APPLICATIONES MATHEMATICAE
31,2 (2004), pp. 127–154

Rolando Cavazos-Cadena (Saltillo)
Raúl Montes-de-Oca (México)

ESTIMATION AND CONTROL IN FINITE MARKOV DECISION PROCESSES WITH THE AVERAGE REWARD CRITERION

Abstract. This work concerns Markov decision chains with finite state and action sets. The transition law satisfies the simultaneous Doeblin condition but is unknown to the controller, and the problem of determining an optimal adaptive policy with respect to the average reward criterion is addressed. A subset of policies is identified so that, when the system evolves under a policy in that class, the frequency estimators of the transition law are consistent on an essential set of admissible state-action pairs, and the non-stationary value iteration scheme is used to select an optimal adaptive policy within that family.

1. Introduction. This note is concerned with discrete-time Markov decision processes (MDPs) with finite state and control spaces. The performance of a control policy is measured by the (long-run expected) average reward criterion, and the main structural assumption is that the simultaneous Doeblin condition (SDC) is satisfied by the transition law (Thomas, 1980), but it is otherwise *unknown* to the controller. In this context, to drive the system in an optimal way, the decision maker must combine the control task with an estimation procedure, so that the actions applied are *adapted* to the available estimate at each decision time; *the main problem* considered below is to build an optimal adaptive policy.

The adaptive control problem studied in this note has recently been addressed, for instance, in Borkar (1996), Duncan *et al.* (1998), Drabik and

2000 *Mathematics Subject Classification*: Primary 90C40, 93E20; Secondary 60J05.

Key words and phrases: optimality equation, Schweitzer's transformation, convergence of non-stationary successive approximations, strong law of large numbers.

This work was supported by the PSF Organization under grant no. 010/300/02-2.

Stettner (2000), and Ren and Krogh (2001). In these papers the conditions imposed on the transition law are such that, under the action of each stationary policy, the *whole* state space is a communicating class. *The main difference* between the result in this work and those already available stems from the conditions imposed on the model, since the SDC used in this note is weaker than those used in the references above; particularly, the presence of transient states is possible under SDC.

Roughly, the analysis performed in the following sections to construct an optimal adaptive policy consists of two steps: First, a class of policies is identified so that, if the system is driven by a policy in that family, then the frequency (or empirical) estimators of the the transition law are consistent at an "essential" set of state-action pairs; the key tool in this part is the strong law of large numbers established in Loève (1977, p. 53), and the results in this direction extend those established in Cavazos-Cadena (1991) concerning completely *communicating* MDPs. Next, the non-stationary value iteration algorithm, extensively studied in Hernández-Lerma (1988), is used to obtain consistent estimates of a solution of the optimality equation on a subset of "essential" states; the analysis in this part avoids the restrictions on the rate of convergence of the estimators of the transition law imposed in Hernández-Lerma (1988, p. 61).

After these steps, the approximate solution of the optimality equation is used to select, at each time n, a stationary policy ϕ_n, and the optimal adaptive policy is constructed using a randomization mechanism under which the probability of applying ϕ_n converges to 1 as n increases. In contrast with the certainty-equivalence approach (Hernández-Lerma, 1988, p. 38, Ren and Krogh, 2001), the implementation of this adaptive policy requires neither *a priori* knowledge of an optimal stationary policy corresponding to each available estimate of the transition law, nor computing such a policy on-line.

The presentation has been organized as follows: In Section 2 the decision model is briefly described, and the sequence $\{P_n\}$ of frequency estimators of the transition law is introduced in Section 3. Next, the consistency of $\{P_n\}$ is studied in Section 4, and the results obtained in this direction are used to analyze the non-stationary value iteration scheme in Section 5. Finally, the optimal adaptive policy is constructed in Section 6.

Notation. Throughout the paper, $\mathbb{N}$ and $\mathbb{R}$ stand for the sets of nonnegative integer and real numbers, respectively. A finite set is always endowed with the discrete topology, and the cartesian product of topological spaces is endowed with the corresponding product topology. Let S and A be finite sets. Given a function $V\colon S \to \mathbb{R}$,

$$\|V\| := \max_{x \in S} |V(x)|$$

is the maximum norm of V, whereas $\mathbb{P}(S)$ denotes the set of all probability measures on S, i.e., $\mathbb{P}(S)$ consists of all functions $\mu\colon S \to [0,1]$ satisfying $\sum_{x\in S}\mu(x) = 1$. Furthermore, $\mathbb{P}(S\,|\,A)$ denotes the class of all stochastic kernels on S given A, that is, $\gamma(\cdot\,|\,\cdot) \in \mathbb{P}(S\,|\,A)$ if and only if $\gamma(\cdot\,|\,a) \in \mathbb{P}(S)$ for each $a \in A$; notice that $\mathbb{P}(S)$ and $\mathbb{P}(S\,|\,A)$ are naturally identified with (compact) subsets of finite-dimensional Euclidean spaces. The total variation distance between $\mu, \mu_1 \in \mathbb{P}(S)$ is defined by

$$\|\mu - \mu_1\| := \sum_{x\in S} |\mu(x) - \mu_1(x)|.$$

Notice the multiple meanings of $\|\cdot\|$; the context will make it clear which one is currently in use. Finally, the indicator function of an event W is denoted by $I[W]$ and, as usual, all the relations involving conditional expectations are supposed to hold almost everywhere with respect to the underlying probability measure.

2. Decision model.

2. Decision model. Let $M = \langle S, A, R, P\rangle$ be a Markov decision chain, where the finite sets S and A are the state and action spaces, respectively, $R\colon \mathbb{K} \to \mathbb{R}$ is the reward function, with the class $\mathbb{K}$ of admissible state-action pairs given by $\mathbb{K} := S \times A$, and $P = [P(y\,|\,x,\cdot)]$ is the controlled transition law. This model M is interpreted as follows: At each time $t \in \mathbb{N}$ the state of a dynamical system is observed, say $X_t = x \in S$, and an action $A_t = a \in A$ is chosen. Then a reward $R(x,a)$ is earned and, regardless of the previous states and actions, the state of the system at time $t+1$ will be $X_{t+1} = y \in S$ with probability $P(y\,|\,x,a)$; this is the Markov property of the decision model. Notice that it is assumed that every $a \in A$ is an admissible action at each state; as noted in Borkar (1984), this latter condition does not imply any loss of generality.

Policies. Given $t \in \mathbb{N}$, let $\mathbb{H}_t$ be the set of possible trajectories (histories) of the state-action process $\{(X_i, A_i)\}$ up to time t, where $\mathbb{H}_0 = S$, and $\mathbb{H}_t = \mathbb{K}\times\mathbb{H}_{t-1}$; a generic element of $\mathbb{H}_t$ is denoted by $\mathfrak{h}_t = (x_0, a_0, x_1, a_1, \ldots, x_{t-1}, a_{t-1}, x_t)$. A control policy $\pi = \{\pi_t(\cdot\,|\,\cdot)\}$ is a sequence of stochastic kernels, where $\pi_t \in \mathbb{P}(A\,|\,\mathbb{H}_t)$; if $t \in \mathbb{N}$ and $B \subset A$, then $\pi_t(B\,|\,\mathfrak{h}_t)$ is the probability of choosing $A_t \in B$ when the observed history is $\mathfrak{h}_t$. A policy π is *randomized stationary* if there exists $\gamma \in \mathbb{P}(A\,|\,S)$ such that the equality $\pi_t(\cdot\,|\,\mathfrak{h}_t) = \gamma(\cdot\,|\,x_t)$ always holds. In this case π and γ are naturally identified, and with this convention $\mathbb{P}(A\,|\,S) \subset \mathcal{P}$. Under the action of policy $\gamma \in \mathbb{P}(A\,|\,S)$ the state process $\{X_t\}$ is a Markov chain with transition probability matrix P^γ determined by

$$P^\gamma_{xy} = \sum_{a\in A} P(y\,|\,x,a)\gamma(a\,|\,x), \qquad x,y \in S,\ \gamma \in \mathbb{P}(A\,|\,S). \tag{2.1}$$

Set $\mathbb{F} := \prod_{x\in S} A$, so that $\mathbb{F}$ consists of all functions $f\colon S \to A$. A policy π is (*non-randomized*) *stationary* if there exists $f \in \mathbb{F}$ such that $\pi_t(\cdot \mid \mathfrak{h}_t)$ is always concentrated at $f(x_t)$, so that, under π, the action applied at time t is simply determined by $A_t = f(X_t)$. The class of stationary policies is identified with $\mathbb{F}$; thus, $\mathbb{F} \subset \mathcal{P}$, and policy $f \in \mathbb{F}$ corresponds to the stochastic kernel $\gamma_f \in \mathbb{P}(A \mid S)$ determined by $\gamma_f(f(x) \mid x) = 1$, $x \in S$. Given $\pi \in \mathcal{P}$ and $x \in S$, a unique probability measure is determined on the Borel σ-field of $\mathbb{H}_\infty := \mathbb{K}^\infty$; this is the distribution of the state-action process $\{(X_t, A_t)\}$ under the action of π when $X_0 = x$, and it is denoted by P_x^π, whereas E_x^π stands for the corresponding expectation operator. For details on this construction see, for instance, Hernández-Lerma (1988), Arapostathis *et al.* (1993), or Puterman (1994).

Performance index. Let $\pi \in \mathcal{P}$ and $x \in S$ be arbitrary. Under policy π the (long-run) expected average reward at state x is defined by

$$J(\pi, x) := \limsup_{n\to\infty} \frac{1}{n+1} E_x^\pi\Big[\sum_{t=0}^{n} R(X_t, A_t)\Big], \tag{2.2}$$

whereas

$$J^*(x) := \sup_{\pi\in\mathcal{P}} J(\pi, x) \tag{2.3}$$

is the optimal average reward at x; a policy π^* is optimal if $J(\pi^*, x) = J^*(x)$ for every $x \in S$. This work focuses on the case of a *constant* optimal value function, result that is ensured under the following *simultaneous Doeblin condition* (Thomas, 1980).

Assumption 2.1. There exist $z \in S$ such that $\max_{x\in S,\, f\in\mathbb{F}} E_x^f[T] < \infty$, where the first passage time T is given by $T := \min\{n > 0 \mid X_n = z\}$.

Lemma 2.1. (i) *Under Assumption* 2.1, *there exist $g \in \mathbb{R}$ and $h\colon S \to \mathbb{R}$ satisfying the average reward optimality equation* (*AROE*):

$$g + h(x) = \max_{a\in A}\Big[R(x, a) + \sum_{y\in S} P(y \mid x, a)h(y)\Big], \qquad x \in S. \tag{2.4}$$

(ii) $J^*(\cdot) = g$, *so that g is uniquely determined.*

(iii) *If $h(\cdot)$ is normalized to satisfy $h(z) = 0$, then $h(\cdot)$ is also unique.*

(iv) *If for each $x \in S$ the term within brackets in* (2.4) *is maximized at $a = f(x) \in A$, then the stationary policy f is optimal.*

A proof of this lemma can be found, for instance, in Hernández-Lerma (1988), Arapostathis *et al.* (1993) or Puterman (1994).

The problem. Throughout the remainder it is assumed that the transition law of the model satisfies Assumption 2.1, but is otherwise *unknown* to the controller and, in this sense, the working context is non-parametric, since no other special structure is assumed on P. Within this framework, to drive the system optimally, the decision maker must combine the control task with an estimation scheme to approximate the unknown transition law, and at each decision time $n \in \mathbb{N}$ the selection of the action to be applied can be visualized as a two-step procedure: First, at each $n \in \mathbb{N}$, the history of the process up to time n is used to build an estimation $P_n = [P_n(\cdot \,|\, \cdot, \cdot)]$ of P, and then the action A_n is (appropriately) chosen based on P_n; a policy combining the control task and an estimation procedure is referred to as an *adaptive* policy, so that *the main problem* of the decision maker is *to determine an optimal adaptive policy.* After analyzing the two steps mentioned above, an adaptive policy solving the controller's problem will be formulated.

The following consequence of Assumption 2.1 will be useful.

Lemma 2.2. *For each policy $\gamma \in \mathbb{P}(A \,|\, S)$, the Markov chain associated to γ has a unique invariant distribution $\mu_\gamma \in \mathbb{P}(S)$ which satisfies $\mu_\gamma(z) > 0$, where z is as in Assumption* 2.1.

Proof. Given $\gamma \in \mathbb{P}(A \,|\, S)$, for each $x \in S$ select $a_x \in A$ such that $\gamma(a_x \,|\, x) > 0$ and set $f(x) := a_x$, $x \in S$, whereas $c := \min_{x \in S} \gamma(a_x \,|\, x)$, so that $c > 0$. With this notation, (2.1) implies that $P^\gamma \geq cP^f$ and then $P_x^\gamma[X_n = z] \geq c^n P_x^f[X_n = z]$ for each state x and $n \in \mathbb{N}$. Next, for each $x \in S$, Assumption 2.1 implies that there exists $n_x \in \mathbb{N}$ satisfying $P_x^f[X_{n_x} = z] > 0$ and in this case $P_x^\gamma[X_{n_x} = z] > 0$. Thus, under the action of policy $\gamma \in \mathbb{P}(S \,|\, A)$, state z is accessible from every state $x \in S$, and then P^γ has a single recurrent class, which contains z, and consequently, a unique invariant distribution $\mu_\gamma \in \mathbb{P}(S)$ that necessarily satisfies $\mu_\gamma(z) > 0$. ■

3. Estimation of the transition law.

In this section the frequency estimators of the transition law are introduced. These estimators have previously been studied under the assumption that the whole state space is a communicating class under the action of each stationary policy; see, for instance, Borkar (1984), where they were used to establish the existence of optimal (non-adaptive) stationary policies for MDPs with denumerable state space and "norm-like" cost function, or Cavazos-Cadena (1991), where estimators P_n below were applied to build an *adaptive* policy which is asymptotically optimal with respect to the discounted criterion.

Definition 3.1. Let $\widetilde{\nu} \in \mathbb{P}(\mathbb{K})$ be arbitrary but fixed.

(i) The sequence $\{\nu_n\}$ of frequency distributions on $\mathbb{K}$ is defined by $\nu_0 = \widetilde{\nu}$ and

$$\nu_n(k) := \frac{1}{n} \sum_{t=0}^{n-1} I[(X_t, A_t) = k], \qquad k \in \mathbb{K},\ n = 1, 2, \ldots.$$

(ii) For each $n \in \mathbb{N}$, $\mu_n \in \mathbb{P}(S)$ is the marginal distribution of ν_n on S; in particular,

$$\mu_n(x) = \frac{1}{n}\sum_{t=0}^{n-1} I[X_t = x], \qquad x \in S,\ n = 1, 2, \ldots.$$

(iii) The conditional distribution $m_n \in \mathbb{P}(A \,|\, S)$ is given by

$$m_n(b \,|\, y) := \begin{cases} \dfrac{v_n(y,b)}{\mu_n(y)} & \text{if } \mu_n(y) \neq 0, \\ \widetilde{m}(b \,|\, y) & \text{otherwise,} \end{cases}$$

where $\widetilde{m} \in \mathbb{P}(A \,|\, S)$ is arbitrary but fixed.

(iv) Let $\widetilde{P} \in \mathbb{P}(S \,|\, \mathbb{K})$ be fixed. The sequence $\{P_n\} \subset \mathbb{P}(S \,|\, \mathbb{K})$ of empirical transition laws is defined by $P_0 = \widetilde{P}$, whereas for each $x \in S$, $k \in \mathbb{K}$ and $n = 1, 2, \ldots,$

$$(3.1) \qquad P_n(x \,|\, k) := \begin{cases} \dfrac{\sum_{t=1}^{n} I[X_t = x, (X_{t-1}, A_{t-1}) = k]}{n\nu_n(k)} & \text{if } \nu_n(k) > 0, \\ \widetilde{P}(x \,|\, k) & \text{otherwise.} \end{cases}$$

Remark 3.1. (i) Throughout the remainder the information vector up to time t is denoted by $I_t = (X_0, A_0, \ldots, X_{t-1}, A_{t-1}, X_t)$ for $t \geq 1$ and $I_0 = X_0$. The σ-fields $\mathcal{F}_t$ are determined by

$$\mathcal{F}_t = \sigma(I_{t-1}, A_{t-1}), \qquad t \geq 1,$$

whereas $\mathcal{F}_0$ is the minimum σ-algebra on $\mathbb{H}_\infty$. Each field $\mathcal{F}_t$ represents all the information available *before* observing state X_t, whereas $\sigma(I_t)$ contains the information available immediately after X_t is observed, and $\mathcal{F}_t \subset \sigma(I_t) \subset \mathcal{F}_{t+1}$ for every $t \in \mathbb{N}$.

(ii) Notice that for each $x, y \in S$ and $a \in A$, $P_n(y \,|\, x, a)$ is a random variable depending on I_n so that, formally, it should be denoted by $P_n(y \,|\, x, a; I_n)$. However, to ease the notational burden, the expression in Definition 3.1 will be used consistently. Similarly, $\nu_n(k)$, $\mu_n(x)$ and $m_n(b \,|\, y)$ always depend on I_{n-1} and A_{n-1}, so that they all are $\mathcal{F}_n$-measurable random variables.

From Definition 3.1 it follows that ν_n can be factored as

$$(3.2) \qquad \nu_n(y,b) = \mu_n(y) m_n(b \,|\, y), \qquad (y,b) \in \mathbb{K},\ n \in \mathbb{N};$$

indeed, when $\mu_n(y) \neq 0$, the equality follows from the definition of the conditional distribution m_n, whereas if $\mu_n(y) = 0$, Definition 3.1(ii) implies that $\nu_n(y,b) = 0$, so that both sides of (3.2) are zero. The analysis of the sequence $\{P_n\}$ relies on the following version of the strong law of large numbers, whose proof can be found in Loève (1977, p. 53).

Lemma 3.1. *Let $\{Y_n\}$ be a sequence of random variables on a probability space $(\Omega, \mathcal{G}, \mathbf{P})$, and suppose that conditions* (a) *and* (b) *below hold*:

(a) *For some finite constant C, $\mathbf{P}[|Y_n| \le C] = 1$ for each $n \in \mathbb{N}$.*

(b) *$\mathcal{G}_k \subset \mathcal{G}_{k+1} \subset \mathcal{G}$, $k \in \mathbb{N}$, and each Y_n is $\mathcal{G}_{n+1}$-measurable.*

In this case, if $\{b_n\}$ is an increasing sequence of positive numbers satisfying $b_n \nearrow \infty$ and $\sum_n 1/b_n^2 < \infty$, then

$$\lim_{n\to\infty} \frac{1}{b_n} \sum_{t=0}^{n} (Y_n - E[Y_n \,|\, \mathcal{G}_n]) = 0$$

with probability 1.

Remark 3.2. Throughout the subsequent development N denotes the number of states, whereas $\{d_n\}$ and $\{c_n\}$ are two sequences of positive numbers satisfying the following properties (a)–(d):

(a) $\{d_n\} \subset (0, 1]$, and $d_n \searrow 0$.

(b) $n/c_n \ge 1$ for each positive integer n.

(c) For each $j = 0, 1, \ldots, N$, $d_n^j c_n \nearrow \infty$ as $n \nearrow \infty$.

(d) $\sum_{n=0}^{\infty} 1/(d_n^N c_n)^2 < \infty$.

These conditions are satisfied, for instance, if we set $c_0 = d_0 = 1$, $c_n = n$ and $d_n = n^{-1/(2N+1)}$ for $n \ge 1$. Notice that (a) and (d) together imply that

(e) $\sum_{n=0}^{\infty} 1/(d_n^j c_n)^2 \le \sum_{n=1}^{\infty} 1/(d_n^N c_n)^2 < \infty$, $j = 0, 1, \ldots, N$.

The key tool to analyze the consistency of the sequence $\{P_n\}$ of frequency estimators is the following consequence of Lemma 3.1; it is an extension of Theorem 5.1 in Cavazos-Cadena (1991), which was derived under conditions stronger than Assumption 2.1.

Lemma 3.2. *For each $w \in S$ and $\pi \in \mathcal{P}$ the following assertions* (i) *and* (ii) *hold P_w^π-a.s.*:

(i) $\liminf_{n\to\infty} \mu_n(z) > 0$, *where $z \in S$ is as in Assumption* 2.1.

(ii) *For each $x \in S$ and $k \in \mathbb{K}$,*

$$\lim_{n\to\infty} \frac{n\nu_n(k)}{d_n^N c_n} [P_n(x \,|\, k) - P(x \,|\, k)] = 0.$$

Proof. (i) Let the σ-fields $\mathcal{F}_t$ be as in Remark 3.1 and let $x \in S$ be arbitrary but fixed. From Definition 3.1(ii) it follows that for each integer $n > 1$,

$$(1+n)\mu_{n+1}(x) = \sum_{t=0}^{n} I[X_t = x] = W_n(x) + \sum_{t=1}^{n} E_w^\pi[I[X_t = x] \,|\, \mathcal{F}_t], \tag{3.3}$$

where the random variable $W_n(x)$ is given by

$$W_n(x) = I[X_0 = x] + \sum_{t=1}^{n}(I[X_t = x] - E_w^\pi[I[X_t = x] \mid \mathcal{F}_t]). \tag{3.4}$$

Observe now that the Markov property implies that $E_w^\pi[I[X_t = x] \mid \mathcal{F}_t] = P(x \mid X_{t-1}, A_{t-1})$ for each $t \geq 1$, so that

$$\sum_{t=1}^{n} E_w^\pi[I[X_t = x] \mid \mathcal{F}_t] = \sum_{t=1}^{n} P(x \mid X_{t-1}, A_{t-1}) = \sum_{t=0}^{n-1} P(x \mid X_t, A_t);$$

from this, Definition 3.1 and (3.2) yield

$$\begin{aligned}\sum_{t=1}^{n} E_w^\pi[I[X_t = x] \mid \mathcal{F}_t] &= n \sum_{(y,b)\in\mathbb{K}} P(x \mid y, b)\nu_n(y, b)\\ &= n \sum_{y\in S} \Big\{ \sum_{b\in A} P(x \mid y, b) m_n(b \mid y) \Big\} \mu_n(y)\\ &= n \sum_{y} \mu_n(y) P_{yx}^{m_n},\end{aligned}$$

where formula (2.1) was used to obtain the last equality. Together with (3.3) this implies

$$(1 + n)\mu_{n+1}(x) = W_n(x) + n \sum_{y} \mu_n(y) P_{yx}^{m_n}.$$

Now let Ω_x be the subset of trajectories for which $\lim_{n\to\infty} W_n(x)/n = 0$. Applying Lemma 3.1 with n, $I[X_n = x]$ and $\mathcal{F}_n$ instead of b_n, Y_n and $\mathcal{G}_n$, respectively, we deduce that $P_w^\pi[\Omega_x] = 1$. Therefore $\widetilde{\Omega} = \bigcap_{x\in S} \Omega_x$ satisfies $P_w^\pi[\widetilde{\Omega}] = 1$, since S is finite, and the above displayed equation yields

$$\text{On the event } \widetilde{\Omega},\ \lim_{n\to\infty} \Big[\mu_{n+1}(x) - \sum_{y} \mu_n(y) P_{yx}^{m_n}\Big] = 0,\ x \in S. \tag{3.5}$$

Next, *consider a fixed trajectory in* $\widetilde{\Omega}$, and let ℓ be the lower limit of the sequence $\{\mu_n(z)\}$ along such a trajectory. In this case there exists a sequence $\{n_k\}$ such that $\lim_{k\to\infty} \mu_{n_k}(z) = \ell$; moreover, since S and A are finite sets, taking a subsequence if necessary, it can be assumed that for some $\mu \in \mathbb{P}(S)$ and $m \in \mathbb{P}(A \mid S)$ the following convergences hold:

$$\lim_{k\to\infty} \mu_{n_k}(y) = \mu(y), \qquad \lim_{k\to\infty} m_{n_k}(b \mid y) = m(b \mid y), \quad y \in S,\ b \in A.$$

Since $\mu_{n_k+1}(x) = [n_k/(n_k + 1)]\mu_{n_k}(x) + I[X_{n_k} = x]/(n_k + 1)$, it follows that $\lim_{k\to\infty} \mu_{n_k+1}(\cdot) = \mu(\cdot)$, whereas (2.1) shows that $\lim_{k\to\infty} P_{yx}^{m_{n_k}} = P_{yx}^m$ for all $x, y \in S$. Therefore, the convergence in (3.5) implies that $\mu(x) = \sum_y \mu(y) P_{yx}^m$ for every $x \in S$, so that $\mu(\cdot)$ is the invariant distribution of matrix P^m, and then $\ell = \mu(z) > 0$, by Lemma 2.2. In short, it has been shown that $\liminf_{n\to\infty}\mu_n(z) > 0$ along each trajectory in $\widetilde{\Omega}$, and the conclusion follows since, as already noted, $P_w^\pi[\widetilde{\Omega}] = 1$.

(ii) Let $x \in S$, $k \in \mathbb{K}$ and the positive integer n be fixed, and notice that

$$n\nu_n(k)P_n(x\,|\,k) = \sum_{t=1}^{n} I[X_t = x, (X_{t-1}, A_{t-1}) = k].$$

Indeed, the right hand side of this equality is bounded above by $n\nu_n(k)$, by Definition 3.1(i), so that both sides of the above equation are null if $\nu_n(k) = 0$, whereas the equality follows from Definition 3.1(iv) if $\nu_n(k) > 0$. Thus,

$$n\nu_n(k)P_n(x\,|\,k) = \widetilde{W}_n + \sum_{t=1}^{n} E_w^\pi[I[X_t = x, (X_{t-1}, A_{t-1}) = k]\,|\,\mathcal{F}_t], \tag{3.6}$$

where the random variable $\widetilde{W}_n$ is given by

$$\begin{aligned}\widetilde{W}_n = \sum_{t=1}^{n}(I[X_t = x, (X_{t-1}, A_{t-1}) = k]&\\ - E_w^\pi[I[X_t = x, (X_{t-1}, A_{t-1}) = k]\,|\,\mathcal{F}_t])&\end{aligned} \tag{3.7}$$

Since the event $[(X_{t-1}, A_{t-1}) = k]$ belongs to $\mathcal{F}_t$, the Markov property implies that $E_w^\pi[I[X_t = x, (X_{t-1}, A_{t-1}) = k]\,|\,\mathcal{F}_t] = P(x\,|\,k)I[(X_{t-1}, A_{t-1}) = k]$, and then

$$\begin{aligned}\sum_{t=1}^{n} E_w^\pi[I[X_t = x, (X_{t-1}, A_{t-1}) = k]\,|\,\mathcal{F}_t] &= \sum_{t=1}^{n} P(x\,|\,k)I[(X_{t-1}, A_{t-1}) = k]\\ &= P(x\,|\,k)\sum_{t=0}^{n-1} I[(X_t, A_t) = k]\\ &= n\nu_n(k)P(x\,|\,k),\end{aligned}$$

by Definition 3.1(i). Together with (3.6), this implies that $n\nu_n(k)[P_n(x\,|\,k) - P(x\,|\,k)] = \widetilde{W}_n$, so that

$$\lim_{n\to\infty}\frac{n\nu_n(k)}{d_n^N c_n}\,[P_n(x\,|\,k) - P(x\,|\,k)] = \lim_{n\to\infty}\frac{\widetilde{W}_n}{d_n^N c_n} = 0 \qquad P_w^\pi\text{-a.s.},$$

where the second equality follows from Lemma 3.1 applied, with $b_t = d_t^N c_t$ and $\mathcal{G}_t = \mathcal{F}_t$, to the variables $Y_t = I[X_t = x, (X_{t-1}, A_{t-1}) = k]$; see (3.7) and Remarks 3.1 and 3.2. ∎

4. Consistency. This section analyzes the consistency of the estimators of the transition law introduced in Definition 3.1. The result in this direction is stated after introducing subsets of the state space and the set of admissible state-action pairs, as well as a subfamily of the class $\mathcal{P}$ of all policies.

Definition 4.1. (i) For each $n \in \mathbb{N}$, the set $S_n \subset S$ is recursively defined as follows: $S_0 = \{z\}$, where z is the fixed state in Assumption 2.1, and for $n \geq 1$,

$$S_n = S_{n-1} \cup \{y \in S \mid P(y \mid x, a) > 0 \text{ for some } (x, a) \in S_{n-1} \times A\}.$$

(ii) The set S^* of essential states is defined by $S^* = \bigcup_{n=0}^{\infty} S_n$, whereas the class $\mathbb{K}^*$ of essential state-action pairs is given by $\mathbb{K}^* = S^* \times A$.

(iii) If $M = [M_{xy}]$ is a matrix whose components are indexed by the elements of S, define

$$\|M\|_* = \max_{x \in S^*} \sum_{y \in S^*} |M_{xy}|. \tag{4.1}$$

The properties of the sets S^* and $\mathbb{K}^*$ stated in the next lemma will be useful.

Lemma 4.1. (i) *The set S^* is closed, i.e., $P(S^* \mid k) = 1$ for each $k \in \mathbb{K}^*$.*

(ii) *$S^* = \bigcup_{n=0}^{N-1} S_n$; recall that N is the number of states.*

(iii) *Let $w \in S$ and $\pi \in \mathcal{P}$ be arbitrary, and for each $k \in \mathbb{K}^*$ define the hitting time T_k by*

$$T_k = \min\{n \geq 0 \mid (X_n, A_n) = k\}. \tag{4.2}$$

With this notation, for each $k \in \mathbb{K}^*$,

$$P_n(S^* \mid k) = 1 \quad P_w^\pi\text{-a.s. on the event } [T_k < n],$$

i.e., $P_w^\pi[T_k < n] = P_w^\pi[[T_k < n] \cap [P_n(S^* \mid k) = 1]]$.

(iv) *$P_w^\pi[X_n \notin S^*] \leq P_w^\pi[T > n] \to 0$ as $n \to \infty$, where T is the hitting time in Assumption 2.1.*

Proof. (i) Let $(x, a) \in S^* \times A = \mathbb{K}^*$ be fixed. In this case there exists i such that $x \in S_i$, and then $P(y \mid x, a) > 0$ implies that $y \in S_{i+1} \subset S^*$.

(ii) Notice that $\{z\} = S_0 \subset S_1 \subset \cdots \subset S_{N-1} \subset S_N$. Since the state space S has N elements, at least one of these inclusions is not strict, so that $S_k = S_{k+1}$ for some $k < N$. From this, an induction argument using Definition 4.1 shows that $S_k = S_{k+r}$ for every $r \in \mathbb{N}$. Therefore, $S^* = \bigcup_{t=0}^{k} S_t = \bigcup_{t=0}^{N-1} S_t$, since $k < N$.

(iii) Let $k \in \mathbb{K}^*$ be arbitrary but fixed. From Definition 3.1 and (4.2), it follows that $[T_k < n] = [\nu_n(k) > 0]$ and this implies, via (3.1), that for each $y \in S$,

$$\begin{aligned}[T_k < n] \cap [P_n(y \mid k) > 0] &= [\nu_n(k) > 0] \cap [P_n(y \mid k) > 0] \\ &\subset \bigcup_{t=0}^{n-1} [X_t = y, (X_{t-1}, A_{t-1}) = k].\end{aligned}$$

Suppose now that $y \in S \setminus S^* = S^{*\mathrm{c}}$. Since $I[(X_{t-1}, A_{t-1}) = k]$ is $\mathcal{F}_t$-measurable, the Markov property yields $P_w^\pi[X_t = y, (X_{t-1}, A_{t-1}) = k \,|\, \mathcal{F}_t] = P(y \,|\, k) I[(X_{t-1}, A_{t-1}) = k] = 0$, where part (i) was used to establish the last equality. Therefore, $P_w^\pi[X_t = y, (X_{t-1}, A_{t-1}) = k] = 0$, and the above displayed inclusion leads to $P_w^\pi[[T_k < n] \cap [P_n(y \,|\, k) > 0]] = 0$ for each $y \in S^{*\mathrm{c}}$. Consequently,

$$\begin{aligned} P_w^\pi[[T_k < n] \cap [P_n(S^{*\mathrm{c}} \,|\, k) > 0]] &= P_w^\pi\Big[[T_k < n] \cap \bigcup_{y \in S^{*\mathrm{c}}} [P_n(y \,|\, k) > 0]\Big] \\ &\le \sum_{y \in S^{*\mathrm{c}}} P_w^\pi[[T_k < n] \cap [P_n(y \,|\, k) > 0]] = 0, \end{aligned}$$

and then $P_w^\pi[T_k < n] = P_w^\pi[[T_k < n] \cap [P_n(S^{*\mathrm{c}} \,|\, k) = 0]] = P_w^\pi[[T_k < n] \cap [P_n(S^* \,|\, k) = 1]]$.

(iv) Since $z \in S^*$, using part (i) it is not difficult to see that, for each $n \in \mathbb{N}$, $P_z^\pi[X_n \notin S^*] = 0$. Observe now that $[T = r] \in \sigma(I_r)$, so that for each $n \ge r$, the Markov property implies that

$$P_w^\pi[T = r, X_n \notin S^* \,|\, I_r] = I[T = r] P_z^{\pi'}[X_{n-r} \notin S^*] = 0,$$

where the shifted policy π' is determined by $\pi_0'(\cdot \,|\, x) = \pi_r(\cdot \,|\, x)$ and $\pi_t'(\cdot \,|\, \mathfrak{h}_t) = \pi_{t+r}(\cdot \,|\, Ir, \mathfrak{h}_t)$ for $t > 0$. Therefore, $P_w^\pi[T = r, X_n \notin S^*] = 0$ when $r \le n$, and thus $P_w^\pi[X_n \notin S^*] = P_w^\pi[X_n \notin S^*, T > n] \le P_w^\pi[T > n] \to 0$ as $n \to \infty$, where the convergence follows from Assumption 2.1 via Markov's inequality. ■

The following class of policies was used in Cavazos-Cadena (1991) to study communicating MDPs with the discounted criterion.

Definition 4.2. Let $\varrho(\cdot)$ be a fixed probability distribution on A satisfying $\varrho(a) > 0$ for all $a \in A$. The family $\mathcal{P}^* \subset \mathcal{P}$ consists of all policies π satisfying

$$\pi_t(\cdot \,|\, \mathfrak{h}_t) \ge d_t \varrho(\cdot), \qquad t \in \mathbb{N},\ \mathfrak{h}_t \in \mathbb{H}_t.$$

Let I_t be the information vector up to time t introduced in Remark 3.1, and notice that the equality $P_x^\pi[A_t = a \,|\, I_t] = \pi_t(a \,|\, I_t)$ always holds, so that

$$E_x^\pi[I[A_t = a] \,|\, I_t] = P_x^\pi[A_t = a \,|\, I_t] \ge d_t \varrho(a), \qquad \pi \in \mathcal{P}^*,\ (x, a) \in \mathbb{K},\ t \in \mathbb{N}. \tag{4.3}$$

The main result of this section can now be stated.

Theorem 4.1. *Let $\{P_n\}$ be the sequence of frequency estimators of the transition law in Definition* 3.1. *For each $\pi \in \mathcal{P}^*$, $w \in S$ and $k \in \mathbb{K}^*$,*

$$P_w^\pi[\lim_{n\to\infty} \|P_n(\cdot \,|\, k) - P(\cdot \,|\, k)\| = 0] = 1.$$

The proof of this theorem is based on the following auxiliary result.

Lemma 4.2. *Let $\pi \in \mathcal{P}^*$ and $w \in S$ be arbitrary.*

(i) *For any $i = 0, 1 \ldots, N-1$ and $k \in S_i \times A$,*

$$\liminf_{n\to\infty} \frac{n\nu_n(k)}{d_n^{i+1} c_n} > 0 \quad P_w^\pi\text{-}a.s. \tag{4.4}$$

(ii) *For each $k \in \mathbb{K}^*$, $P_w^\pi[T_k < \infty] = 1$; see* (4.2).

Proof. (i) [By induction on i.] Let $\pi \in \mathcal{P}^*$ and $w \in S$ be fixed. To establish the case $i = 0$ of (4.4), let $(z, b) \in S_0 \times A = \{z\} \times A$ be fixed and observe that for each positive integer n,

$$n\nu_n(z,b) = \sum_{t=0}^{n-1} I[X_t = z, A_t = b] = \widetilde{W}_n + \sum_{t=0}^{n-1} E_w^\pi[I[X_t = z, A_t = b] \mid I_t],$$

where

$$\widetilde{W}_n := \sum_{t=0}^{n-1} (I[X_t = z, A_t = b] - E_w^\pi[I[X_t = z, A_t = b] \mid I_t]). \tag{4.5}$$

Since $I[X_t = z]$ is $\sigma(I_t)$-measurable, (4.3) yields

$$E_w^\pi[I[X_t = z, A_t = b] \mid I_t] = I[X_t = z] E_w^\pi[I[A_t = b] \mid I_t] \geq I[X_t = z] d_t \varrho(a)$$

and thus

$$\begin{aligned} n\nu_n(z,b) &\geq \widetilde{W}_n + \sum_{t=0}^{n-1} I[X_t = z] d_t \varrho(a) \\ &\geq \widetilde{W}_n + d_n \varrho(a) \sum_{t=0}^{n-1} I[X_t = z] = \widetilde{W}_n + n d_n \varrho(a) \mu_n(z), \end{aligned}$$

where the second inequality used the fact that $\{d_t\}$ is a decreasing sequence, and the equality is due to Definition 3.1. Therefore,

$$\frac{n\nu_n(z,b)}{d_n c_n} \geq \frac{\widetilde{W}_n}{d_n c_n} + \frac{n}{c_n} \varrho(a) \mu_n(z).$$

Set $Y_t = I[X_t = z, A_t = a]$ and $b_t = d_t c_t$; then parts (c) and (e) of Remark 3.2 and (4.5) together allow one to apply Lemma 3.1 with $\mathcal{G}_t = \sigma(I_t)$ to obtain

$$\frac{\widetilde{W}_n}{d_n c_n} \to 0 \quad P_w^\pi\text{-a.s.}$$

Also, since $n/c_n \geq 1$ and $\varrho(a) > 0$ (see Remark 3.2 and Definition 4.2), Lemma 3.2(i) leads to

$$\liminf_{n\to\infty} \frac{n}{c_n} \varrho(a) \mu_n(z) > 0 \quad P_w^\pi\text{-a.s.},$$

Combining the last three displayed relations yields

$$\liminf_{n\to\infty} \frac{n\nu_n(z,b)}{d_n c_n} > 0 \qquad P_w^\pi\text{-a.s.},$$

establishing the case $i = 0$ of (4.4).

Assume now that (4.4) holds for some non-negative integer $i = j - 1 < N - 1$ and let $(y, b) \in S_j \times A$ be arbitrary. In this case, by Definition 4.1(i), there exists $(x, a) \in S_{j-1} \times A$ such that

$$P(y \mid x, a) > 0. \tag{4.6}$$

Next, since the event $[X_t = y, (X_{t-1}, A_{t-1}) = (x, a)]$ belongs to $\sigma(I_t)$ for each $t > 0$,

$$\begin{aligned} E_w^\pi[I[(X_t, A_t) = (y, b), (X_{t-1}, A_{t-1}) = (x, a)] \mid I_t] \\ &= I[X_t = y, (X_{t-1}, A_{t-1}) = (x, a)] E_w^\pi[I[A_t = b] \mid I_t] \\ &\geq I[X_t = y, (X_{t-1}, A_{t-1}) = (x, a)] d_t \varrho(b) \end{aligned}$$

where (4.3) was used in the last step; if we recall that $\mathcal{F}_t \subset \sigma(I_t)$, this inequality and the inclusion $[(X_{t-1}, A_{t-1}) = (x, a)] \in \mathcal{F}_t$ together imply

$$\begin{aligned} E_w^\pi[I[(X_t, A_t) = (y, b), (X_{t-1}, A_{t-1}) = (x, a)] \mid \mathcal{F}_t] \\ &\geq d_t \varrho(b) E_w^\pi[I[X_t = y, (X_{t-1}, A_{t-1}) = (x, a)] \mid \mathcal{F}_t] \\ &= d_t \varrho(b) I[(X_{t-1}, A_{t-1}) = (x, a)] E_w^\pi[I[X_t = y] \mid \mathcal{F}_t] \\ &= d_t \varrho(b) I[(X_{t-1}, A_{t-1}) = (x, a)] P(y \mid X_{t-1}, A_{t-1}), \end{aligned}$$

where the second equality is due to the Markov property and the definition of the σ-field $\mathcal{F}_t$; see Remark 3.1. Thus,

$$\begin{aligned} E_w^\pi[I[(X_t, A_t) = (y, b), (X_{t-1}, A_{t-1}) = (x, a)] \mid \mathcal{F}_t] \\ \geq d_t P(y \mid x, a) \varrho(b) I[(X_{t-1}, A_{t-1}) = (x, a)]. \end{aligned} \tag{4.7}$$

To continue, notice that Definition 3.1(i) shows that for each $n > 0$,

$$\begin{aligned} n\nu_n(y, b) &\geq \sum_{t=1}^{n-1} I[X_t = y, A_t = b] \\ &= \widehat{W}_n + \sum_{t=1}^{n-1} E_w^\pi[I[X_t = y, A_t = b] \mid \mathcal{F}_t], \end{aligned} \tag{4.8}$$

where

$$\widehat{W}_n := \sum_{t=1}^{n-1} (I[X_t = y, A_t = b] - E_w^\pi[I[X_t = y, A_t = b] \mid \mathcal{F}_t]). \tag{4.9}$$

Since

$$\begin{aligned} E_w^\pi[I[(X_t, A_t) = (y, b) \mid \mathcal{F}_t] \\ \geq E_w^\pi[I[(X_t, A_t) = (y, b), (X_{t-1}, A_{t-1}) = (x, a)] \mid \mathcal{F}_t], \end{aligned}$$

from (4.7) and Definition 3.1 it follows that for $n > 1$,

$$\begin{aligned}\sum_{t=1}^{n-1} E_w^\pi[I[(X_t, A_t) = (y, b)] \,|\, \mathcal{F}_t] \\ &\geq P(y \,|\, x, a)\varrho(b) \sum_{t=1}^{n-1} d_t I[(X_{t-1}, A_{t-1}) = (x, a)] \\ &\geq P(y \,|\, x, a)\varrho(b) d_n \sum_{t=0}^{n-2} I[(X_t, A_t) = (x, a)] \\ &= P(y \,|\, x, a)\varrho(b) d_n \{n\nu_n(x, a) - I[(X_{n-1}, A_{n-1}) = (x, a)]\} \\ &\geq P(y \,|\, x, a)\varrho(b) d_n n\nu_n(x, a) - 1,\end{aligned}$$

where the second inequality used the fact that $\{d_n\}$ is a decreasing sequence. Together with (4.8) this implies that

$$\text{(4.10)} \qquad \frac{n\nu_n(y, b)}{d_n^{j+1} c_n} \geq \frac{\widehat{W}_n - 1}{d_n^{j+1} c_n} + P(y \,|\, x, a)\varrho(b) \,\frac{n\nu_n(x, a)}{d_n^j c_n}.$$

Since $d_n^{j+1} c_n \nearrow \infty$ and $\sum_n 1/(d_n^{j+1} c_n)^2 < \infty$ (see Remark 3.2), from the definition of $\widehat{W}_n$ in (4.9), an application of Lemma 3.1 yields

$$\lim_{n\to\infty} \frac{\widehat{W}_n - 1}{d_n^{j+1} c_n} = 0 \qquad P_w^\pi\text{-a.s.},$$

whereas the inclusion $(x, a) \in S_{j-1} \times A$ implies

$$\liminf_{n\to\infty} \frac{n\nu_n(x, a)}{d_n^j c_n} > 0 \qquad P_w^\pi\text{-a.s.},$$

by the induction hypothesis. Since $\varrho(\cdot) > 0$, these last two statements, (4.6), and inequality (4.10) imply that

$$\liminf_{n\to\infty} \frac{n\nu_n(y, b)}{d_n^{j+1} c_n} > 0 \qquad P_w^\pi\text{-a.s.};$$

since $(y, b) \in S_j \times A$ is arbitrary, this establishes the case $i = j$ of (4.4) and completes the induction argument.

(ii) Let $k \in \mathbb{K}^*$ be arbitrary so that $k \in S_i \times A$ for some $i < N$, by Lemma 4.1(ii). From (4.2) and Definition 3.1(i) it follows that

$$\begin{aligned}P_w^\pi[T_k = \infty] &= P_w^\pi[(X_t, A_t) \neq k \text{ for all } t \in \mathbb{N}] \\ &= P_w^\pi[\nu_t(k) = 0 \text{ for all } t = 1, 2, \ldots] = 0,\end{aligned}$$

where the last equality is due to part (i). ■

Proof of Theorem 4.1. Let $\pi \in \mathcal{P}^*$ and $w \in S$ be arbitrary. Since $d_n \in (0, 1]$ it follows that $1/(d_n^N c_n) \geq 1/(d_n^{i+1} c_n)$ for any $i = 0, 1, 2, \ldots, N - 1$, so that Lemma 4.2 implies that

$$\liminf_{n\to\infty} \frac{n\nu_n(k)}{d_n^N c_n} > 0 \quad P_x^\pi\text{-a.s.,} \tag{4.11}$$

for every $k \in \bigcup_{i=0}^{N-1}(S_i \times A) = \mathbb{K}^*$. Therefore, Lemma 3.2(ii) shows that for any $x \in S$ and $k \in \mathbb{K}^*$, $\lim_{n\to\infty}[P_n(x \,|\, k) - P(x \,|\, k)] = 0$ P_w^π-a.s., and the conclusion follows. ■

5. Non-stationary value iteration. This section introduces a form of the value iteration algorithm, a device that will be used later to formulate an optimal adaptive policy. According to an advice in Puterman (1994), firstly the original model will be modified by applying the following Schweitzer's transformation to the transition law P.

Definition 5.1 (Schweitzer, 1971). Let $M = \langle S, A, R, P\rangle$ be the MDP described in Section 2, and let $\alpha \in (0,1)$ be fixed.

(i) The modified transition law $Q = [Q(y \,|\, x, \cdot)]$ is defined by

$$Q(y \,|\, x, a) = (1-\alpha)\delta(x,y) + \alpha P(y \,|\, x, a), \qquad x, y \in S,\ a \in A, \tag{5.1}$$

where $\delta(\cdot,\cdot)$ is the Kronecker symbol on S, i.e., for $x, y \in S$, $\delta(x,x) = 1$ and $\delta(x,y) = 0$ if $x \neq y$. Similarly, for each $n \in \mathbb{N}$ set

$$Q_n(y \,|\, x, a) = (1-\alpha)\delta(x,y) + \alpha P_n(y \,|\, x, a), \qquad x, y \in S,\ a \in A, \tag{5.2}$$

where P_n is the estimator of the transition law P introduced in Definition 3.1(iv).

(ii) The transformed MDP $\widetilde{M}$ is given by $\widetilde{M} = \langle S, A, R, Q\rangle$.

The following lemma, which follows from Definition 5.1 via direct calculations, shows that the solution to the optimality equations associated to models M and $\widetilde{M}$ are simply related.

Lemma 5.1. (i) *Suppose that $g \in \mathbb{R}$ and $h\colon S \to \mathbb{R}$ satisfy* (2.4) *and define $H\colon S \to \mathbb{R}$ by*

$$H(x) = \frac{h(x)}{\alpha}, \qquad x \in S.$$

In this case $(g, H(\cdot))$ is a solution to the AROE associated to the transformed model $\widetilde{M}$:

$$g + H(x) = \sup_{a\in A}\Big[R(x,a) + \sum_{y\in S} Q(y \,|\, x, a) H(y)\Big], \qquad x \in S. \tag{5.3}$$

(ii) *Suppose that* (5.3) *is valid for the pair $(g, H(\cdot))$ and set $h(\cdot) = \alpha H(\cdot)$. In this case $(g, h(\cdot))$ satisfies the AROE* (2.4) *for the original model M.*

Remark 5.1. Throughout the remainder $(g, h(\cdot))$ and $(g, H(\cdot))$ stand for pairs satisfying (2.4) and (5.3), respectively, and $H(z) = 0 = h(z)$; such pairs exist and are unique, by Lemmas 2.1 and 5.1, and $\alpha H(\cdot) = h(\cdot)$.

Definition 5.2 (Federgruen and Schweitzer, 1981; Hernández-Lerma, 1988).

(i) The non-stationary value iteration functions $\{V_n\colon S \to \mathbb{R} \mid n = -1, 0, 1, 2, \dots\}$ are defined as follows: $V_{-1}(\cdot) = 0$, and

$$V_n(x) = \max_{a\in A}\Big[R(x,a) + \sum_{y\in S} Q_n(y\,|\,x,a)V_{n-1}(y)\Big], \qquad x \in S,\ n \in \mathbb{N}. \tag{5.4}$$

(ii) Let $z \in S$ be as in Assumption 2.1. The nth relative value function $H_n\colon S \to \mathbb{R}$ is given as follows:

$$H_n(x) = V_n(x) - V_n(z), \qquad x \in S,\ n = -1, 0, 1, 2, \dots, \tag{5.5}$$

whereas the nth differential reward $g_n \in \mathbb{R}$ is defined by

$$g_n = V_n(z) - V_{n-1}(z), \qquad n \in \mathbb{N}. \tag{5.6}$$

From this definition, it is not difficult to see that

$$\|V_{n-1}\| \le n\|R\|, \qquad n \in \mathbb{N}, \tag{5.7}$$

and after some computations using (5.6) and (5.5), equation (5.4) can be equivalently written as

$$g_n + H_n(x) = \max_{a\in A}\Big[R(x,a) + \sum_{y\in S} Q_n(y\,|\,x,a)H_{n-1}(y)\Big], \qquad n \in \mathbb{N},\ x \in S, \tag{5.8}$$

which resembles the AROE (5.3).

Remark 5.2. Notice that for each $k \in \mathbb{K}$, $x \in S$, and $n \in \mathbb{N}$, $Q_n(x\,|\,k)$ is a function of I_n; see Remark 3.1(ii). Consequently, g_n, $V_n(x)$ and $H_n(x)$ are always $\sigma(I_n)$-measurable random variables.

The main result of this section can now be stated as follows.

Theorem 5.1. *For each $x \in S^*$, $w \in S$ and $\pi \in \mathcal{P}^*$ assertions* (i) *and* (ii) *below hold P_w^π-a.s.*:

(i) $\lim_{n\to\infty} g_n = g$.

(ii) $\lim_{n\to\infty} H_n(x) = H(x)$; *equivalently*, $\lim_{n\to\infty} \alpha H_n(x) = h(x)$ (*see Remark* 5.1).

To establish this theorem it is convenient to introduce the following notation.

DEFINITION 5.3. (i) For each $n \in \mathbb{N}$, the random variable ε_n is given by

$$\varepsilon_n := \max_{k\in\mathbb{K}^*}\Big|\sum_{w\in S}[Q_n(w\,|\,k) - Q(w\,|\,k)]V_{n-1}(w)\Big|.$$

(ii) Given $f \in \mathbb{F}$ and $n \in \mathbb{N}$, matrices Q^f and $Q^{n;f}$ are determined as follows: for each $x, y \in S$,

$$Q^f_{xy} := Q(y\,|\,x, f(x)), \qquad Q^{n;f}_{xy} := Q_n(y\,|\,x, f(x)).$$

(iii) For $f_1, \ldots, f_k \in \mathbb{F}$, set $f_1^k := (f_1, \ldots, f_k)$ and define

$$M^k(f_1^k) := \prod_{i=1}^k Q^{f_i}, \qquad M_n^k(f_1^k) := \prod_{i=1}^k Q^{n-i+1;f_i} \quad \text{when } n \ge k.$$

The proof of Theorem 5.1 has been divided into four lemmas. The first one was stated as Theorem 5.1 in Cavazos-Cadena (1998), where it was proved using the inequality $Q(x\,|\,x,a) \ge 1-\alpha > 0$, which follows from (5.2).

LEMMA 5.2. *Let $z \in S$ be as in Assumption* 2.1. *There exist a positive integer N_0 and $\Delta \in (0,1)$ such that if $f_i \in \mathbb{F}$ for $i = 1, \ldots, N_0$, then*

$$M^{N_0}(f_1^{N_0})_{xz} > 2\Delta, \qquad x \in S.$$

LEMMA 5.3. *Let N_0 and Δ be as in Lemma* 5.2. *There exists a random variable L: $\mathbb{H}_\infty \to [N_0, \infty]$ such that*:

(i) *When $n > L$, the set S^* of essential states is closed with respect to each matrix $Q^{n;f}$, that is, given $f \in \mathbb{F}$, $x \in S^*$ and $n > L$, we have $Q^{n;f}_{xw} = 0$ for each $w \in S^{*\mathrm{c}}$.*

(ii) *For each $x \in S^*$ and $f_1, \ldots, f_{N_0} \in \mathbb{F}$, if $n > L$ then $M_n^{N_0}(f_1^{N_0})_{xy} > 0 \Rightarrow y \in S^*$, and*

$$M_n^{N_0}(f_1^{N_0})_{xz} > \Delta. \tag{5.9}$$

(iii) *For each $w \in S$ and $\pi \in \mathcal{P}^*$, $P_w^\pi[L < \infty] = 1$.*

Proof. For each $r \in \mathbb{N}$, define the event

$$\begin{aligned}\Omega_r &:= \bigcap_{n\ge r}\bigcap_{k\in\mathbb{K}^*}[P_n(S^*\,|\,k) = 1, \text{ and } \|P_n(\cdot\,|\,k) - P(\cdot\,|\,k)\| \le \Delta/(\alpha N_0)] \\ &= \bigcap_{n\ge r}\bigcap_{k\in\mathbb{K}^*}[Q_n(S^*\,|\,k) = 1, \text{ and } \|Q_n(\cdot\,|\,k) - Q(\cdot\,|\,k)\| \le \Delta/N_0];\end{aligned} \tag{5.10}$$

see Definition 5.1. With this notation and via Definition 5.3(ii), it is not difficult to see that the following assertion holds:

(a) Along a trajectory in Ω_r, for each $x \in S^*$ and $f \in \mathbb{F}$, if $n \ge r$ then $\sum_{w\in S^*} Q^{n;f}_{xw} = 1$.

From this, the product formula for $M_n^{N_0}(f_1^{N_0})$ in Definition 5.3(iii) allows us to obtain

(b) On the event Ω_r, if $n \geq N_0 + r$ and $x \in S^*$, then $M_n^{N_0}(f_1^{N_0})_{xy} > 0$ implies that $y \in S^*$.

Recall now that $P(S^* \mid k) = 1 = Q(S^* \mid k)$ for each $k \in \mathbb{K}^*$, by Lemma 4.1(i) and (5.2). Since along a trajectory in Ω_r the equality $Q_n(S^* \mid k) = 1$ holds when $n \geq r$, an induction argument shows that for each positive integer t the following is true (see (4.1)):

$$\text{(5.11)} \quad \text{On the event } \Omega_r,\ \|M_n^t(f_1^t) - M^t(f_1^t)\|_* \leq \sum_{i=1}^{t} \|Q^{n-i+1;f_i} - Q^{f_i}\|_*, \qquad f_1, \dots, f_t \in \mathbb{F},\ n \geq t + r.$$

On the other hand, combining Definition 5.3(ii) and (4.1) yields

$$\|Q^{n-i+1;f_i} - Q^{f_i}\|_* \leq \max_{k \in \mathbb{K}^*} \|Q_{n-i+1}(\cdot \mid k) - Q(\cdot \mid k)\|,$$

so that along trajectories in Ω_r,

$$\|Q^{n-i+1;f_i} - Q^{f_i}\|_* \leq \Delta/N_0, \qquad n - i + 1 \geq r,\ f_i \in \mathbb{F};$$

see (5.10). Next, let $x \in S^*$ be arbitrary but fixed. Since $z \in S^*$, (4.1) yields

$$|M_n^{N_0}(f_1^{N_0})_{x\,z} - M^{N_0}(f_1^{N_0})_{xz}| \leq \|M_n^{N_0}(f_1^{N_0}) - M^{N_0}(f_1^{N_0})\|_*$$

and combining these relations with the case $t = N_0$ of (5.11) shows that, on trajectories in Ω_r, if $x \in S^*$, then $|M_n^{N_0}(f_1^{N_0})_{xz} - M^{N_0}(f_1^{N_0})_{xz}| \leq \Delta$ is always valid when $n \geq N_0 + r$; since $M^{N_0}(f_1^{N_0})_{xz} > 2\Delta$, by Lemma 5.2, the next claim follows:

(c) For each $x \in S^*$ and $f_1, \dots, f_{N_0}$, if $n \geq N_0 + r$, then $M_n^{N_0}(f_1^{N_0})_{xz} > \Delta$ on the event Ω_r.

Now, observe that $\Omega_t \subset \Omega_{t+1}$, and define the random variable $L\colon \mathbb{H}_\infty \to [N_0, \infty]$ by

$$L := \begin{cases} N_0 + r & \text{on } \Omega_r \setminus \bigcup_{i<r} \Omega_i, \\ \infty & \text{on } \mathbb{H}_\infty \setminus \bigcup_{i=1}^{\infty} \Omega_i. \end{cases}$$

In this case assertions (a)–(c) above show that parts (i) and (ii) hold with this variable L. To conclude, let $w \in S$ and $\pi \in \mathcal{P}^*$ be arbitrary but fixed, and notice that Theorem 4.1 implies that

$$\lim_{r \to \infty} P_w^\pi\Big[\bigcap_{n \geq r} \bigcap_{k \in \mathbb{K}^*} \|P_n(\cdot \mid k) - P(\cdot \mid k)\| \leq \Delta/(\alpha N_0)\Big] = 1.$$

Next, set $T^* = \max\{T_k \mid k \in \mathbb{K}^*\}$ (see (4.2)) and observe that $P_w^\pi[T^* < \infty] = 1$, by Lemma 4.2(ii) and the finiteness of $\mathbb{K}^*$, and thus $P_w^\pi[T^* < r] \nearrow 1$ as $r \nearrow 1$. Since $[T^* < r] \subset \bigcap_{n \geq r} \bigcap_{k \in \mathbb{K}^*} [P_n(S^* \mid k) = 1]$, by Lemma 4.1(iii), it follows that

$$\lim_{r \to \infty} P_w^\pi\Big[\bigcap_{n \geq r} \bigcap_{k \in \mathbb{K}^*} [P_n(S^* \mid k) = 1]\Big] = 1,$$

so that $\lim_{r\to\infty} P_w^\pi[\Omega_r] = 1$, by (5.10), and thus

$$P_w^\pi[L<\infty] = P_w^\pi\Big[\bigcup_r \Omega_r\Big] = 1. \quad \blacksquare$$

To continue, for each $W\colon S\to\mathbb{R}$ define

$$\operatorname{sp}^*(W) = \max_{x\in S^*} W(x) - \min_{x\in S^*} W(x) \tag{5.12}$$

and notice that, via (5.7),

$$\operatorname{sp}^*(V_n) \le 2\|R\|(n+1), \qquad n\in\mathbb{N}. \tag{5.13}$$

LEMMA 5.4. *Let the random variable L be as in Lemma* 5.3. *On the event* $[L<\infty]$, *the sequence* $\{\operatorname{sp}^*(V_n)\}$ *is* (*pointwise*) *bounded.*

Proof. Given $n\in\mathbb{N}$, (5.4) and the finiteness of A ensure that there exists a policy $\phi_n\in\mathbb{F}$ such that

$$V_n(x) = R(x,\phi_n(x)) + \sum_w Q_n(w\,|\,x,\phi_n(x))V_{n-1}(w), \qquad x\in S; \tag{5.14}$$

notice that, for each state x, $\phi_n(x)$ is a random variable depending on I_n, by Remark 5.2. This equality implies that for every state x,

$$-\|R\| + \sum_w Q_n(w\,|\,x,\phi_n(x))V_{n-1}(w) \le V_n(x) \le \|R\| + \sum_w Q_n(w\,|\,x,\phi_n(x))V_{n-1}(w).$$

If we identify V_n with a column vector and use the notation of Definition 5.3, it follows that $-\|R\|\mathbb{1} + Q^{n;\phi_n}V_{n-1} \le V_n \le \|R\|\mathbb{1} + Q^{n;\phi_n}V_{n-1}$, where $\mathbb{1}$ is the vector of ones, and an induction argument shows that, for each $t\in\mathbb{N}$,

$$-(t+1)\|R\|\mathbb{1} + \Big[\prod_{i=0}^t Q^{n-i;\phi_{n-i}}\Big]V_{n-t-1} \le V_n \le (t+1)\|R\|\mathbb{1} + \Big[\prod_{i=0}^t Q^{n-i;\phi_{n-i}}\Big]V_{n-t-1}, \qquad n>t.$$

Now, *let n be an integer satisfying $n > L$* ($\ge N_0$). In this context, set $(f_1,\dots,f_{N_0})\colon= (\phi_n,\phi_{n-1},\dots,\phi_{n-N_0+1})$ and observe that, by Definition 5.3, the previous inequality with $t = N_0-1$ leads to

$$-N_0\|R\|\mathbb{1} + M_n^{N_0}(f_1^{N_0})V_{n-N_0} \le V_n \le N_0\|R\|\mathbb{1} + M_n^{N_0}(f_1^{N_0})V_{n-N_0}.$$

Next, *let $x,y\in S^*$ be arbitrary but fixed.* The above displayed relation implies that

$$V_n(x) - V_n(y) \le 2N_0\|R\| + \sum_{w\in S} M_n^{N_0}(f_1^{N_0})_{xw}V_{n-N_0}(w) - \sum_{w\in S} M_n^{N_0}(f_1^{N_0})_{yw}V_{n-N_0}(w), \tag{5.15}$$

and since $M_n^{N_0}(f_1^{N_0})_{xw} = 0$ when $w \notin S$, by Lemma 5.3(ii),

$$\begin{aligned}\sum_{w\in S} M_n^{N_0}(f_1^{N_0})_{xw} V_{n-N_0}(w)& \\ &= \sum_{w\in S^*} M_n^{N_0}(f_1^{N_0})_{xw} V_{n-N_0}(w)\\ &= \sum_{w\in S^*,\, w\neq z} M_n^{N_0}(f_1^{N_0})_{xw} V_{n-N_0}(w)\\ &\quad + (M_n^{N_0}(f_1^{N_0})_{xz} - \varDelta)V_{n-N_0}(z) + \varDelta V_{n-N_0}(z);\end{aligned}$$

since $M_n^{N_0}(f_1^{N_0})_{xz} - \varDelta > 0$, by (5.9), it follows that

$$\begin{aligned}\sum_{w\in S^*,\, w\neq z} & M_n^{N_0}(f_1^{N_0})_{xw} V_{n-N_0}(w) + (M_n^{N_0}(f_1^{N_0})_{xz} - \varDelta)V_{n-N_0}(z)\\ &\le \Big[\sum_{w\in S^*,\, w\neq z} M_n^{N_0}(f_1^{N_0})_{xw} + (M_n^{N_0}(f_1^{N_0})_{xz} - \varDelta)\Big] \max_{w\in S^*} V_{n-N_0}(w)\\ &= (1-\varDelta)\max_{w\in S^*} V_{n-N_0}(w)\end{aligned}$$

so that

$$\sum_{w\in S} M_n^{N_0}(f_1^{N_0})_{xw} V_{n-N_0}(w) \le (1-\varDelta)\max_{w\in S^*} V_{n-N_0}(w) + \varDelta V_{n-N_0}(z).$$

Similarly, it can be established that

$$\sum_{w\in S} M_n^{N_0}(f_1^{N_0})_{y\,w} V_{n-N_0}(w) \ge (1-\varDelta)\min_{w\in S^*} V_{n-N_0}(w) + \varDelta V_{n-N_0}(z).$$

The last two displayed inequalities together with (5.15) and (5.12) lead to $V_n(x) - V_n(y) \le 2N_0\|R\| + (1-\varDelta)\mathrm{sp}^*(V_{n-N_0})$, and since $x, y \in S^*$ and $n > L$ are arbitrary, it follows that

$$\mathrm{sp}^*(V_n) \le 2N_0\|R\| + (1-\varDelta)\mathrm{sp}^*(V_{n-N_0}), \qquad n > L.$$

Next, given $n > L$, let r be the first positive integer such that $n - rN_0 \le L$. Repeated application of the above inequality allows us to obtain

$$\begin{aligned}\mathrm{sp}^*(V_n) &\le 2N_0\|R\| \sum_{i=0}^{r-1}(1-\varDelta)^i + (1-\varDelta)^r \mathrm{sp}^*(V_{n-rN_0})\\ &\le \frac{2N_0\|R\|}{\varDelta} + \mathrm{sp}^*(V_{n-rN_0});\end{aligned}$$

since $n - rN_0 \le L$, via (5.13) it follows that $\mathrm{sp}^*(V_{n-rN_0}) \le 2(L+1)\|R\|$, and then

$$\mathrm{sp}^*(V_n) \le \frac{2N_0\|R\|}{\varDelta} + 2(L+1)\|R\|, \qquad n > L;$$

together with (5.13), this implies that $\{\mathrm{sp}^*(V_n)\}$ is pointwise bounded on the set $[L<\infty]$. ■

The following is the last step before the proof of Theorem 5.1.

LEMMA 5.5. *Let the random variable L be as in the previous lemmas, define the event Ω' by*

$$\Omega' = [L<\infty, \text{ and } \lim_{n\to\infty}\|Q_n(\cdot\,|\,k)-Q(\cdot\,|\,k)\| = 0 \text{ for all } k\in\mathbb{K}^*], \tag{5.16}$$

and set

$$D_n(\cdot) = V_n(\cdot) - ng - H(\cdot) \tag{5.17}$$

(*see Remark* 5.1). *The following convergences* (i) *and* (ii) *hold on the event Ω'*:

(i) $\lim_{n\to\infty}\varepsilon_n = 0$, *where ε_n is as in Definition* 5.3(i).

(ii) $\lim_{n\to\infty}\mathrm{sp}^*(D_n)=0$.

Proof. (i) Observe that

$$Q(S^*\,|\,k)=1, \quad \text{and} \quad Q_n(S^*\,|\,k)=1 \quad \text{when } n>L,\ k\in\mathbb{K}^*; \tag{5.18}$$

see Lemma 4.1(i), (5.2), and Lemma 5.3(i). Since $z\in S^*$, it follows that for $k\in\mathbb{K}^*$,

$$\begin{aligned}\Big|\sum_{w\in S}&[Q_n(w\,|\,k)-Q(w\,|\,k)]V_{n-1}(w)\Big|\\ &= \Big|\sum_{w\in S^*}[Q_n(w\,|\,k)-Q(w\,|\,k)](V_{n-1}(w)-V_{n-1}(z))\Big|\\ &\le \sum_{w\in S^*}|Q_n(w\,|\,k)-Q(w\,|\,k)|\mathrm{sp}^*(V_{n-1})\\ &= \mathrm{sp}^*(V_{n-1})\|Q_n(\cdot\,|\,k)-Q(\cdot\,|\,k)\|.\end{aligned}$$

Thus, $\varepsilon_n \le \mathrm{sp}^*(V_{n-1})\max_{k\in\mathbb{K}^*}\|Q_n(\cdot\,|\,k)-Q(\cdot\,|\,k)\|$ for $n>L$ and, via Lemma 5.4, the conclusion follows from the specification of Ω'.

(ii) Given a positive integer n, let $\phi_n\in\mathbb{F}$ be as in (5.14). From (5.3), it follows that $ng+h(x)\le R(x,\phi_n(x))+\sum_{w\in S}Q(w\,|\,x,\phi_n(x))[(n-1)g+h(w)]$, which together with (5.14) and (5.17) yields

$$\begin{aligned}D_n(x) \ge \sum_{w\in S}&Q_n(w\,|\,x,\phi_n(x))V_{n-1}(w)\\ &-\sum_{w\in S}Q(w\,|\,x,\phi_n(x))[(n-1)g+h(w)], \qquad x\in S.\end{aligned} \tag{5.19}$$

This is equivalent to

$$\begin{aligned} D_n(x) \geq & \sum_{w\in S}[Q_n(w\,|\,x,\phi_n(x)) - Q(w\,|\,x,\phi_n(x))]V_{n-1}(w) \\ & + \sum_{w\in S} Q(w\,|\,x,\phi_n(x))D_{n-1}(w) \end{aligned}$$

and thus

$$D_n(x) \geq -\varepsilon_n + \sum_{w\in S^*} Q^{\phi_n}_{xw} D_{n-1}(w), \qquad x \in S^*;$$

see Definition 5.3 and observe that $Q^{\phi_n}_{xw} = Q(w\,|\,x,\phi_n(x)) = 0$ when $x \in S^*$ and $w \notin S^*$, by Lemma 4.1(i) and (5.1). From this, an induction argument shows that for each $t \in \mathbb{N}$,

$$D_n(x) \geq -\sum_{i=0}^{t-1} \varepsilon_{n-i} + \sum_{w\in S^*} \Big[\prod_{i=0}^{t} Q^{\phi_{n-i}}\Big]_{xw} D_{n-t-}(w), \qquad x \in S^*,\ n > t,$$

or, with the notation in Definition 5.3,

$$D_n(x) \geq -\sum_{i=0}^{t-1} \varepsilon_{n-i} + \sum_{w\in S^*} M^t(f_1^t)_{xw} D_{n-t-1}(w), \qquad x \in S^*,\ n > t, \tag{5.20}$$

where $(f_1,\ldots,f_t) := (\phi_n,\phi_{n-1},\ldots,\phi_{n-t+1})$. Observe now that $M^{N_0}(f_1^{N_0})_{xz} > \Delta$, by Lemma 5.2, so that

$$\begin{aligned} & \sum_{w\in S^*} M^{N_0}(f_1^{N_0})_{xw} D_{n-N_0}(w) \\ & = \sum_{w\in S^*,\, w\neq z} M^{N_0}(f_1^{N_0})_{x\,w} D_{n-N_0}(w) \\ & \quad + (M^{N_0}(f_1^{N_0})_{x\,w} - \Delta) D_{n-N_0}(z) + \Delta D_{n-N_0}(z) \\ & \geq \Big[\sum_{w\in S^*,\, w\neq z} M^{N_0}(f_1^{N_0})_{xw} + (M^{N_0}(f_1^{N_0})_{xz} - \Delta)\Big] \min_{w\in S^*} D_{n-N_0}(w) \\ & \quad + \Delta D_{n-N_0}(z) \\ & = (1-\Delta)\min_{w\in S^*} D_{n-N_0}(w) + \Delta D_{n-N_0}(z), \end{aligned}$$

and together with the case $t = N_0$ of (5.20) this implies

$$D_n(x) \geq -\sum_{i=0}^{N_0-1} \varepsilon_{n-i} + (1-\Delta)\min_{w\in S^*} D_{n-N_0}(w) + \Delta V_{n-N_0}(z), \qquad x \in S^*,\ n > N_0. \tag{5.21}$$

Next, let $f \in \mathbb{F}$ be such that, for each $x \in S$, $f(x)$ maximizes the right hand side of (5.3), so that

$$ng + h(y) = R(y, f(y)) + \sum_{w\in S} Q(w \,|\, y, f(y))[(n-1)g + h(w)]$$

for every $y \in S$ and $n > 0$. Since

$$V_n(y) \le R(y, f(y)) + \sum_{w\in S} Q_n(w \,|\, y, f(y))V_{n-1}(y),$$

by (5.4), it follows that for each $n > 0$,

$$\begin{aligned} D_n(y) \le \; & \sum_{w\in S} Q_n(w \,|\, y, f(y))V_{n-1}(y) \\ & - \sum_{w\in S} Q(w \,|\, y, f(y))[(n-1)g + h(y)], \qquad y \in S. \end{aligned}$$

Paralleling the argument used to go from (5.19) to (5.21), it can be established that

$$D_n(y) \le \sum_{i=0}^{N_0-1} \varepsilon_{n-i} + (1-\varDelta) \max_{w\in S^*} D_{n-N_0}(w) + \varDelta V_{n-N_0}(z), \qquad y \in S^*,\ n > N_0.$$

Combining this with (5.21) and (5.12) implies that, for each $x, y \in S^*$, $D_n(y) - D_n(x) \le 2\sum_{i=0}^{N_0-1} \varepsilon_{n-i} + (1-\varDelta)\mathrm{sp}^*(D_{n-N_0})$, and thus

$$\mathrm{sp}^*(D_n) \le 2 \sum_{i=0}^{N_0-1} \varepsilon_{n-i} + (1-\varDelta)\mathrm{sp}^*(D_{n-N_0}), \qquad n > N_0. \tag{5.22}$$

To conclude, let $\mathfrak{h}_\infty \in \varOmega'$ be arbitrary but fixed, and notice that part (i) implies that $\{\varepsilon_n\}$ is bounded along this trajectory, say $\varepsilon_n \le b \equiv b(\mathfrak{h}_\infty)$ for each $n \in \mathbb{N}$. In this case, (5.22) leads to $\mathrm{sp}^*(D_n) \le 2bN_0 + (1-\varDelta)\mathrm{sp}^*(D_{n-N_0})$, for each $n > n_0$, and along the same lines as in the proof of Lemma 5.4 this implies that $\{\mathrm{sp}^*(D_n)\}$ is bounded, so that $\limsup_{n\to\infty} \mathrm{sp}^*(D_n) < \infty$. Taking the upper limit on both sides of (5.22), via part (i) it follows that

$$\limsup_{n\to\infty}\ \mathrm{sp}^*(D_n) \le (1-\varDelta) \limsup_{n\to\infty}\ \mathrm{sp}^*(D_n),$$

and then $\limsup_{n\to\infty} \mathrm{sp}^*(D_n) = 0$, since $\varDelta$ is positive. ■

Proof of Theorem 5.1. Since $H(z) = 0$ (see Remark 5.1), (5.5) and (5.17) yield

$$\begin{aligned} H_n(x) - H(x) &= [V_n(x) - V_n(z)] - H(x) + H(z) \\ &= [V_n(x) - ng - H(x)] - [V_n(z) - ng - H(z)] \\ &= D_n(x) - D_n(z), \end{aligned}$$

and as $z \in S^*$, from (5.12) it follows that $|H_n(x) - H(x)| \le \mathrm{sp}^*(D_n)$ for every $x \in S^*$ and thus, by Lemma 5.5(ii),

$$\text{(5.23)} \qquad \text{on the event } \Omega',\ \lim_{n\to\infty} H_n(x) = H(x),\ x \in S^*.$$

On the other hand, on Ω' the random variable L is finite, and by (5.16), $\lim_{n\to\infty} \|Q_n(\cdot\,|\,z,a) - Q(\cdot\,|\,z,a)\| = 0$. Therefore, since A is finite, via (5.18) it follows that along trajectories in Ω',

$$\begin{aligned}\lim_{n\to\infty} \max_{a\in A} \Big[R(z,a) + \sum_{y\in S} Q_n(y\,|\,z,a)H_{n-1}(y)\Big]& \\ &= \lim_{n\to\infty} \max_{a\in A} \Big[R(z,a) + \sum_{y\in S^*} Q_n(y\,|\,z,a)H_{n-1}(y)\Big] \\ &= \max_{a\in A} \Big[R(z,a) + \sum_{y\in S^*} Q(y\,|\,z,a)H(y)\Big] \\ &= \max_{a\in A} \Big[R(z,a) + \sum_{y\in S} Q(y\,|\,z,a)H(y)\Big],\end{aligned}$$

and together with (5.23), (5.8) and (5.3), this shows that

$$\text{(5.24)} \qquad \lim_{n\to\infty} g_n = g \qquad \text{on the event } \Omega'.$$

To conclude, observe that $P_w^\pi[\lim_{n\to\infty} \|Q_n(\cdot\,|\,k) - Q(\cdot\,|\,k)\| = 0$ for all $k \in \mathbb{K}^*] = 1$ by Theorem 4.1 and (5.2), whereas $P_w^\pi[L < \infty] = 1$ by Lemma 5.3(ii). Thus, formula (5.16) leads to $P_w^\pi[\Omega'] = 1$, so that (5.23) and (5.24) yield the conclusions of Theorem 5.1. ∎

7. The adaptive policy.

The results in the previous sections will now be used to construct an optimal adaptive policy. On a given trajectory in $\mathbb{H}_\infty$ let the policy ϕ_n be as in (5.14), and notice that, as already observed in the proof of Lemma 5.4, $\phi_n(x)$ is a function of I_n for each $x \in S$.

Definition 6.1. The *NVI adaptive policy* π^* is specified as follows: For each $t \in \mathbb{N}$ and $\mathfrak{h}_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t) \in \mathbb{H}_t$,

$$\pi_t^*(B\,|\,\mathfrak{h}_t) = (1-d_t)\delta_{\phi_t(x_t)}(B) + d_t\varrho(B), \qquad B \subset A,$$

where, for each $a \in A$, δ_a stands for the Dirac probability measure concentrated at a, and $\varrho(\cdot)$ is as in Definition 4.2.

When the system is driven by π^*, the actions are selected using a random mechanism: If state $X_n = x$ is observed, then with probability $1-d_n$ the action applied is $A_n = \phi_n(x)$, whereas with probability d_n the control A_n is selected among all the available actions according to the probability distribution $\varrho(\cdot)$; notice that $\pi^* \in \mathcal{P}^*$. The main result of this note is the following.

Theorem 6.1. *The NVI adaptive policy π^* is optimal.*

The proof of this theorem is based on the two lemmas below; as usual (Hernández-Lerma, 1988), the argument uses the *discrepancy function* Φ: $\mathbb{K} \to \mathbb{R}$ given by

$$\Phi(x,a) = g + h(x) - R(x,a) - \sum_{y\in S} P(y\,|\,x,a)h(y), \qquad (x,a)\in\mathbb{K}, \tag{6.1}$$

where $(g, h(\cdot))$ is the unique solution of (2.4) satisfying $h(z) = 0$.

Lemma 6.1. *For each $w \in S$ and $\pi \in \mathcal{P}^*$,*

(i) $\lim_{n\to\infty} \Phi(X_n, \phi_n(X_n))I[X_n \in S^*] = 0$ P_w^π*-a.s., and consequently,*
(ii) $\lim_{n\to\infty} E_w^\pi[\Phi(X_n, \phi_n(X_n))] = 0$.

Proof. From (5.2) and Definition 5.2, it is not difficult to see that (5.14) is equivalent to

$$\begin{aligned} g_n + \alpha H_n(x) &+ (1-\alpha)[H_n(x) - H_{n-1}(x)] \\ &= R(x,\phi_n(x)) + \sum_{y\in S} P_n(y\,|\,x,\phi_n(x))[\alpha H_{n-1}(y)], \end{aligned}$$

and via (6.1) it follows that for every state x,

$$\begin{aligned} \Phi(x,\phi_n(x)) = [g - g_n] &+ [h(x) - \alpha H_n(x)] \\ &- (1-\alpha)[H_n(x) - H_{n-1}(x)] \\ &+ \alpha\sum_{y\in S}[P_n(y\,|\,x,\phi_n(x)) - P(y\,|\,x,\phi_n(x))]H_{n-1}(y) \\ &+ \sum_{y\in S} P(y\,|\,x,\phi_n(x))[\alpha H_{n-1}(y) - h(y)]. \end{aligned} \tag{6.2}$$

Next, let $w \in S$ and $\pi \in \mathcal{P}^*$ be arbitrary, and recall that $P(S^*\,|\,x,a) = 1$ when $(x,a) \in \mathbb{K}^* = S^* \times A$, by Lemma 4.1(i), so that

$$\begin{aligned} &\Big|\sum_{y\in S} P(y\,|\,x,\phi_n(x))[\alpha H_{n-1}(y) - h(y)]\Big| \\ &\qquad = \Big|\sum_{y\in S^*} P(y\,|\,x,\phi_n(x))[\alpha H_{n-1}(y) - h(y)]\Big| \\ &\qquad \le \sum_{y\in S^*} |\alpha H_{n-1}(y) - h(y)| \to 0 \quad P_w^\pi\text{-a.s., } x\in S^*, \end{aligned} \tag{6.3}$$

by Theorem 5.1(ii). On the other hand, via (5.2) and (5.5), it follows that

$$\begin{aligned}\alpha\sum_{y\in S}&[P_n(y\,|\,x,\phi_n(x))-P(y\,|\,x,\phi_n(x))]H_{n-1}(y)\\&=\sum_{y\in S}[Q_n(y\,|\,x,\phi_n(x))-Q(y\,|\,x,\phi_n(x))](V_{n-1}(y)-V_{n-1}(z))\\&=\sum_{y\in S}[Q_n(y\,|\,x,\phi_n(x))-Q(y\,|\,x,\phi_n(x))]V_{n-1}(y)\end{aligned}$$

and then for each $x\in S^*$,

$$\Big|\alpha\sum_{y\in S}[P_n(y\,|\,x,\phi_n(x))-P(y\,|\,x,\phi_n(x))]H_{n-1}(y)\Big|\le\varepsilon_n\to 0\qquad P_w^\pi\text{-a.s.},$$

by Definition 5.3(i) and Lemma 5.5(i). Combining this with (6.3) and Theorem 5.1, and letting $n\to\infty$ on both sides of (6.2) we deduce that $P_w^\pi[\lim_{n\to\infty}\Phi(x,\phi_n(x))=0]=1$ for each $x\in S^*$, and part (i) is a consequence of $|\Phi_n(X_n,\phi(X_n))I[X_n\in S^*]|\le\sum_{x\in S^*}\Phi(x,\phi_n(x))$. Now, the bounded convergence theorem implies $E_w^\pi[\Phi_n(X_n,\phi(X_n))I[X_n\in S^*]]\to 0$, and (ii) follows by observing that

$$|E_w^\pi[\Phi_n(X_n,\phi(X_n))I[X_n\notin S^*]]|\le\|\Phi\|P_w^\pi[X_n\notin S^*]\to 0,$$

by Lemma 4.1(iv). ■

Lemma 6.2. *Let* π^* *be the NVI adaptive policy in Definition* 6.1. *For each* $w\in S$,

$$\lim_{n\to\infty}E_w^{\pi^*}[\Phi(X_n,A_n)]=0.$$

Proof. Let $n\in\mathbb{N}$ and $w\in S$ be arbitrary. By the Markov property, the specification of π^* yields

$$\begin{aligned}E_w^{\pi^*}[\Phi(X_n,A_n)\,|\,I_n]=\Phi(X_n,\phi_n(X_n))&\\+d_n\Big(\sum_{a\in A}\varrho(a)\Phi(X_n,a)-\Phi(X_n,\phi_n(X_n))\Big)&\end{aligned}$$

so that $|E_w^{\pi^*}[\Phi(X_n,A_n)\,|\,I_n]-\Phi(X_n,\phi_n(X_n))|\le 2d_n\|\Phi\|$, and thus

$$\begin{aligned}|E_w^{\pi^*}&[\Phi(X_n,A_n)]-E_w^{\pi^*}[\Phi(X_n,\phi_n(X_n))]|\\&=|E_w^{\pi^*}[E_w^{\pi^*}[\Phi(X_n,A_n)\,|\,I_n]-\Phi(X_n,\phi_n(X_n))]|\\&\le E_w^{\pi^*}[|E_w^{\pi^*}[\Phi(X_n,A_n)\,|\,I_n]-\Phi(X_n,\phi_n(X_n))|]\le 2d_n\|\Phi\|.\end{aligned}$$

Since $d_n\to 0$, this last inequality and Lemma 6.1(ii) applied to policy $\pi^*\in\mathcal{P}^*$ together imply that $\lim_{n\to\infty}E_w^{\pi^*}[\Phi(X_n,A_n)]=0$. ■

Proof of Theorem 6.1. An induction argument using (6.1) shows that for each $w \in S$ and $n \in \mathbb{N}$,

$$g + \frac{h(x)}{n+1} = \frac{1}{n+1} E_w^{\pi}\Big[\sum_{t=0}^{n} [R(X_n, A_n) + \varPhi(X_n, A_n)]\Big]. \tag{6.4}$$

Observe now that Lemma 6.2 implies that

$$\lim_{n\to\infty} \frac{1}{n+1} E_w^{\pi^*}\Big[\sum_{t=0}^{n} \varPhi(X_n, A_n)\Big] = 0, \qquad w \in S,$$

and thus, replacing π by π^* in (6.4) and letting $n \to \infty$ on both sides of the resulting equality, we conclude that for each $w \in S$,

$$g = \lim_{n\to\infty} \frac{1}{n+1} E_w^{\pi^*}\Big[\sum_{t=0}^{n} R(X_n, A_n)\Big],$$

so that π^* is optimal; see (2.2), (2.3) and Lemma 2.1(ii). ■

References

A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh and S. I. Marcus (1993), *Discrete-time controlled Markov processes with average cost criterion: a survey*, SIAM J. Control Optim. 31, 282–334.

V. S. Borkar (1984), *On minimum cost per unit time control of Markov chains*, ibid. 22, 965–978.

V. S. Borkar (1996), *Recursive self-tuning of finite Markov chains*, Appl. Math. (Warsaw) 24, 169–188.

R. Cavazos-Cadena (1991), *Nonparametric estimation and adaptive control in a class of finite Markov decision chains*, Ann. Oper. Res. 28, 169–184.

R. Cavazos-Cadena (1996), *Value iteration in a class of communicating Markov decision chains with the average cost criterion*, SIAM J. Control Optim. 34, 1848–1873.

R. Cavazos-Cadena (1998), *A note on the convergence rate of the value iteration scheme in controlled Markov chains*, Systems Control Lett. 33, 221–230.

E. Drabik and Ł. Stettner (2000), *On adaptive control of Markov chains using nonparametric estimation*, Appl. Math. (Warsaw) 27, 143–152.

T. E. Duncan, B. Pasik-Duncan and Ł. Stettner (1998), *Discretized maximum likelihood and almost optimal adaptive control of ergodic Markov models*, SIAM J. Control Optim. 36, 422–446.

A. Federgruen and P. J. Schweitzer (1981), *Nonstationary Markov decision problems with converging parameters*, J. Optim. Theory Appl., 34, 207–241.

O. Hernández-Lerma (1988), *Adaptive Markov Control Processes*, Springer, New York.

M. Loève (1977), *Probability Theory I*, Springer, New York.

M. L. Puterman (1994), *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, New York.

Z. Ren and B. H. Krogh (2001), *Adaptive control of Markov chains with average cost*, IEEE Trans. Automat. Control 46, 613–617.

P. J. Schweitzer (1971), *Iterative solution of the functional equations of undiscounted Markov renewal programming*, J. Math. Anal. Appl. 34, 495–501.

L. C. Thomas (1980), *Connectedness conditions for denumerable state Markov decision processes*, in: Recent Developments in Markov Decision Processes, R. Hartley, L. C. Thomas and D. J. White (eds.), Academic Press, New York, 181–204.

Departamento de Estadística y Cálculo
Universidad Autónoma Agraria Antonio Narro
Buenavista, Saltillo COAH 25315, México
E-mail: rcavazos@narro.uaaan.mx

Departamento de Matemáticas
Universidad Autónoma Metropolitana
Campus Iztapalapa
Avenida San Rafael Atlixco #186
Colonia Vicentina
México 09340, D.F., México
E-mail: momr@xanum.uam.mx

Received on 25.11.2003 (1716)