Contents lists available at ScienceDirect

Journal of Mathematical Analysis and Applications

www.elsevier.com/locate/jmaa

On the adaptive control of a class of partially observed Markov decision processes [☆]

Shun-Pin Hsu [a,*], Ari Arapostathis [b]

[a] Department of Electrical Engineering, National Chung Hsing University, 250, Kuo-Kuang Rd., Taichung 402, Taiwan
[b] Department of Electrical and Computer Engineering, The University of Texas at Austin, 1 University Station C0803, Austin, TX 78712-0240, USA

ARTICLE INFO

Article history:
Received 10 March 2009
Available online 15 March 2011
Submitted by Goong Chen

Keywords:
Adaptive control
Markov decision processes
Average-cost optimality

ABSTRACT

This paper is concerned with the adaptive control problem, over the infinite horizon, for partially observable Markov decision processes whose transition functions are parameterized by an unknown vector. We treat finite models and impose relatively mild assumptions on the transition function. Provided that a sequence of parameter estimates converging in probability to the true parameter value is available, we show that the certainty equivalence adaptive policy is optimal in the long-run average sense.

1. Introduction

Considerable effort has been invested in the study of stochastic adaptive control. Special attention has been paid to systems with incomplete or noisy state observation, and in particular, to discrete-time, partially observable Markov decision processes (POMDPs) with the transition probability matrix depending on some unknown parameter vector θ. A common approach is to decompose the problem into an adaptive estimation problem treated as in [1], and a control synthesis problem based on the parameter estimates provided by the estimation algorithm. The adaptive estimation algorithm is usually analyzed via the ordinary differential equation method, and convergence of the sequence of estimates to the true parameter, in probability, can be asserted under suitable conditions [2]. For an uncontrolled process, the aim of adaptive estimation algorithm is to successfully track the system state, or in other words to make the state estimates obtained under parameter uncertainty asymptotic to those obtained under the true parameter. This might fail, if feedback control is also introduced. What usually holds for the controlled process is convergence of the parameter estimates to the true parameter in probability. Thus, from the point of view of the control synthesis, the problem that remains is to synthesize a control policy based on the convergent sequence of parameter estimates, so as to achieve the control objectives. Studies along these lines, which establish the optimality of a certainty equivalence adaptive policy, under varying hypotheses, can be found in [3–6]. The key assumptions that are common in these and other studies are that (a) for each value of the unknown parameter there exists a stationary optimal policy, and (b) the information state under the true parameter and the estimated one become asymptotic to each other.

For the long-run average cost problem, the positivity of the transition matrices in [4] and a *renewability* property in [6] are enough guarantee (a). Existence of a stationary optimal policy, and also uniqueness of the solution to the Hamilton–Jacobi–Bellman (HJB) equation, has been established in [7] for POMDPs with finite state space and compact action space

[☆] This research was supported in part by the Office of Naval Research through the Electric Ship Research and Development Consortium, and in part by the National Science Council of Taiwan under grant NSC 97-2221-E-005-062.

[*] Corresponding author.
E-mail addresses: shsu@nchu.edu.tw (S.-P. Hsu), ari@ece.utexas.edu (A. Arapostathis).

doi:10.1016/j.jmaa.2011.03.004

under very mild assumptions. In this paper we employ one of the hypotheses in [7], and show that under parameter uncertainty, (b) holds for the process governed by a certainty equivalence policy. We then proceed to show that the certainty equivalence policy is optimal.

This paper is organized as follows. Section 2 reviews the basics of the POMDP model. Section 3 describes the assumptions imposed on the model and the adaptive policy. Section 4 summarizes some results concerning products of substochastic matrices, which are employed in Section 5 to derive the main results of the paper.

2. The POMDP model

The most commonly used model for a discrete-time, partially observable Markov decision process consists of the set of objects $(\mathcal{S}\times\mathcal{Y},\mathbb{U},Q,c)$, where: $\mathcal{S}=\{1,2,\dots,N\}$ is the *core state space*, $\mathcal{Y}$ is a finite *observation space*, $\mathbb{U}$ is a finite *action space*, Q is the transition matrix of the process, and $c\colon\mathcal{S}\times\mathbb{U}\to\mathbb{R}$ is the *running cost*. The system evolves as follows: if the system state at time t is X_t and a control action U_t is taken, a cost $c(X_t,U_t)$ is incurred, and the system moves to next state X_{t+1}, and observation Y_{t+1} is recorded, according to the transition matrix Q that may be interpreted as

$$Q(y,u)_{ij}=\mathrm{Prob}(X_{t+1}=j,\ Y_{t+1}=y\mid X_t=i,\ U_t=u).$$

Note that by definition the elements of Q are nonnegative and satisfy

$$\sum_{y\in\mathcal{Y},\,j\in\mathcal{S}}Q(y,u)_{ij}=1,\quad \forall i\in\mathcal{S},\ \forall u\in\mathbb{U}.$$

Only the process $\{Y_t\}$ is available for control synthesis—the core state $\{X_t\}$ is unavailable. It is well known that this model can be transformed into a completely observed one, which is equivalent to the original one as far as optimal control with a single stage running cost is concerned [8–10]. The state of the equivalent model at time t is chosen as the conditional distribution Ψ_t of X_t given the past observations $\{Y_1,Y_2,\dots,Y_t\}$. Thus, Ψ_t is a probability distribution on $\mathcal{S}$. In other words, Ψ_t takes values in Δ, the set of probability vectors of dimension N, defined by

$$\Delta\triangleq\biggl\{\psi=\bigl(\psi(1),\dots,\psi(N)\bigr)\in\mathbb{R}^N_+\colon\sum_{i\in\mathcal{S}}\psi(i)=1\biggr\}.$$

Here, $\mathbb{R}_+$ denotes the nonnegative real line. The evolution of the process Ψ_t can be written in a recursive form, via the well known filtering equation

$$\psi_{t+1}=\frac{\psi_t Q(y_{t+1},u_t)}{\|\psi_t Q(y_{t+1},u_t)\|}.\tag{1}$$

Eq. (1) is expressed along the paths of the process. Throughout this paper we use upper case letter for random variables, and lower case letters for the elements of the space they live in. For example $\{y_1,\dots,y_k\}$ is an element of $\mathcal{Y}^k$ and represents a particular sample path of length k of the process $\{Y_t\}$. We also fix a norm $\|\cdot\|$ on $\mathbb{R}^N$, defined by

$$\|z\|\triangleq\sum_{i=1}^N|z_i|,\quad z\in\mathbb{R}^N.$$

A restriction on the control U_t is that it has to be non-anticipative, or in other words the control variable U_t has to be measurable with respect to $\sigma(Y_1,\dots,Y_t)$, the 'sigma field' generated by the observations up to time t. In accordance with this requirement, an *admissible* action at time t may be chosen as a function that maps $\{y_1,\dots,y_t\}\in\mathcal{Y}^k$ to $\mathbb{U}$, and we define an *admissible* policy π as a sequence $\pi=(\pi_0,\pi_1,\dots)$ of admissible actions. The set of all admissible policies is denoted by Π. We refrain from introducing the concept of randomized actions, as it is not explicitly needed in this paper. Given a policy $\pi\in\Pi$ and an initial value ψ_0, there is a unique probability measure $\mathbb{P}^\pi_{\psi_0}$ on the path space of the process (Ψ_t,Y_t,U_t), which is generated via the transition kernel induced by the filter in (1) [8,9,11]. We let $\mathbb{E}^\pi_{\psi_0}$ denote the corresponding expectation operator. There is a particular class of policies that has special significance. If $\pi_t\colon\Delta\to\mathbb{U}$ is a measurable function, then the sequence $\{U_t=\pi_t(\Psi_t),\ t\geqslant 0\}$ constitutes a sequence of admissible actions. Moreover, under $\mathbb{P}^\pi_{\psi_0}$, $\{\Psi_t\}$ is a Markov chain. As a result, the corresponding policy π is called *Markov*. If each π_t is independent of t, i.e., $\pi_t=f$, where $f:\Delta\to\mathbb{U}$, the policy is called *stationary Markov*, and abusing the notation, we identify the policy π with the function f. The set of stationary Markov policies is denoted by Π_{SM}. If $\pi\in\Pi_{\mathrm{SM}}$, then $\{\Psi_t\}$ is a stationary Markov chain under $\mathbb{P}^\pi_{\psi_0}$.

2.1. Ergodic control

The objective of ergodic control, in its average formulation, is to synthesize a policy π that minimizes the incurred long-run average cost:

$$J(\psi_0,\pi)\triangleq\limsup_{T\to\infty}\frac{1}{T}\mathbb{E}^\pi_{\psi_0}\Biggl[\sum_{t=0}^{T-1}\tilde{c}(\Psi_t,U_t)\Biggr].$$

Here, $\tilde{c}$ is the transformed running cost under the equivalent completely observed model, and can be calculated from the original c by

$$\tilde{c}(\psi, u) = \sum_{i \in \mathcal{S}} c(i, u)\psi(i), \quad \psi \in \Delta,\ u \in \mathbb{U}.$$

Under the assumptions of the model considered in this paper, J does not depend on ψ_0, so we write this as $J(\pi)$. We set

$$\varrho^* \triangleq \inf_{\pi \in \Pi} J(\pi), \tag{2}$$

and we say that $\pi^* \in \Pi$ is *average-cost optimal* if $J(\pi^*) = \varrho^*$, and we refer to ϱ^* as the *optimal cost*.

3. Adaptive control

If the transition matrix Q depends on an unknown parameter $\theta^* \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a subset of a Euclidean space, the control problem is more complicated. A common approach is to construct a recursive algorithm that generates estimates $\{\hat{\Theta}_t,\ t \geqslant 0\}$ of the true parameter $\theta^* \in \boldsymbol{\Theta}$, and incorporate these estimates in the control synthesis. Stochastic approximation-type estimation algorithms have been widely used for this purpose. Naturally, $\hat{\Theta}_t$ should depend only on the past history of the observation process, or in other words be $\sigma(Y_1, \ldots, Y_t)$-measurable.

In general, estimation and control are interleaved, and the convergence of an estimation scheme may be conditional upon the control policy chosen. However, there are many cases where an estimation algorithm can be constructed that converges under any admissible policy, and while the performance of the estimation may depend on the policy utilized, qualitatively speaking, convergence is guaranteed. This is the scenario we address in this paper. We assume that there exists a recursive algorithm that generates estimates $\{\hat{\Theta}_t\}$ of the parameter, based on the past history of the observation process, such that $\hat{\Theta}_t \to \theta^*$ in probability, under any admissible control. Then, since the parameter estimation is non-anticipative, these estimates can be incorporated in the control design. We focus on a *certainty equivalence* policy, which we describe later in this section.

3.1. Assumptions on the model

We include the dependence of Q on θ explicitly, by denoting the transition matrix as Q_θ. Also ϱ^*_θ denotes the optimal cost, defined in (2), when the value of the parameter is θ.

Recall that a *row-allowable matrix* is a nonnegative matrix with all of its row sums positive. We make the following structural assumption on the model: $Q_\theta(y, u)$ is row-allowable for each $y \in \mathcal{Y}$, $u \in \mathbb{U}$, and $\theta \in \boldsymbol{\Theta}$. Since $\mathcal{Y}$ and $\mathbb{U}$ are finite and $\boldsymbol{\Theta}$ is compact, this implies that there exists $\varepsilon > 0$, such that

$$\sum_{j \in \mathcal{S}} \bigl[Q_\theta(y, u)\bigr]_{ij} \geqslant \varepsilon, \quad \text{for all } j \in \mathcal{S},\ y \in \mathcal{Y},\ u \in \mathbb{U}, \text{ and } \theta \in \boldsymbol{\Theta}.$$

We need to guarantee the existence of a stationary Markov optimal policy. To this end, we impose a hypothesis that is among the weakest in the literature. For a comparison of commonly used conditions to guarantee existence solutions to the HJB equation, we refer the reader to [7]. We state precisely the hypotheses on the model in Assumption 3.1 below. We need the following definition.

Definition 3.1. For a matrix A, let $A_{i\cdot}$ denote its ith row. For $\varepsilon > 0$, we let $\mathcal{M}(\varepsilon)$ denote the set of nonnegative, row-allowable $N \times N$ matrices A satisfying

$$\sum_{j \in \mathcal{S}} A_{ij} \geqslant \varepsilon,$$

and $\overline{\mathcal{M}}(\varepsilon)$ denote the subset of $\mathcal{M}(\varepsilon)$ satisfying

$$A_{i_1 j} \geqslant \varepsilon A_{i_2 j}, \quad \forall i_1, i_2, j \in \mathcal{S}. \tag{3}$$

Assumption 3.1. The parameter space $\boldsymbol{\Theta}$ is a compact subset of a Euclidean space, and the map $\theta \to Q_\theta$ is Lipschitz continuous. In addition we assume, that for some $\varepsilon > 0$,

(i) $Q_\theta(y, u) \in \mathcal{M}(\varepsilon)$, for all $y \in \mathcal{Y}$, $u \in \mathbb{U}$, and $\theta \in \boldsymbol{\Theta}$.

(ii) There exists a constant $\kappa \in \mathbb{N}$ such that for every pair of sequences $\{y_1, y_2, \ldots\}$ and $\{u_0, u_1, \ldots\}$,

$$Q^k_\theta\bigl(y^k, u^{k-1}\bigr) \in \overline{\mathcal{M}}(\varepsilon), \quad \text{for some } k \in \{1, 2, \ldots, \kappa\},$$

where $y^k \triangleq (y_1, \ldots, y_k) \in \mathcal{Y}^k$, $u^k \triangleq (u_0, \ldots, u_k) \in \mathbb{U}^k$, and Q^k_θ denotes the k-step transition matrix

$$Q^k_\theta\bigl(y^k, u^{k-1}\bigr) \triangleq Q_\theta(y_1, u_0) \cdots Q_\theta(y_k, u_{k-1}).$$

Remark 3.1. An example that satisfies Assumption 3.1 is in the following. Consider a transportation system with $\mathcal{S} = \mathcal{Y} = \{0, 1, 2\}$ and $\mathbb{U} = \{0, 1\}$. The transition matrix $Q_\theta(y, u) = P(u)O_\theta(y)$ where

$$P(0) = \begin{bmatrix} p_{00} & p_{01} & 1 - p_{00} - p_{01} \\ p_{10} & p_{11} & 1 - p_{10} - p_{11} \\ p_{20} & p_{21} & 1 - p_{20} - p_{21} \end{bmatrix}, \qquad P(1) = \begin{bmatrix} 1 - p_3 & p_3 & 0 \\ 1 - p_4 & p_4 & 0 \\ 1 - p_5 & p_5 & 0 \end{bmatrix} \tag{4}$$

and

$$O_\theta(0) = \begin{bmatrix} \theta & 0 & 0 \\ 0 & 1 - \theta & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad O_\theta(1) = \begin{bmatrix} 1 - \theta & 0 & 0 \\ 0 & \theta & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad O_\theta(2) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{5}$$

All the entries in $P(0)$ and $P(1)$ are non-zero except those in the third column of $P(1)$. The three states 0, 1, and 2 are interpreted as *normal*, *busy*, and *jammed* cases of the transportation system, respectively. Action 1 or 0 means to execute the traffic control or not, respectively. The only unknown parameter is θ. It lies in $[0.5, 1]$ and represents the probability of correct observation. Clearly the structure of the transition matrix satisfies the assumption.

In the theorem that follows, we let $e_0 \triangleq (1, 0, \dots, 0) \in \Delta$ (or any other fixed element of Δ).

Theorem 3.1. *Suppose Assumption* 3.1 *holds. Then, corresponding to each fixed $\theta \in \boldsymbol{\Theta}$, there exist a constant ϱ_θ, and a concave function $\mathcal{V}_\theta : \Delta \to \mathbb{R}$, with $\mathcal{V}_\theta(e_0) = 0$, such that the pair $(\varrho_\theta, \mathcal{V}_\theta)$ is a solution of the HJB equation*

$$\varrho_\theta + \mathcal{V}_\theta(\psi) = \min_{u \in \mathbb{U}} \bigl\{ \mathcal{L}_\theta(\psi, u) \bigr\}, \tag{6}$$

where

$$\mathcal{L}_\theta(\psi, u) \triangleq \tilde{c}(\psi, u) + \sum_{y \in \mathcal{Y}} \bigl\| \psi Q_\theta(y, u) \bigr\| \mathcal{V}_\theta \left(\frac{\psi Q_\theta(y, u)}{\| \psi Q_\theta(y, u \|} \right).$$

Moreover,

(i) *ϱ_θ is the optimal average cost, i.e., $\varrho_\theta = \varrho_\theta^*$, and $\pi \in \Pi_{\mathrm{SM}}$ is average-cost optimal if and only if it is a selector from the minimizer of* (6).
(ii) *If $(\hat{\varrho}, \hat{\mathcal{V}})$ is any solution of* (6)*, corresponding to $\theta \in \boldsymbol{\Theta}$, and satisfying $\hat{\mathcal{V}}(e_0) = 0$, then $\hat{\varrho} = \varrho_\theta$ and $\hat{\mathcal{V}} = \mathcal{V}_\theta$.*
(iii) *The maps $\theta \mapsto \varrho_\theta$, and $(\psi, \theta) \mapsto \mathcal{V}_\theta(\psi)$ are continuous.*
(iv) *In general, a policy $\pi \in \Pi$ (not necessarily stationary), is average-cost optimal if and only if it satisfies*

$$\lim_{t \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_{\psi_0}^{\pi} \bigl[\mathcal{D}_\theta(\Psi_t, U_t) \bigr] = 0,$$

where $\mathcal{D}_\theta : \Delta \times \mathbb{U} \to \mathbb{R}$ is the discrepancy function, *and is defined by*

$$\mathcal{D}_\theta(\psi, u) \triangleq \mathcal{L}_\theta(\psi, u) - \varrho_\theta - \mathcal{V}_\theta(\psi).$$

Proof. The assertions in (i) and (ii) follow by [7, Theorem 11], and (iv) follows from [10, Proposition 5.5.5]. For (iii) we argue as in [6]. Using the technique in [7], following the vanishing discount approach, one can show that $\psi \to \mathcal{V}_\theta(\psi)$ is equi-continuous, uniformly over $\theta \in \boldsymbol{\Theta}$. Therefore, if $\theta_n \to \theta$, as $n \to \infty$, then along some subsequence also denoted by $\{n\}$, $\mathcal{V}_{\theta_n} \to \hat{\mathcal{V}}$, uniformly on Δ, as $n \to \infty$, for some $\hat{\mathcal{V}} : \Delta \to \mathbb{R}$. Since ϱ_{θ_n} is bounded, the subsequence may be selected such that ϱ_{θ_n} converges to some constant $\hat{\varrho}$, as $n \to \infty$. Thus, $(\hat{\varrho}, \hat{\mathcal{V}})$ is a solution of (6), and satisfies $\hat{\mathcal{V}}(e_0) = 0$. By uniqueness, $\hat{\mathcal{V}} = \mathcal{V}_\theta$, and $\hat{\varrho} = \varrho_\theta$, thus establishing (iii). □

3.2. The adaptive policy

Suppose we have at our disposal a sequence of estimates $\{\hat{\Theta}_t,\ t \geqslant 0\}$ of the true parameter θ^* satisfying the following condition:

Assumption 3.2. The sequence $\{\hat{\Theta}_t\}_{t=0}^{\infty}$ satisfies

(i) $\hat{\Theta}_t$ is $\sigma(Y_1, \dots, Y_t)$-measurable, for each $t \geqslant 0$.
(ii) $\hat{\Theta}_t \to \theta^*$ in $\mathbb{P}_{\psi_0}^{\pi}$, as $t \to \infty$, for all $\pi \in \Pi$.

Remark 3.2. Consider the example in Remark 3.1. In practical situations, if the traffic flow is in the normal case with probability lower than some threshold, the traffic control will be executed. Otherwise, it will not. Note that the jammed traffic case is completely observable. Whenever it is observed, the traffic control will be executed and the next observation Y follows

$$P(Y=0)=(1-p_5)\theta+p_5(1-\theta)=1-P(Y=1). \tag{7}$$

Suppose $c(2,u)>c(i,u)$ for all $u\in\mathbb{U}$ and all $i=0,1$. Then $Y=0$ or 1 occurs infinitely often and from the nonlinear filter it is evident that $Y=1$ occurs infinitely often. The probability that $Y=1$ can be estimated by its occurrence frequency after each time $Y=2$ is observed. As time evolves we obtain a sequence of estimators that converges to the true value almost surely by the strong law of large numbers. However, in order to make every estimator fall in $\boldsymbol{\Theta}$, a projection scheme is required. By following the standard stochastic approximation-type recursive algorithm, the sequence of the estimators converges to the set of limit pints of some ordinary differential equation. For details, please see [6, p. 990] and the reference therein.

For $\theta\in\boldsymbol{\Theta}$, let $\pi^*(\cdot,\theta)$ be a measurable selector from the minimizer in (6). Note that this may be selected so that $(\psi,\theta)\mapsto\pi^*(\psi,\theta)$ is measurable. By Lemma 4.1 below, it follows that $Q_\theta^\kappa(y^\kappa,u^{\kappa-1})\in\overline{\mathcal{M}}(\varepsilon)$, for all $y^\kappa\in\mathcal{Y}^\kappa$ and $u^{\kappa-1}\in\mathbb{U}^\kappa$, where $\kappa\in\mathbb{N}$ is the constant in Assumption 3.1. We freeze the estimate of the parameter over consecutive blocks of time of length κ. In other words, if $\lfloor r\rfloor$ denotes the integer part of a real number r, we let

$$\check{\Theta}_t\triangleq\hat{\Theta}_{\lfloor t/\kappa\rfloor\kappa}.$$

It is evident that $\check{\Theta}_t$ is $\sigma(Y_1,\dots,Y_t)$-measurable, and since it is a subsequence of $\{\hat{\Theta}_t\}$, it also converges to θ^*. We use the sequence $\{\check{\Theta}_t\}$ to define an adaptive policy $\hat{\pi}=\{\hat{\pi}_t,\ t\geqslant 0\}$, along each sample path, recursively for $t=0,1,\dots,$ by

$$\begin{aligned}\hat{\pi}_t&\triangleq\pi^*(\hat{\psi}_t,\check{\theta}_t),\\ \hat{\psi}_{t+1}&\triangleq\frac{\hat{\psi}_t Q_{\check{\theta}_t}(y_{t+1},\pi^*(\hat{\psi}_t,\check{\theta}_t))}{\|\hat{\psi}_t Q_{\check{\theta}_t}(y_{t+1},\pi^*(\hat{\psi}_t,\check{\theta}_t))\|},\end{aligned} \tag{8}$$

with $\hat{\psi}_0\equiv\psi_0$. It is clear that $\hat{\pi}$ is admissible, i.e., $\hat{\pi}\in\Pi$. In this manner we also obtain a process $\{\hat{\Psi}_t,\ t\geqslant 0\}$. The process $\{\Psi_t,\ t\geqslant 0\}$ still depends, of course, on the true parameter θ^*, so it is governed by the filter

$$\psi_{t+1}=\frac{\psi_t Q_{\theta^*}(y_{t+1},\pi^*(\hat{\psi}_t,\check{\theta}_t))}{\|\psi_t Q_{\theta^*}(y_{t+1},\pi^*(\hat{\psi}_t,\check{\theta}_t))\|}.$$

The main goal of this paper is to show that $\hat{\pi}$ is optimal, i.e., $J(\hat{\pi})=\varrho^*_{\theta^*}$. In other words, $\hat{\pi}$ attains the optimal cost $\varrho^*_{\theta^*}$, corresponding to the true parameter θ^*.

Remark 3.3. Note that since $\check{\theta}$ is constant over time blocks of the form $\{t\colon\ \kappa(\ell-1)\leqslant t<\kappa\ell\}$, for $\ell\geqslant 1$, then, by Assumption 3.1,

$$Q_{\check{\theta}_{\kappa(\ell-1)}}(y_{\kappa(\ell-1)+1},\hat{\pi}_{\kappa(\ell-1)})\cdots Q_{\check{\theta}_{\kappa\ell-1}}(y_{\kappa\ell},\hat{\pi}_{\kappa\ell-1})\in\overline{\mathcal{M}}(\varepsilon),\quad\forall\ell\geqslant 1.$$

4. Products of nonnegative matrices

In this section, we summarize some properties of products of matrices belonging to the classes $\mathcal{M}(\varepsilon)$ and $\overline{\mathcal{M}}(\varepsilon)$. These relate to results on weak ergodicity of products of nonnegative matrices [12]. The key result is in Lemma 4.2 below, which is a variation of [13, Lemma 6.2].

Recall that the *oscillation* of a function $f:\mathcal{S}\to\mathbb{R}$, is defined by

$$\operatorname*{osc}_{s\in\mathcal{S}} f(s)\triangleq\max_{s\in\mathcal{S}} f(s)-\min_{s\in\mathcal{S}} f(s).$$

This is used in the next lemma.

Lemma 4.1. *If $A\in\overline{\mathcal{M}}(\varepsilon)$ and $B\in\mathcal{M}(\varepsilon')$, with $\varepsilon,\varepsilon'>0$, then $AB\in\overline{\mathcal{M}}(\varepsilon)$, and*

$$\frac{\|A_{i_1j}B_{j\cdot}\|}{\|(AB)_{i_1\cdot}\|}\geqslant\varepsilon^2\frac{\|A_{i_2j}B_{j\cdot}\|}{\|(AB)_{i_2\cdot}\|},\quad\textit{for all } i_1,i_2,j\in\mathcal{S}. \tag{9}$$

Moreover,

$$\operatorname*{osc}_{s\in\mathcal{S}}\frac{(AB)_{sj}}{\|(AB)_{s\cdot}\|}\leqslant\left(1-\varepsilon^2\right)\operatorname*{osc}_{s\in\mathcal{S}}\frac{B_{sj}}{\|B_{s\cdot}\|},\quad\forall j\in\mathcal{S}. \tag{10}$$

Proof. A direct computation yields, by employing (3),

$$(AB)_{i_1 j} = \sum_{k=1}^{N} A_{i_1 k} B_{kj} \geqslant \sum_{k=1}^{N} \varepsilon A_{i_2 k} B_{kj} = \varepsilon (AB)_{i_2 j}.$$

Hence $AB \in \overline{\mathcal{M}}(\varepsilon)$. Therefore, $\|(AB)_{i_2\cdot}\| \geqslant \varepsilon \|(AB)_{i_1\cdot}\|$, and (9) follows by comparing numerators and denominators.

To prove (10), define

$$\tilde{A}_{ij} \triangleq \frac{A_{ij} B_{j\cdot}}{\|(AB)_{i\cdot}\|}.$$

Then, by (9), and $|\cdot|^+$ denoting the positive part of a real number,

$$\sum_{s\in\mathcal{S}} |\tilde{A}_{i_1 s} - \tilde{A}_{i_2 s}|^+ \leqslant \sum_{s\in\mathcal{S}} \bigl(1-\varepsilon^2\bigr)\tilde{A}_{i_1 s} \leqslant \bigl(1-\varepsilon^2\bigr). \tag{11}$$

Also, since $\sum_j \tilde{A}_{ij} = 1$,

$$\sum_{s\in\mathcal{S}} |\tilde{A}_{i_1 s} - \tilde{A}_{i_2 s}|^+ = \sum_{s\in\mathcal{S}} |\tilde{A}_{i_2 s} - \tilde{A}_{i_1 s}|^+ = \frac{1}{2}\sum_{s\in\mathcal{S}} |\tilde{A}_{i_1 s} - \tilde{A}_{i_2 s}|. \tag{12}$$

By (11) and (12),

$$\begin{aligned}
\frac{(AB)_{i_1 j}}{\|(AB)_{i_1\cdot}\|} - \frac{(AB)_{i_2 j}}{\|(AB)_{i_2\cdot}\|} &= \sum_{s\in\mathcal{S}} \frac{B_{sj}}{\|B_{s\cdot}\|}\left(\frac{\|A_{i_1 s}B_{s\cdot}\|}{\|(AB)_{i_1\cdot}\|} - \frac{\|A_{i_2 s}B_{s\cdot}\|}{\|(AB)_{i_2\cdot}\|}\right)\\
&= \sum_{s\in\mathcal{S}} \frac{B_{sj}}{\|B_{s\cdot}\|}\bigl(|\tilde{A}_{i_1 s} - \tilde{A}_{i_2 s}|^+ - |\tilde{A}_{i_2 s} - \tilde{A}_{i_1 s}|^+\bigr)\\
&\leqslant \left(\operatorname*{osc}_{s\in\mathcal{S}} \frac{B_{sj}}{\|B_{s\cdot}\|}\right)\sum_{s\in\mathcal{S}} |\tilde{A}_{i_1 s} - \tilde{A}_{i_2 s}|^+\\
&\leqslant \bigl(1-\varepsilon^2\bigr)\operatorname*{osc}_{s\in\mathcal{S}} \frac{B_{sj}}{\|B_{s\cdot}\|},
\end{aligned}$$

thus establishing (10). □

Lemma 4.2. *Let $\{A_1,\dots,A_n\} \subset \overline{\mathcal{M}}(\varepsilon)$, and $M \triangleq A_1 A_2 \cdots A_n$. Then*

$$\left\|\frac{pM}{\|pM\|} - \frac{p'M}{\|p'M\|}\right\| \leqslant \frac{N}{\varepsilon}\bigl(1-\varepsilon^2\bigr)^{n-1}\|p - p'\|, \quad \forall p, p' \in \Delta. \tag{13}$$

Proof. Define $\tilde{p} = (\tilde{p}_1,\dots,\tilde{p}_N)$ by

$$\tilde{p}_i \triangleq \frac{p_i\|M_{i\cdot}\|}{\|pM\|}, \quad i \in \mathcal{S},$$

and similarly for $\tilde{p}'$. Expanding, and using Lemma 4.1,

$$\begin{aligned}
\left\|\frac{pM}{\|pM\|} - \frac{p'M}{\|p'M\|}\right\| &= \sum_{j\in\mathcal{S}}\left|\sum_{i\in\mathcal{S}}\left(\frac{p_i M_{ij}}{\|pM\|} - \frac{p'_i M_{ij}}{\|p'M\|}\right)\right|\\
&= \sum_{j\in\mathcal{S}}\left|\sum_{i\in\mathcal{S}} \frac{M_{ij}}{\|M_{i\cdot}\|}\left(\frac{p_i\|M_{i\cdot}\|}{\|pM\|} - \frac{p'_i\|M_{i\cdot}\|}{\|p'M\|}\right)\right|\\
&= \sum_{j\in\mathcal{S}}\left|\sum_{i\in\mathcal{S}} \frac{M_{ij}}{\|M_{i\cdot}\|}\bigl(\tilde{p}_i - \tilde{p}'_i\bigr)\right|\\
&\leqslant \frac{1}{2}\|\tilde{p} - \tilde{p}'\|\sum_{j\in\mathcal{S}}\left(\operatorname*{osc}_{i\in\mathcal{S}} \frac{M_{ij}}{\|M_{i\cdot}\|}\right)\\
&\leqslant \frac{N}{2}\bigl(1-\varepsilon^2\bigr)^{n-1}\|\tilde{p} - \tilde{p}'\|.
\end{aligned} \tag{14}$$

A straightforward calculation yields

$$\begin{aligned}\sum_{s\in\mathcal{S}}\left|\frac{p'_s\|M_{s\cdot}\|}{\|p'M\|}-\frac{p'_s\|M_{s\cdot}\|}{\|pM\|}\right| &\leqslant \sum_{s\in\mathcal{S}}\left|\frac{\|(p-p')M\|p'_s\|M_{s\cdot}\|}{\|pM\|\|p'M\|}\right| \\ &= \left|\frac{\|(p-p')M\|}{\|pM\|}\right| \\ &\leqslant \sum_{s\in\mathcal{S}}\left|\frac{p_s\|M_{s\cdot}\|}{\|pM\|}-\frac{p'_s\|M_{s\cdot}\|}{\|pM\|}\right|. \end{aligned} \tag{15}$$

Forming a triangle inequality and applying (15), we obtain

$$\|\tilde{p}-\tilde{p}'\| \leqslant 2\sum_{s\in\mathcal{S}}\frac{|p_s-p'_s|\|M_{s\cdot}\|}{\|pM\|} \leqslant 2\max_{j\in\mathcal{S}}\frac{\|M_{j\cdot}\|}{\|pM\|}\sum_{s\in\mathcal{S}}|p_s-p'_s|. \tag{16}$$

Since $\{A_1,\dots,A_n\}\subset\overline{\mathcal{M}}(\varepsilon)$, then by (9), $M\in\overline{\mathcal{M}}(\varepsilon)$, and it follows from (16) that

$$\|\tilde{p}-\tilde{p}'\| \leqslant \frac{2}{\varepsilon}\|p-p'\|. \tag{17}$$

Therefore, (13) follows by (14) and (17). □

5. Main results

In this section we show that the adaptive policy $\hat{\pi}$ defined in Section 3.2 is average-cost optimal.

Theorem 5.1. *Suppose Assumptions* 3.1 *and* 3.2 *hold. Then*

$$\mathbb{E}^{\hat{\pi}}_{\psi_0}\big[\|\Psi_t-\hat{\Psi}_t\|\big]\xrightarrow[t\to\infty]{}0,\quad \forall\psi_0\in\Delta. \tag{18}$$

Proof. In the interest of notational economy we define

$$\llbracket t\rrbracket \triangleq \lfloor t/\kappa\rfloor.$$

We evaluate along sample paths, and for this reason, in order to simplify the calculations we define

$$\hat{B}^t_\ell \triangleq \begin{cases} Q_{\check{\theta}_{\kappa(\ell-1)}}(y_{\kappa(\ell-1)+1},\hat{\pi}_{\kappa(\ell-1)})\cdots Q_{\check{\theta}_{\kappa\ell-1}}(y_{\kappa\ell},\hat{\pi}_{\kappa\ell-1}) & \text{if } \ell<\llbracket t\rrbracket, \\ Q_{\check{\theta}_{\kappa(\ell-1)}}(y_{\kappa(\ell-1)+1},\hat{\pi}_{\kappa(\ell-1)})\cdots Q_{\check{\theta}_{t-1}}(y_t,\hat{\pi}_{t-1}) & \text{for } \ell=\llbracket t\rrbracket. \end{cases}$$

Matrices $\{B^t_k\}$ are analogously defined, with $\check{\theta}$ replaced by the true parameter θ^*. By Assumption 3.1 and Remark 3.3, $\hat{B}^t_k, B^t_k\in\overline{\mathcal{M}}(\varepsilon)$, for all $k<\llbracket t\rrbracket$. Also, for $k=\llbracket t\rrbracket$, since $\hat{B}^t_k$ can be written as $\hat{B}^t_k=\hat{B}^{\kappa k}_k D$, with $D\in\mathcal{M}(\varepsilon)$, then by Lemma 4.1, $\hat{B}^t_k\in\overline{\mathcal{M}}(\varepsilon)$. Similarly $B^t_k\in\overline{\mathcal{M}}(\varepsilon)$, for $k=\llbracket t\rrbracket$. Let $\hat{p}_\ell\triangleq\hat{\psi}_{\kappa\ell}$. Note that

$$\hat{p}_\ell=\frac{\hat{p}_{\ell-1}\hat{B}^t_\ell}{\|\hat{p}_{\ell-1}\hat{B}^t_\ell\|},\quad \ell<\llbracket t\rrbracket.$$

For $\ell<\llbracket t\rrbracket$, we define

$$\bar{p}_\ell\triangleq\frac{\hat{p}_{\ell-1}B^t_\ell}{\|\hat{p}_{\ell-1}B^t_\ell\|},$$

and

$$\hat{D}^t_\ell\triangleq\hat{B}^t_\ell\hat{B}^t_{\ell+1}\cdots\hat{B}^t_{\llbracket t\rrbracket},\qquad D^t_\ell\triangleq B^t_\ell B^t_{\ell+1}\cdots B^t_{\llbracket t\rrbracket}.$$

Recall that $\hat{\psi}_0\equiv\psi_0$. We use the decomposition

$$\|\psi_t-\hat{\psi}_t\|=\sum_{\ell=1}^{\llbracket t\rrbracket-1}\left\|\frac{\bar{p}_\ell D^t_{\ell+1}}{\|\bar{p}_\ell D^t_{\ell+1}\|}-\frac{\hat{p}_\ell D^t_{\ell+1}}{\|\hat{p}_\ell D^t_{\ell+1}\|}\right\|+\left\|\frac{\hat{p}_{\llbracket t\rrbracket-1}D^t_{\llbracket t\rrbracket}}{\|\hat{p}_{\llbracket t\rrbracket-1}D^t_{\llbracket t\rrbracket}\|}-\frac{\hat{p}_{\llbracket t\rrbracket-1}\hat{D}^t_{\llbracket t\rrbracket}}{\|\hat{p}_{\llbracket t\rrbracket-1}\hat{D}^t_{\llbracket t\rrbracket}\|}\right\|. \tag{19}$$

By Lemma 4.2, for $1\leqslant\ell<\llbracket t\rrbracket$,

$$\left\|\frac{\bar{p}_\ell D^t_{\ell+1}}{\|\bar{p}_\ell D^t_{\ell+1}\|}-\frac{\hat{p}_\ell D^t_{\ell+1}}{\|\hat{p}_\ell D^t_{\ell+1}\|}\right\|\leqslant\frac{N}{\varepsilon}\big(1-\varepsilon^2\big)^{\llbracket t\rrbracket-\ell}\|\bar{p}_\ell-\hat{p}_\ell\|. \tag{20}$$

By the Lipschitz continuity of $\theta \mapsto Q_\theta(y,u)$, and since $\mathcal{Y}$ and $\mathbb{U}$ are finite, there exists a positive constant C_{Lip}, such that

$$\begin{aligned}
&\|\bar{p}_\ell - \hat{p}_\ell\| \leqslant C_{\mathrm{Lip}} \big\|\check{\theta}_{\kappa\ell-1} - \theta^*\big\|_{\boldsymbol{\Theta}}, \quad \ell < [\![t]\!], \\
&\left\| \frac{\hat{p}_{[\![t]\!]-1} D^t_{[\![t]\!]}}{\|\hat{p}_{[\![t]\!]-1} D^t_{[\![t]\!]}\|} - \frac{\hat{p}_{[\![t]\!]-1} \hat{D}^t_{[\![t]\!]}}{\|\hat{p}_{[\![t]\!]-1} \hat{D}^t_{[\![t]\!]}\|} \right\| \leqslant C_{\mathrm{Lip}} \sum_{i=0,1} \big\|\check{\theta}_{\kappa[\![t]\!]-i} - \theta^*\big\|_{\boldsymbol{\Theta}},
\end{aligned} \tag{21}$$

where $\|\cdot\|_{\boldsymbol{\Theta}}$ is the Euclidean norm in the parameter space. Replacing the terms on the right-hand side of (19) by their bounds in (20)–(21), we obtain

$$\begin{aligned}
\|\psi_t - \hat{\psi}_t\| &\leqslant \frac{NC_{\mathrm{Lip}}}{\varepsilon} \sum_{\ell=1}^{[\![t]\!]-1} \big(1-\varepsilon^2\big)^{[\![t]\!]-\ell} \big\|\check{\theta}_{\kappa\ell-1} - \theta^*\big\|_{\boldsymbol{\Theta}} + C_{\mathrm{Lip}} \sum_{i=0,1} \big\|\check{\theta}_{\kappa[\![t]\!]-i} - \theta^*\big\|_{\boldsymbol{\Theta}} \\
&\leqslant C_0 \sum_{\ell=1}^{[\![t]\!]+1} \big(1-\varepsilon^2\big)^{[\![t]\!]-\ell} \big\|\check{\theta}_{\kappa\ell-1} - \theta^*\big\|_{\boldsymbol{\Theta}},
\end{aligned} \tag{22}$$

where

$$C_0 \triangleq \frac{NC_{\mathrm{Lip}}}{\varepsilon}.$$

Since (22) holds for every sample path, then

$$\|\Psi_t - \hat{\Psi}_t\| \leqslant C_0 \sum_{\ell=1}^{[\![t]\!]+1} \big(1-\varepsilon^2\big)^{[\![t]\!]-\ell} \big\|\check{\Theta}_{\kappa\ell-1} - \theta^*\big\|_{\boldsymbol{\Theta}}. \tag{23}$$

Since $\check{\Theta}_t \to \theta^*$ in probability, under any $\pi \in \Pi$, as $t \to \infty$, and $\boldsymbol{\Theta}$ is compact,

$$\mathbb{E}^{\pi}_{\psi_0}\big[\big\|\check{\Theta}_{\kappa\ell-1} - \theta^*\big\|_{\boldsymbol{\Theta}}\big] \xrightarrow[\ell\to\infty]{} 0, \tag{24}$$

and therefore (18) follows by (23) and (24). □

Theorem 5.2. *Let Assumptions* 3.1 *and* 3.2 *hold. For a given unknown true parameter vector* $\theta^* \in \boldsymbol{\Theta}$, *the adaptive policy* $\hat{\pi}$ *described in* (8) *is average-cost optimal.*

Proof. By Theorem 3.1(iii), the map $(\psi,\theta) \mapsto \mathcal{D}_\theta(\psi,\cdot)$ is continuous. Moreover, as mentioned earlier, $\psi \mapsto \mathcal{D}_\theta(\psi,u)$ is Lipschitz continuous in ψ. Therefore, by Theorem 5.1, for each $u \in \mathbb{U}$,

$$\mathcal{D}_{\check{\Theta}_t}(\hat{\Psi}_t,u) - \mathcal{D}_{\theta^*}(\Psi_t,u) \xrightarrow[t\to\infty]{\mathbb{P}^{\hat{\pi}}_{\psi_0}} 0,$$

and since $\mathcal{D}$ is bounded, it converges also in $\mathbb{E}^{\hat{\pi}}_{\psi_0}$. Since $\mathbb{U}$ is finite, it follows that

$$\mathbb{E}^{\hat{\pi}}_{\psi_0}\big[\big|\mathcal{D}_{\check{\Theta}_t}\big(\hat{\Psi}_t,\pi^*(\hat{\Psi}_t,\check{\Theta}_t)\big) - \mathcal{D}_{\theta^*}\big(\Psi_t,\pi^*(\hat{\Psi}_t,\check{\Theta}_t)\big)\big|\big] \underset{t\to\infty]}{\longrightarrow} 0. \tag{25}$$

On the other hand, by definition, π^* satisfies $\mathcal{D}_\theta(\psi,\pi^*(\psi,\theta)) = 0$, for all $\psi \in \Delta$ and $\theta \in \boldsymbol{\Theta}$. Therefore,

$$\mathcal{D}_{\check{\Theta}_t}\big(\hat{\Psi}_t,\pi^*(\hat{\Psi}_t,\check{\Theta}_t)\big) = 0, \quad \mathbb{P}^{\hat{\pi}}_{\psi_0}\text{-a.s.}$$

In turn, by (25),

$$\mathbb{E}^{\hat{\pi}}_{\psi_0}\big[\big|\mathcal{D}_{\theta^*}\big(\hat{\Psi}_t,\pi^*(\hat{\Psi}_t,\check{\Theta}_t)\big)\big|\big] \xrightarrow[t\to\infty]{} 0,$$

and the result follows from Theorem 3.1(iv). □

6. Conclusion

A positive attribute of this paper is that there is essentially only one structural assumption (Assumption 3.1) imposed on the system model, and this takes a simple and verifiable form. Nevertheless, the reader will notice that existence of a solution to the HJB equation is guaranteed even under weaker hypotheses in [7]. It would be interesting to pursue the results in this paper under Assumption 4 in [7]. Also, extending the results to models with compact action spaces is desirable.

References

[1] A. Arapostathis, S.I. Marcus, Analysis of an identification algorithm arising in the adaptive estimation of Markov chains, Math. Control Signals Systems 3 (1) (1990) 1–29.

[2] H.J. Kushner, G.G. Yin, Stochastic Approximation Algorithms and Applications, Appl. Math. (N.Y.), vol. 35, Springer-Verlag, New York, 1997.

[3] G.B. Di Masi, Ł. Stettner, Bayesian adaptive control of discrete-time Markov processes with long-run average cost, Systems Control Lett. 34 (1–2) (1998) 55–62.

[4] T.E. Duncan, B. Pasik-Duncan, L. Stettner, Adaptive control of a partially observed discrete time Markov process, Appl. Math. Optim. 37 (3) (1998) 269–293.

[5] E. Fernández-Gaucherand, A. Arapostathis, S. Marcus, A methodology for the adaptive control of Markov chains under partial state information, in: Proceedings of the 31st IEEE Conference on Decision and Control, vol. 3, 1992, pp. 2750–2751.

[6] E. Fernández-Gaucherand, A. Arapostathis, S.I. Marcus, Analysis of an adaptive control scheme for a partially observed controlled Markov chain, IEEE Trans. Automat. Control 38 (6) (1993) 987–993.

[7] S.-P. Hsu, D.-M. Chuang, A. Arapostathis, On the existence of stationary optimal policies for partially observed MDPs under the long-run average cost criterion, Systems Control Lett. 55 (2) (2006) 165–173.

[8] A. Arapostathis, V.S. Borkar, E. Fernández-Gaucherand, M.K. Ghosh, S.I. Marcus, Discrete-time controlled Markov processes with average cost criterion: A survey, SIAM J. Control Optim. 31 (2) (1993) 282–344.

[9] E.B. Dynkin, A.A. Yushkevich, Controlled Markov Processes, Grundlehren Math. Wiss., vol. 235, Springer-Verlag, New York, 1979.

[10] O. Hernández-Lerma, J.B. Lasserre, Discrete-Time Markov Control Processes: Basic Optimality Criteria, Appl. Math. (N.Y.), vol. 30, Springer-Verlag, New York, 1996.

[11] D.P. Bertsekas, S.E. Shreve, Stochastic Optimal Control. The Discrete Time Case, Math. Sci. Eng., vol. 139, Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1978.

[12] E. Seneta, Nonnegative Matrices and Markov Chains, second ed., Springer Ser. Statist., Springer-Verlag, New York, 1981.

[13] T. Kaijser, A limit theorem for partially observed Markov chains, Ann. Probab. 3 (4) (1975) 677–696.