. Let $X^t = X([0,t])$, $t \geq 0$, $X^\infty = X([0,\infty))$ and let $\mu_t(d\theta \,|\, X^t)$ be the regular conditional law of $\zeta$ given $X^t$ for $t \in [0,\infty)$. As before, $\mu_t(d\theta \,|\, X^t) \to \mu_\infty(d\theta \,|\, X^\infty)$ a.s. in $P(A)$ as $t \to \infty$. Let

$$\begin{aligned} \Lambda_t(\theta) = \exp\Big[ \int_0^t \langle \sigma^{-1}(X(s))(m(X(s),u(s),\theta) - m(x(s),u(s),\theta_0)), dW(s) \rangle \\ - \frac{1}{2} \int_0^t \|\sigma^{-1}(X(s))(m(X(s),u(s),\theta) - m(x(s),u(s),\theta_0))\|^2 \, ds \Big] \end{aligned}$$

for $t \geq 0$. A simple Bayes rule argument using Girsanov's theorem leads to

$$\mu_t(d\theta \,|\, X^t) = \alpha_t(\theta) d\theta, \qquad t \geq 0,$$

with

$$\alpha_t(\theta) = \Lambda_t(\theta) \big/ \textstyle\int \Lambda_t(\theta') \, \mu_t(d\theta').$$

Let $A_0 = A$ and $A_t = \{\theta \in A \mid \alpha_t(\theta) \geq 1/t\}$, $t > 0$. Let

$$\widetilde{\theta}_t = \operatorname{argmin}_{A_t} \beta(\theta)$$

where any tie for the argmin is resolved according to some prescribed priority rule, say lexicographic, which ensures a measurable version of $t \to \widetilde{\theta}_t$. Let $y_n$, $n > 0$, be a prescribed deterministic sequence of positive numbers satisfying $\sum_n y_n^{-l} < \infty$ for some integer $l \geq 1$. Let $0 < r_1 < r_2 < \infty$ and $B_1, B_2$ be

balls of radii $r_1, r_2$ resp. in $\mathbb{R}^d$ with centre at the origin. Let $\partial B_i$, $i = 1, 2$, be the respective boundaries. Define stopping times $\{\tau_i\}$ as follows: $\tau_0 = 0$ and

$$\tau_{n+1} = (\inf\{t > \tau_n \mid X(t) \in \partial B_1 \text{ and } X(s) \in \partial B_2 \text{ for some } s \in [\tau_n, t]\}) \wedge (\tau_n + y_n), \qquad n \geq 0.$$

Let $[t] =$ the $\tau_n$ for which $\tau_n \leq t \leq \tau_{n+1}$. Our adaptive control scheme will be

$$u(t) = v(\widehat{\theta}(t), X(t)), \qquad t \geq 0,$$

where $\widehat{\theta}(t) = \widetilde{\theta}_{[t]}$. Let

$$M_t = \int_0^t \langle \sigma^{-1}(X(s))(m(X(s), u(s), \theta) - m(X(s), u(s), \theta_0)),\, dW(s) \rangle,$$

$$\langle M \rangle_t(\theta) = \int_0^t \|\sigma^{-1}(X(s))(m(X(s), u(s), \theta) - m(X(s), u(s), \theta_0))\|^2\, ds.$$

For each $\theta$, $(M_t(\theta), \sigma(X([0, t])))$, $t \geq 0$, is a zero mean square-integrable martingale with continuous paths and $\langle M \rangle_t(\theta)$, $t \geq 0$, the associated quadratic variation process.

Lemma 5.1. *The map $(\theta, t) \to M_t(\theta)/t$ has a jointly continuous version which is uniformly continuous in $\theta$, uniformly with respect to $t$, and*

$$\lim_{t \to \infty} \sup_{\theta} |M_t(\theta)/t| = 0 \qquad \text{a.s.}$$

Proof. This follows exactly as in Lemmas 5.1 and 5.2, p. 134 of [7]. In particular, it follows that $M_t(\theta) = o(\langle M \rangle_t(\theta))$ a.s. on $\{\langle M \rangle_\infty(\theta) = \infty\}$. ∎

Define $\langle M \rangle_\infty(\theta) = \lim_{t \to \infty} \langle M \rangle_t(\theta)$ (possibly $\infty$) and

$$B_1(X^\infty) = \{\theta \in A \mid \langle M \rangle_\infty(\theta) < \infty\},$$
$$B_2(X^\infty) = \{\theta \in A \mid \langle M \rangle_t(\theta)/t \to 0\}.$$

Theorem 5.1. *Almost surely, $\mu_\infty(B_2(X^\infty) \mid X^\infty) = 1$. For finite $A$, this can be improved to $\mu_\infty(B_1(X^\infty) \mid X^\infty) = 1$.*

Proof. The first claim follows as in Theorem 3.1 in view of the preceding lemma. For the second claim, as in Theorem 3.2, $\mu_\infty(\cdot \mid X^\infty)$ is supported on $H = \{\theta \mid \Lambda_\infty(\theta) = \lim_{t \to \infty} \Lambda_t(\theta) > 0\}$ a.s. Since for each $\theta$, $M_t(\theta)$ converges a.s. on $\{\langle M \rangle_\infty(\theta) < \infty\}$ (see Lemma 5.1 of [7]), it follows that

$$\Lambda_t(\theta) = \exp\big[-\tfrac{1}{2}\langle M \rangle_t(\theta)(1 - 2M_t(\theta)/\langle M \rangle_t(\theta))\big] \to \exp\big(\lim_{t \to \infty} M_t(\theta) - \tfrac{1}{2}\langle M \rangle_\infty(\theta)\big)$$

a.s. on $\{\langle M \rangle_\infty(\theta) < \infty\}$ and tends to 0 a.s. on $\{\langle M \rangle_\infty(\theta) = \infty\}$.

Thus $H = B_1(X^\infty)$ a.s. ∎

This describes the asymptotic behaviour of the Bayes scheme in continuous time case along the lines of Section III. Coming back to the adaptive control scheme we have:

LEMMA 5.2.

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \|\sigma^{-1}(X(s))(m(X(s), u(s), \widetilde{\theta}_t) - m(X(s), u(s), \theta_0))\|^2 \, ds = 0 \text{ a.s.}$$

This again follows along the lines of Lemma 4.2 using Lemma 5.1 above. Finally, we have the following analog of Lemma 4.1.

LEMMA 5.3. *Almost surely,* $\beta(\widetilde{\theta}_t) \leq \beta(\theta_0)$ *from some* $t$ *onwards.*

Proof. It suffices to prove that almost surely, $\theta_0 \in A_t$ from some $t$ onwards. This follows as in Lemma 4.1. ∎

The rest of the argument leading to a.s. optimality of our adaptive control scheme imitates Section IV, the details being supplied by [7]. First, one deduces as in Section 4 of [7] that almost surely $\beta(\bar{\theta}) = \beta(\theta_0)$ for any frequent limit point $\bar{\theta}$ of $\{\widehat{\theta}_t\}$, Lemma 5.2 above playing the role of Lemma 4.2, p. 125 of [7]. In view of Lemma 5.3 above, one then has $\beta(\bar{\theta}) = \beta(\theta_0)$, which replaces Lemma 5.3, p. 135 of [7]. The rest of the proof is identical to that of [7], pp. 135–136, leading to:

THEOREM 5.2. *The adaptive control scheme proposed here is a.s. optimal.*

## References


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

[2] A. Barron, *Are Bayes rules consistent in information?*, in: *Problems in Communication and Computation*, T. M. Cover and B. Gopinath (eds.), Springer, New York, 1987, 85–91.

[3] R. N. Bhattacharya, *Asymptotic behaviour of several dimensional diffusions*, in: *Stochastic Nonlinear Systems*, L. Arnold and R. Lefever (eds.), Springer, New York, 1981, 86–91.

[4] D. Blackwell and L. Dubins, *Merging of opinions with increasing information*, Ann. Math. Statist. 33 (1962), 882–887.

[5] V. S. Borkar, *Control of Markov chains with long run average cost criterion*, in: *Stochastic Differential Systems*, *Stochastic Control Theory and Applications*, W. H. Fleming and P. L. Lions (eds.), Springer, New York, 1987, 57–77.

[6] V. S. Borkar, *The Kumar–Becker–Lin scheme revisited*, J. Optim. Theory Appl. 66 (1990), 289–309.

[7] —, *Self-tuning control of diffusions without the identifiability condition*, ibid. 68 (1991), 117–137.

[8] —, *On the Milito–Cruz adaptive control scheme for Markov chains*, ibid. 77 (1993), 387–397.

[9] V. S. Borkar, *A modified self-tuner for controlled diffusions with an unknown parameter*, in: *Mathematical Theory of Control* (*Bombay*, *1990*), A. V. Balakrishnan and M. C. Joshi (eds.), Marcel Dekker, 1992, 57–67.
[10] V. S. Borkar and M. K. Ghosh, *Ergodic and adaptive control of nearest neighbour motions*, Math. Control Signals and Systems 4 (1991), 81–98.
[11] —, —, *Ergodic control of multidimensional diffusions II*: *adaptive control*, Appl. Math. Optim. 21 (1990), 191–220.
[12] V. S. Borkar and P. P. Varaiya, *Identification and adaptive control of Markov chains I*: *finite parameter case*, IEEE Trans. Automatic Control 24 (1979), 953–957.
[13] —, —, *Identification and adaptive control of Markov chains*, SIAM J. Control Optim. 20 (1982), 470–488.
[14] E. K. P. Chong and P. J. Ramadge, *Stochastic optimization of regenerative systems using infinitesimal perturbation analysis*, IEEE Trans. Automatic Control 39 (1994), 1400–1410.
[15] Y. S. Chow and H. Teicher, *Probability Theory*: *Independence*, *Interchangeability*, *Martingales*, Springer, New York, 1979.
[16] G. B. Di Masi and Ł. Stettner, *Bayesian ergodic adaptive control of discrete time Markov processes*, Stochastics Stochastic Reports 54 (1995), 301–316.
[17] B. Doshi and S. E. Shreve, *Randomized self-tuning control of Markov chains*, J. Appl. Probab. 17 (1980), 726–734.
[18] B. Hajek, *Hitting-time and occupation-time bounds implied by drift analysis with applications*, Adv. Appl. Probab. 14 (1982), 502–525.
[19] P. R. Kumar and A. Becker, *A new family of optimal adaptive controllers for Markov chains*, IEEE Trans. Automatic Control 27 (1982), 137–142.
[20] P. R. Kumar and W. Lin, *Optimal adaptive controllers for Markov chains*, ibid. 27 (1982), 756–774.
[21] P. R. Kumar and P. P. Varaiya, *Stochastic Systems—Estimation*, *Identification and Adaptive Control*, Prentice-Hall, 1986.
[22] P. Mandl, *Estimation and control in Markov chains*, Adv. Appl. Probab. 6 (1974), 40–60.
[23] R. Milito and J. B. Cruz, Jr., *An optimization oriented approach to adaptive control of Markov chains*, IEEE Trans. Automatic Control 32 (1987), 754–762.
[24] J. N. Tsitsiklis, *Asynchronous stochastic approaximation and Q-learning*, Machine Learning 16 (1994), 195–202.
[25] K. Van Hee, *Bayesian Control of Markov Chains*, Math. Center Tracts, 95, Math. Center, Amsterdam, 1978.

V. S. Borkar
Department of Computer Science and Automation
Indian Institute of Science
Bangalore 560012, India
E-mail: borkar@csa.iisc.ernet.in

S. M. Mundra
Associate, OA Division (SW Team)
Samsung Electronics Co. Ltd.
Suwon, P.O.B. 105, Kyungki-Do
South Korea 440600
E-mail: mundra@atom.info.samsung.co.kr

*Received on 20.8.1997;*
*revised version on 6.1.1998*