J. Appl. Prob. **52,** 419–440 (2015)
Printed in England

SAMPLE-PATH OPTIMAL STATIONARY POLICIES IN STABLE MARKOV DECISION CHAINS WITH THE AVERAGE REWARD CRITERION

ROLANDO CAVAZOS-CADENA,* *Universidad Autónoma Agraria Antonio Narro*
RAÚL MONTES-DE-OCA,** *Universidad Autónoma Metropolitana-Iztapalapa*
KAREL SLADKÝ,*** *Institute of Information Theory and Automation*

Abstract

This paper concerns discrete-time Markov decision chains with denumerable state and compact action sets. Besides standard continuity requirements, the main assumption on the model is that it admits a Lyapunov function ℓ. In this context the average reward criterion is analyzed from the sample-path point of view. The main conclusion is that if the expected average reward associated to ℓ^2 is finite under any policy then a stationary policy obtained from the optimality equation in the standard way is sample-path average optimal in a strong sense.

Keywords: Dominated convergence theorem for the expected average criterion; discrepancy function; Kolmogorov inequality; innovations; strong sample-path optimality

2010 Mathematics Subject Classification: Primary 90C40
Secondary 93E20; 60J05

1. Introduction

This paper is concerned with discrete-time Markov decision processes (MDPs) with denumerable state-space and time-invariant transition mechanism. Within this context, the existence of optimal stationary policies with respect to a strong sample-path average index is analyzed. This problem has been studied in the literature, and the available results can be briefly described as follows: conditions on the model are imposed such that the expected average cost optimality equation has a solution, which generates an expected average optimal stationary policy f in the standard way. Then it is proved that such a policy f is also sample-path average optimal. Roughly, the requirements used to obtain such a conclusion involve, either a special structure on the cost function, or conditions on the transition law implying geometric ergodicity with respect to a (certain weighted) norm. In this paper the average reward criterion is studied, and the main difference with respect to the available results is that neither a special structure on the reward function is imposed, nor the stability assumptions used in this paper imply geometric ergodicity.

When the performance of a control strategy is measured by an expected average criterion, the analysis of the model is based on the optimality equation, which can be solved under diverse

Received 9 January 2014; revision received 12 June 2014.
* Postal address: Departamento de Estadística y Cálculo, Universidad Autónoma Agraria Antonio Narro, Buenavista, Saltillo, COAH, 25315, México. Email address: r.cavazoscadena@gmail.com
** Postal address: Departamento de Matemáticas, Universidad Autónoma Metropolitana, Campus Iztapalapa, Avenida San Rafael Atlixco #186, Colonia Vicentina, México 09340, D. F. México.
*** Postal address: Institute of Information Theory and Automation, Pod Vodárenskou věží 4, CZ-182 08, Praha 8, Czech Republic.

communication-stability conditions (see Thomas (1980) and Arapostathis *et al.* (1993)) and, in some sense, such conditions are also necessary (see Cavazos-Cadena (1988), (1989)). Among the different requirements ensuring that the optimality equation has a solution rendering an expected average optimal stationary policy f, the most general one is the so-called *Lyapunov function condition* which, extending ideas by Foster (1953) on uncontrolled Markov chains, was formulated by Hordijk (1974). In addition to standard continuity-compactness requirements, the basic structural assumption in this work is that the system has a Lyapunov function ℓ.

On the other hand, an expected average criterion is quite appropriate if the controller repeats the underlying random dynamical experiment many times under similar conditions, but not for a single trial. As already stated, in this paper the average index is studied from a *sample-path perspective*, and the analysis involves the following idea.

A policy π^* is average optimal in the sample-path sense if there exists a constant, say g^*, such that under the action of π^* and regardless of the initial state, the average of the observed rewards over a finite horizon t converges to g^* as $t \to \infty$ with probability 1, whereas under any other policy, the *superior limit* of such averages is always bounded above by g^* almost surely (a.s.).

It was recently shown in Cavazos-Cadena *et al.* (2014) that, under the sole assumption that the system admits a Lyapunov function, the existence of a sample-path average optimal stationary policy cannot be ensured. *The main result* of this note can be briefly described as follows. If the MDP admits a Lyapunov function ℓ and, regardless of the initial state and the policy employed, the expected average reward corresponding to ℓ^2 is finite, then a stationary policy f obtained from the optimality equation in the standard way is also sample-path average optimal.

The theory and applications of MDPs have been extensively studied; see, for instance, Hernández-Lerma (1989), Puterman (1994), Sennott (1999), and Bäuerle and Rieder (2010), (2011). Concerning the idea of sample-path average optimality, it is known that if the optimality equation has a bounded solution then the stationary policy f referred to above is optimal in the sample-path sense (Arapostathis *et al.* (1993)). On the other hand, for MDPs with denumerable state-space and endowed with the average cost criterion (see Borkar (1984), (1991)) it was proved that if the cost function has a 'penalized structure', in the sense that it is sufficiently large outside a compact set, then a sample-path average optimal stationary policy exists, a conclusion that has been extended to models evolving on Borel spaces in Lasserre (1999) and Vega-Amaya (1999). Under geometric ergodicity conditions, the existence of sample-path optimal stationary policies was established in Hernández-Lerma *et al.* (1999) and in Zhu and Guo (2006) for models on Borel spaces, whereas Hunt (2005) considered MDPs with denumerable state-space and finite action sets; the sample-path perspective in a continuous-time framework is used in Dai Pra *et al.* (1999).

The approach used below to establish the aforementioned result is based on (i) a dominated convergence theorem for the average reward criterion, and (ii) a direct analysis of the trajectories of the state-action process using Kolmogorov's inequality and the first Borel–Cantelli lemma.

The organization of the subsequent material is as follows: in Section 2 the decision model is presented and the superior and inferior limit expected average criteria, as well as the corresponding optimality equation, are briefly discussed. Next, in Section 3 the idea of a Lyapunov function is introduced and some basic consequences of the existence of such a mapping are established, whereas in Section 4 the main result on sample-path average optimal stationary policies is stated as Theorem 4.1. From this point onwards, the remainder of the paper is dedicated to proving that result, and the rather involved argument has been divided into four

parts: Sections 5 and 6 concern the necessary technical tools involving the *expected* average reward optimality equation, whereas in Section 7 we present a direct analysis of the trajectories of the state-action process. The final step is established in Section 8, where the proof of the main result is presented.

Notation. Throughout this paper $\mathbb{N}$ stands for the set of all nonnegative integers. For a topological space $\mathbb{K}$, the class of all continuous functions defined on $\mathbb{K}$ and the Borel σ-field of $\mathbb{K}$ are denoted by $\mathcal{C}(\mathbb{K})$ and $\mathcal{B}(\mathbb{K})$, respectively.

2. The decision model

Let $\mathcal{M} = (S, A, \{A(x)\}_{x\in S}, R, \mathbb{P})$ be the usual MDP, where the state-space S is a denumerable set endowed with the discrete topology and the action set A is a metric space. For each $x \in S$, $A(x) \subset A$ is the nonempty subset of admissible actions, and

$$R \in \mathcal{C}(\mathbb{K}) \tag{2.1}$$

is the reward function, where $\mathbb{K} := \{(x, a) \mid x \in S, a \in A(x)\}$ is the space of admissible pairs. On the other hand, $\mathbb{P} = (\mathbb{P}_{xy}\{\cdot\})$ is the controlled transition law on S given $\mathbb{K}$, that is for all $(x, a) \in \mathbb{K}$ and $y \in S$ the relations $\mathbb{P}_{xy}(a) \geq 0$ and $\sum_{y\in S} \mathbb{P}_{xy}(a) = 1$ are satisfied. In this model $\mathcal{M}$ is interpreted as follows: at each time $t \in \mathbb{N}$ the decision maker knows the previous states and actions and observes the current state, say $X_t = x \in S$. Using that information, the controller selects an action (control) $A_t = a \in A(x)$ and two things happen: a reward $R(x, a)$ is obtained by the controller, and the system moves to a new state $X_{t+1} = y \in S$ with probability $\mathbb{P}_{xy}(a)$. In what follows several continuous reward functions will be considered, but all of the other components of $\mathcal{M}$ will be fixed. The following condition will be enforced even without explicit reference.

Assumption 2.1. (i) *For each $x \in S$, $A(x)$ is a compact subset of A.*

(ii) *For every $x, y \in S$, the mapping $a \mapsto \mathbb{P}_{xy}(a)$ is continuous in $a \in A(x)$.*

Policies. The set $\mathbb{H}_t$ of possible histories up to time $t \in \mathbb{N}$ is defined by $\mathbb{H}_0 := S$ and $\mathbb{H}_t := \mathbb{K}^t \times S$ for $t \geq 1$; a generic element of $\mathbb{H}_t$ is denoted by $\boldsymbol{h}_t = (x_0, a_0, \ldots, x_i, a_i, \ldots, x_t)$, where $a_i \in A(x_i)$ and $x_i \in S$. A policy $\pi = \{\pi_t\}$ is a special sequence of stochastic kernels: for each $t \in \mathbb{N}$ and $\boldsymbol{h}_t \in \mathbb{H}_t$, $\pi_t\{\cdot \mid \boldsymbol{h}_t\}$ is a probability measure on $\mathcal{B}(A)$ satisfying (i) $\pi_t\{A(x_t) \mid \boldsymbol{h}_t\} = 1$, and (ii) for each $B \in \mathcal{B}(A)$, the mapping $\boldsymbol{h}_t \mapsto \pi_t\{B \mid \boldsymbol{h}_t\}$, $\boldsymbol{h}_t \in \mathbb{H}_t$, is Borel-measurable. When the controller chooses actions according to π, the control A_t applied at time t belongs to $B \subset A$ with probability $\pi_t\{B \mid \boldsymbol{h}_t\}$, where $\boldsymbol{h}_t$ is the observed history of the process up to time t. The class of all policies is denoted by $\mathcal{P}$ and, given the policy π being used for choosing actions and the initial state $X_0 = x$, the distribution of the state-action process $\{(X_t, A_t)\}$ is uniquely determined (Puterman (1994)); such a distribution and the corresponding expectation operator are denoted by $\mathbb{P}_x^\pi$ and $\mathbb{E}_x^\pi$, respectively. Next, define $\mathbb{F} := \prod_{x\in S} A(x)$ and notice that $\mathbb{F}$ is a compact metric space, which consists of all functions $f : S \to A$ such that $f(x) \in A(x)$ for each $x \in S$. A policy π is Markovian if there exists a sequence $\{f_t\} \subset \mathbb{F}$ such that the probability measure $\pi_t\{\cdot \mid \boldsymbol{h}_t\}$ is always concentrated at $f_t(x_t)$, and if $f_t \equiv f$ for every t, the Markovian policy π is referred to as stationary. The classes of stationary and Markovian policies are naturally identified with $\mathbb{F}$ and $\mathbb{M} := \prod_{t=0}^{\infty} \mathbb{F}$, respectively, and with these conventions $\mathbb{F} \subset \mathbb{M} \subset \mathcal{P}$.

Expected average criteria. Assume that $R(X_t, A_t)$ has finite expectation with respect to every distribution $\mathbb{P}_x^\pi$. In this context, the (long-run superior limit) average reward criterion

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

corresponding to $\pi \in \mathcal{P}$ at state $x \in S$ is defined by

$$J(x,\pi) := \limsup_{k\to\infty} \frac{1}{k}\mathbb{E}_x^\pi\left\{\sum_{t=0}^{k-1} R(X_t, A_t)\right\}, \tag{2.2}$$

whereas the corresponding optimal value function is

$$J^*(x) := \sup_{\pi\in\mathcal{P}} J(x,\pi), \qquad x \in S;$$

a policy $\pi^* \in \mathcal{P}$ is (limsup) expected average optimal if $J(x,\pi^*) = J^*(x)$ for every $x \in S$. The criterion (2.2) represents an optimistic perspective of the decision maker, since the performance of a policy is evaluated by the largest among the limit points of the expected average rewards in finite times. The pessimistic point of view is represented by the following index, assessing the performance of a policy in terms of the smallest of such limit points:

$$J_-(x,\pi) := \liminf_{k\to\infty} \frac{1}{k}\mathbb{E}_x^\pi\left\{\sum_{t=0}^{k-1} R(X_t, A_t)\right\} \tag{2.3}$$

is the (long-run) inferior limit average index associated to $\pi \in \mathcal{P}$ at state x, and

$$J_-^*(x) := \sup_{\pi\in\mathcal{P}} J_-(x,\pi), \qquad x \in S$$

is the corresponding optimal value function; from this specification it follows that

$$J_-^*(\cdot) \le J^*(\cdot).$$

3. Lyapunov functions

A fundamental instrument to analyze the above criteria is the following optimality equation:

$$g + h(x) = \sup_{a\in A(x)}\left[R(x,a) + \sum_{y\in S}\mathbb{P}_{xy}(a)h(y)\right], \qquad x \in S, \tag{3.1}$$

where $g \in \mathbb{R}$ and $h \in \mathcal{C}(S)$ is a given function. Assume that the pair $(g, h(\cdot)) \in \mathbb{R}\times\mathcal{C}(S)$ satisfies (3.1) and that the following properties are valid. For each $x \in S$ and $\pi \in \mathcal{P}$,

(i) $\mathbb{E}_x^\pi\{|h(X_n)|\} < \infty$ for each $n = 1, 2, 3, \ldots$,

(ii)
$$\frac{\mathbb{E}_x^\pi\{|h(X_n)|\}}{n} \to 0 \quad \text{as } n \to \infty, \tag{3.2}$$

(iii) and the mapping $a \mapsto \sum_{y\in S}\mathbb{P}_{xy}(a)h(y)$, $a \in A(x)$ is continuous.

Under these requirements, using that the reward function R is continuous, it can be shown that the following conclusions (a) and (b) hold (Cavazos-Cadena and Montes-de-Oca (2012)).

(a) The superior and inferior limit average criteria render the same optimal value function, and the optimal average cost is equal to g:

$$J_-^*(x) = J^*(x) = g, \qquad x \in S.$$

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

(b) There exists a stationary policy $f \in \mathbb{F}$ satisfying

$$g + h(x) = R(x, f(x)) + \sum_{y \in S} \mathbb{P}_{xy}(f(x))h(y), \qquad x \in S, \tag{3.3}$$

and such a policy is optimal with respect to the limsup and liminf average reward criteria, that is,

$$J^*(x) = J(x; f) = g = J_-(x; f) = J_-^*(x), \qquad x \in S. \tag{3.4}$$

The existence of a pair $(g, h(\cdot))$ satisfying (3.1) as well as (3.2) requires some communication and stability condition (see, for instance, Thomas (1980) or Cavazos-Cadena (1988), (1989)), and a general requirement in this direction is presented below. Throughout this paper the remainder $z \in S$ is a fixed state, whereas T stands for the first return time to state z, i.e.

$$T := \min\{n > 0 \mid X_n = z\}, \tag{3.5}$$

where, by convention, the minimum of the empty set is ∞. The following idea was introduced in Hordijk (1974) and several alternative formulations were analyzed in Cavazos-Cadena and Hernández-Lerma (1992).

Definition 3.1. Let $D \in \mathcal{C}(\mathbb{K})$ and $\ell\colon S \to [1, \infty)$ be given functions. The mapping ℓ is a Lyapunov function for D if the following conditions occur:

(i) $1 + |D(x, a)| + \sum_{y \neq z} \mathbb{P}_{xy}(a)\ell(y) \leq \ell(x)$ for all $(x, a) \in \mathbb{K}$,

(ii) for each $x \in S$, the mapping $a \mapsto \sum_y \mathbb{P}_{xy}(a)\ell(y)$ is continuous in $a \in A(x)$,

(iii) for every $f \in \mathbb{F}$ and $x \in S$, the convergence $\lim_{n\to\infty} \mathbb{E}_x^f\{\ell(X_n)\,\mathbf{1}_{\{T>n\}}\} = 0$ holds, where $\mathbf{1}_{\{\cdot\}}$ is the indicator function.

Using condition (i) in this definition it is not difficult to see that, regardless of the initial state, the inequality $\mathbb{E}_x^\pi\{\sum_{t=0}^{T-1}[1 + D(X_t, A_t)]\} \leq \ell(x)$ holds for every policy π; in particular,

$$\mathbb{P}_x^\pi\{T < \infty\} = 1, \qquad x \in S,\ \pi \in \mathcal{P}. \tag{3.6}$$

As is shown in the following result by Hordijk (1974), the existence of a Lyapunov function for the reward function R has important implications for the analysis of the average criteria in (2.2) and (2.3).

Lemma 3.1. *Suppose that Assumption 2.1 holds, and that the reward function $R \in \mathcal{C}(\mathbb{K})$ has a Lyapunov function ℓ. In this context,*

(i) *there exists a unique pair $(g_R, h_R(\cdot)) \equiv (g, h(\cdot)) \in \mathbb{R} \times \mathcal{C}(S)$ such that*

(a) *$h(z) = 0$ and $|h(\cdot)| \leq \alpha\ell(\cdot)$ for some constant $\alpha > 0$, and*

(b) *the optimality equation (3.1) corresponding to R is satisfied by $(g, h(\cdot))$.*

(ii) *The following conclusions are valid:*

(a) *$|g| \leq \ell(z)$, and $|h(x)| \leq (1 + \ell(z))\ell(x)$ for all $x \in S$,*

(b) *the relations in (3.2) are satisfied by $h_R \equiv h$, so that $g = J^*(\cdot)$,*

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

(c) *there exists* $f \in \mathbb{F}$ *such that (3.3) holds. Such a policy is optimal and satisfies (3.4), so that*

$$\lim_{k\to\infty} \frac{1}{k}\mathbb{E}_x^f\left\{\sum_{t=0}^{k-1} R(X_t, A_t)\right\} = g, \qquad x \in S, \tag{3.7}$$

and

$$\limsup_{k\to\infty} \frac{1}{k}\mathbb{E}_x^\pi\left\{\sum_{t=0}^{k-1} R(X_t, A_t)\right\} \le g, \qquad x \in S,\ \pi \in \mathcal{P}. \tag{3.8}$$

(iii) *The function h is given by the following expression:*

$$h(x) = \sup_{\pi\in\mathcal{P}} \mathbb{E}_x^\pi\left\{\sum_{t=0}^{T-1}(R_t - g)\right\} = \mathbb{E}_x^f\left\{\sum_{t=0}^{T-1}(R_t - g)\right\}, \qquad x \in S. \tag{3.9}$$

A proof of this result can be essentially found in Hordijk (1974); also, see Cavazos-Cadena and Fernández-Gaucherand (1995) for a proof of (3.9). The remainder of this paper is dedicated to studying the validity of the sample-path versions of (3.7) and (3.8), which are obtained by replacing the expected averages by observed averages along sample trajectories. The following simple properties of Lyapunov functions will be useful.

Remark 3.1. Let $D_1, D_2 \in \mathcal{C}(\mathbb{K})$ be such that D_1 and D_2 have Lyapunov functions ℓ_1 and ℓ_2, respectively. In this case (Cavazos-Cadena and Montes-de-Oca (2012)),

(i) if $D \in \mathcal{C}(\mathbb{K})$ satisfies that $|D| \le |D_1|$ then ℓ_1 is a Lyapunov function for D,

(ii) for $a_0, a_1 \in \mathbb{R}$, $(\max\{|a_0|, |a_1|\} + 1)\ell_1$ is a Lyapunov function for $a_0 + a_1 D_1$.

(iii) for $a_1, a_2 \in \mathbb{R}$, the mapping $\max\{|a_1|, 1\}|\ell_1 + \max\{|a_2|, 1\}\ell_2$ is a Lyapunov function for $a_1 D_1 + a_2 D_2$.

To conclude this section, sufficient conditions are given to ensure that the functional part of a solution of the optimality equation is bounded above or below.

Lemma 3.2. *Under Assumption 2.1, assume that* $R \in \mathcal{C}(\mathbb{K})$ *has a Lyapunov function* ℓ *and let* $(g, h(\cdot))$ *be the solution of the optimality equation (3.1) as in Lemma 3.1(i).*

(i) *Suppose that there exists a finite set* $F \subset S$ *such that*

$$\inf_{a\in A(x)} R(x, a) \ge g, \qquad x \in S \setminus F. \tag{3.10}$$

In this case, there exists a constant $b \in \mathbb{R}$ *such that* $h(\cdot) \ge b$.

(ii) *If for a finite set* $F \subset S$ *the property*

$$\sup_{a\in A(x)} R(x, a) \le g \qquad x \in S \setminus F$$

holds then there exists a constant $b \in \mathbb{R}$ *such that* $h(\cdot) \le b$.

Proof. Let $F \subset S$ be a finite set such that (3.10) holds and, without loss of generality, assume that $z \in F$. It will be proved that for every $x \in S$,

$$h(x) \ge b := -(1 + \ell(z)) \max_{y\in F} \ell(y), \tag{3.11}$$

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

an inequality that, using the bound $|h(\cdot)| \le (1 + \ell(z))\ell(\cdot)$ in Lemma 3.1(ii), is valid if $x \in F$. To establish (3.11) when x is not an element of F, let T_F be the time of the first visit to F, i.e.

$$T_F := \min\{t \ge 0 \mid X_t \in F\},$$

where the minimum of the empty set is ∞; note that the inclusion $z \in F$ implies that

$$T_F \le T \tag{3.12}$$

(see (3.5)) and then with probability 1, T_F is finite regardless of the initial state and the policy used to drive the system, by (3.6). Now, let $f \in \mathbb{F}$ be as in (3.3), select $x \in S \setminus F$ and note that in this case $\mathbb{P}_x^f\{T_F > 0\} = 1$; since $X_t \notin F$ for $t < T_F$, it follows that (3.10) yields

$$R(X_t, A_t) \ge g, \qquad 0 \le t < T_F, \quad \mathbb{P}_x^f\text{-a.s.} \tag{3.13}$$

so that

$$\mathbb{E}_x^f\left\{\mathbf{1}_{\{T_F=T\}} \sum_{t=0}^{T-1}[R(X_t, A_t) - g]\right\} = \mathbb{E}_x^f\left\{\mathbf{1}_{\{T_F=T\}} \sum_{t=0}^{T_F-1}[R(X_t, A_t) - g]\right\} \ge 0. \tag{3.14}$$

Now, let k be a positive integer and note that

$$[T_F = k < T] \in \mathcal{P}_k = \sigma(X_0, A_0, \ldots, X_{k-1}, A_{k-1}, X_k);$$

thus, an application of the Markov property yields

$$\begin{aligned}
&\mathbb{E}_x^f\left\{\mathbf{1}_{\{T_F=k<T\}} \sum_{t=0}^{T-1}[R(X_t, A_t) - g] \;\middle|\; \mathcal{P}_k\right\} \\
&\quad = \mathbf{1}_{\{T_F=k<T\}} \sum_{t=0}^{k-1}[R(X_t, f(X_t)) - g] + \mathbf{1}_{\{T_F=k<T\}}\, \mathbb{E}_{X_k}^f\left\{\sum_{t=0}^{T-1}[R(X_t, A_t) - g]\right\} \\
&\quad \ge \mathbf{1}_{\{T_F=k<T\}}\, \mathbb{E}_{X_k}^f\left\{\sum_{t=0}^{T-1}[R(X_t, A_t) - g]\right\} \\
&\quad = \mathbf{1}_{\{T_F=k<T\}}\, h(X_k),
\end{aligned}$$

where the inequality is due to (3.13), and the second equality in (3.9) was used in the last step. Thus, since $X_k \in F$ when $T_F = k$, using the bound for $h(\cdot)$ in part (a) of (3.1), it follows that

$$\mathbb{E}_x^f\left\{\mathbf{1}_{\{T_F=k<T\}} \sum_{t=0}^{T-1}[R(X_t, A_t) - g] \;\middle|\; \mathcal{P}_k\right\} \ge -\mathbf{1}_{\{T_F=k<T\}}(1 + \ell(z)) \max_{y\in F} \ell(y),$$

and since k is arbitrary, this yields

$$\mathbb{E}_x^f\left\{\mathbf{1}_{\{T_F<T\}} \sum_{t=0}^{T-1}[R(X_t, A_t) - g]\right\} \ge -\mathbb{P}_x^f\{T_F < T\}(1 + \ell(z)) \max_{y\in F} \ell(y).$$

Combining this inequality with (3.12) and (3.14), it follows that (3.11) is also valid when $x \in S \setminus F$. This completes the proof of part (i), while assertion (ii) can be obtained along similar lines.

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

4. Sample-path optimality

In this section we formally introduce the idea of the (strong) sample-path average optimal policy and state the main existence result of this paper.

Definition 4.1. A policy $\pi^* \in \mathcal{P}$ is sample-path average optimal with optimal value $g^* \in \mathbb{R}$ if the following conditions hold:

(i) for each state $x \in S$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} R(X_t, A_t) = g^*, \qquad \mathbb{P}_x^{\pi^*}\text{-a.s.}$$

(ii) for every $\pi \in \mathcal{P}$ and $x \in S$,

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} R(X_t, A_t) \leq g^*, \qquad \mathbb{P}_x^{\pi}\text{-a.s.}$$

The existence of sample-path optimal stationary policies will be derived under the following condition.

Assumption 4.1. *The reward function $R \in \mathcal{C}(\mathbb{K})$ has a Lyapunov function ℓ such that, under the action of any policy and regardless of the initial state, the (superior limit) average reward corresponding to ℓ^2 is finite, that is,*

$$\limsup_{n\to\infty} \frac{1}{n+1} \mathbb{E}_x^{\pi} \left\{ \sum_{k=0}^{n} \ell^2(X_k) \right\} < \infty, \qquad x \in S,\ \pi \in \mathcal{P}. \tag{4.1}$$

Theorem 4.1. *Suppose that Assumptions 2.1 and 4.1 hold, and let $(g, h(\cdot))$ be the solution of the optimality equation guaranteed by Lemma 3.1. In this case, if the stationary policy f satisfies (3.3) then f is sample-path average optimal with optimal value g. More explicitly, for each $x \in S$ and $\pi \in \mathcal{P}$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} R(X_t, A_t) = g, \qquad \mathbb{P}_x^{f}\textit{-a.s.}$$

and

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} R(X_t, A_t) \leq g, \qquad P_x^{\pi}\textit{-a.s.} \tag{4.2}$$

Remark 4.1. (i) The above result is related to Theorem 4.1 in Cavazos-Cadena and Fernández-Gaucherand (1995), where it was proved that the existence of a Lyapunov function for R implies, without any additional requirement, that

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} R(X_t, A_t) \leq g, \qquad \mathbb{P}_x^{\pi}\text{-a.s.}$$

for each $\pi \in \mathcal{P}$ and $x \in S$, and that the equality holds with limit instead of inferior limit whenever $\pi = f \in \mathbb{F}$ satisfies (3.3). In Theorem 4.1 above, the existence of a Lyapunov function for the reward function R is complemented with the requirement (4.1), and in that context the conclusion (4.2) is obtained, which is stronger than the one in the above display.

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

(ii) Theorem 4.1 above generalizes a result of Cavazos-Cadena and Montes-de-Oca (2012), where the sample path average optimality of the policy f in (3.3) was obtained when the condition (4.1) is replaced by the following requirement:

$$\text{for some } \beta > 2, \text{ the function } \ell^\beta \text{ has a Lyapunov function.}$$

This condition ensures that the optimal reward function associated to ℓ^β is finite, and then, since $\ell \geq 1$ and $\beta > 2$ the (superior limit) optimal average index corresponding to ℓ^2 is finite; thus, the above displayed requirement is stronger that (4.1) (costs, instead of rewards, were considered in the aforementioned paper).

(iii) A class of queueing system satisfying the conditions of Theorem 4.1 can be constructed along the lines of Cavazos-Cadena and Montes-de-Oca (2012).

The rather technical proof of Theorem 4.1 has been divided into four steps. The first two steps, contained in Sections 5 and 6, involve the optimality equation (3.1). Next, the third step concerns a direct analysis of the sample trajectories of the state-action process $\{(X_t, A_t)\}$ and is presented in Section 7, whereas the final step combines the tools in Sections 5–7 and is presented in Section 8, just before the proof of the main result.

5. A continuity property

This section presents the first auxiliary result that will be used in the proof of Theorem 4.1. The main objective is to establish a sort of dominated convergence theorem, which can be described as follows. Suppose that a sequence $\{D_n\} \subset \mathcal{C}(\mathbb{K})$ is such that the functions D_n have a common Lyapunov function (the dominance condition), and let $\{g_{D_n}\}$ be the corresponding sequence of optimal average rewards. In this case, if the sequence $\{D_n\}$ converges in an appropriate sense then $\{g_{D_n}\}$ converges to the optimal average reward associated to the limit function.

Throughout the remainder of the paper $\{S_k\}$ is a sequence of nonempty and finite subsets of S such that

$$S_k \subset S_{k+1},\ k = 1, 2, 3, \ldots, \quad \text{and} \quad \bigcup_{k=1}^{\infty} S_k = S. \tag{5.1}$$

Theorem 5.1. *Let the sequence $\{D_n\} \subset \mathcal{C}(\mathbb{K})$ and $D \in \mathcal{C}(\mathbb{K})$ be such that the following conditions are satisfied:*

(a) *for each $x \in S$, $\lim_{n\to\infty} \sup_{a\in A(x)} |D_n(x, a) - D(x, a)| = 0$,*

(b) *there exists $\ell\colon S \to [1, \infty)$ such that for every $n \in \mathbb{N}$, ℓ is a Lyapunov function for D_n.*

In this context

$$\lim_{n\to\infty} g_{D_n} = g_D, \tag{5.2}$$

where g_{D_n} and g_D are the optimal average rewards corresponding to the functions D_n and D, respectively.

Proof. Since ℓ is a Lyapunov function for each mapping D_n, from condition (a) it is not difficult to see that ℓ is also a Lyapunov function for D. Now, let $(g_{D_n}, h_{D_n}) \equiv (g_n, h_n(\cdot))$ be the unique pair solving the optimality equation corresponding to the reward function D_n as

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

described in Lemma 3.1, so that

$$g_n + h_n(x) = \sup_{a\in A(x)}\left[D_n(x,a) + \sum_{y\in S}\mathbb{P}_{xy}(a)h_n(y)\right], \qquad x \in S, \tag{5.3}$$

as well as

$$|g_n| \le \ell(z),\ h_n(z) = 0, \quad \text{and} \quad |h_n(x)| \in [0, (1+\ell(z))\ell(x)],\ x \in S. \tag{5.4}$$

To establish (5.2), let $g \in [-\ell(z), \ell(z)]$ be an arbitrary limit point of the sequence $\{g_n\}$, and select an increasing sequence of positive integers $\{n_k\}$ such that

$$g_{n_k} \to g \quad \text{as } k \to \infty. \tag{5.5}$$

Next, using the second inequality in (5.4), taking a subsequence (if necessary), without loss of generality assume that

$$\lim_{k\to\infty} h_{n_k}(x) = h(x) \in [-(1+\ell(z))\ell(x), (1+\ell(z))\ell(x)], \qquad x \in S, \tag{5.6}$$

so that the pair (g,h) satisfies relations similar to those in (5.4). Now, let $x \in S$ be arbitrary and observe that

(i) for each integer k, the finiteness of S_k and Assumption 2.1 together yield that the mapping $a \mapsto \sum_{y\in S_k}\mathbb{P}_{xy}(a)\ell(y)$ is continuous in $a \in A(x)$ and, because of the positivity of ℓ,

(ii) $\sum_{y\in S_k}\mathbb{P}_{xy}(a)\ell(y) \nearrow \sum_{y\in S}\mathbb{P}_{xy}(a)\ell(y)$ as $k \nearrow \infty$ for every $a \in A(x)$.

Recalling that $\sum_{y\in S}\mathbb{P}_{xy}(a)\ell(y)$ is a continuous function of $a \in A(x)$, by Definition 3.1(ii), and that the action set $A(x)$ is compact, by Assumption 2.1, then Dini's theorem (Ash (1972)) yields

$$\sup_{a\in A(x)}\left|\sum_{y\in S_k}\mathbb{P}_{xy}(a)\ell(y) - \sum_{y\in S}\mathbb{P}_{xy}(a)\ell(y)\right| = \sup_{a\in A(x)}\left|\sum_{y\in S\setminus S_k}\mathbb{P}_{xy}(a)\ell(y)\right| \to 0 \quad \text{as } k\to\infty. \tag{5.7}$$

Note that for all positive integers n and k,

$$\begin{aligned}
&\sup_{a\in A(x)}\left|\sum_{y\in S}\mathbb{P}_{xy}(a)h_n(y) - \sum_{y\in S}\mathbb{P}_{xy}(a)h(y)\right| \\
&\quad\le \sup_{a\in A(x)}\left|\sum_{y\in S_k}\mathbb{P}_{xy}(a)(h_n(y)-h(y))\right| + \sup_{a\in A(x)}\left|\sum_{y\in S\setminus S_k}\mathbb{P}_{xy}(a)(h_n(y)-h(y))\right| \\
&\quad\le \sum_{y\in S_k}\left|h_n(y)-h(y)\right| + 2(1+\ell(z))\sup_{a\in A(x)}\left|\sum_{y\in S\setminus S_k}\mathbb{P}_{xy}(a)\ell(y)\right|,
\end{aligned}$$

where the inclusions in (5.4) and (5.6) were used to set the second inequality; since the set S_k is finite, this last display and (5.6) lead to

$$\limsup_{n\to\infty}\sup_{a\in A(x)}\left|\sum_{y\in S}\mathbb{P}_{xy}(a)h_n(y) - \sum_{y\in S}\mathbb{P}_{xy}(a)h(y)\right| \le 2(1+\ell(z))\sup_{a\in A(x)}\left|\sum_{y\in S\setminus S_k}\mathbb{P}_{xy}(a)\ell(y)\right|,$$

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

and letting k increase to ∞, (5.7) implies that

$$\sup_{a\in A(x)}\left|\sum_{y\in S}\mathbb{P}_{xy}(a)h_n(y)-\sum_{y\in S}\mathbb{P}_{xy}(a)h(y)\right|\to 0\quad\text{as } n\to\infty. \tag{5.8}$$

Next, observe that

$$\begin{aligned}&\left|\sup_{a\in A(x)}\left[D_n(x,a)+\sum_{y\in S}\mathbb{P}_{xy}(a)h_n(y)\right]-\sup_{a\in A(x)}\left[D(x,a)+\sum_{y\in S}\mathbb{P}_{xy}(a)h(y)\right]\right|\\ &\quad\le\sup_{a\in A(x)}\left|\left[D_n(x,a)+\sum_{y\in S}\mathbb{P}_{xy}(a)h_n(y)\right]-\left[D(x,a)+\sum_{y\in S}\mathbb{P}_{xy}(a)h(y)\right]\right|\\ &\quad\le\sup_{a\in A(x)}|D_n(x,a)-D(x,a)|+\sup_{a\in A(x)}\left|\sum_{y\in S}\mathbb{P}_{xy}(a)h_n(y)-\sum_{y\in S}\mathbb{P}_{xy}(a)h(y)\right|.\end{aligned}$$

combining this fact with (5.8) and condition (a) in the statement of the theorem, it follows that, as $n\to\infty$,

$$\sup_{a\in A(x)}\left[D_n(x,a)+\sum_{y\in S}\mathbb{P}_{xy}(a)h_n(y)\right]\to\sup_{a\in A(x)}\left[D(x,a)+\sum_{y\in S}\mathbb{P}_{xy}(a)h(y)\right],$$

and then taking the limit as n goes to ∞ in both sides of (5.3), together (5.5) and (5.6) imply that

$$g+h(x)=\sup_{a\in A(x)}\left[D(x,a)+\sum_{y\in S}\mathbb{P}_{xy}(a)(y)\right],\qquad x\in S.$$

Therefore, by Lemma 3.1, g coincides with the optimal average cost g_D corresponding to the reward function D; the conclusion (5.2) follows since g is an arbitrary limit point of $\{g_n\}\equiv\{g_{D_n}\}$.

6. The discrepancy function

This section contains the second auxiliary result that will be used to establish Theorem 4.1. The main conclusions stated below involve the following idea.

Definition 6.1. Suppose that Assumption 2.1 holds and that $R\in\mathcal{C}(\mathbb{K})$ has a Lyapunov function ℓ. In this context, the discrepancy function $\Phi_R\colon\mathbb{K}\to\mathbb{R}$ corresponding to R is defined by

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

where

$$(g,h(\cdot))\equiv(g_R,h_R(\cdot))$$

is the unique solution of the optimality equation corresponding to R as described in Lemma 3.1.

Recalling that the function $h(\cdot)$ satisfies the requirements in (3.2), the continuity of R yields $\Phi_R\in\mathcal{C}(\mathbb{K})$; also, observe that the optimality equation (3.1) implies that

$$\Phi_R(x,a)\ge 0,\qquad(x,a)\in\mathbb{K}$$

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

and that (6.1) can be equivalently written as

$$g + h(x) = R(x,a) + \Phi_R(x,a) + \sum_{y\in S} \mathbb{P}_{xy}(a)h(y), \qquad (x,a) \in \mathbb{K}. \tag{6.2}$$

The simple properties below will be useful.

Lemma 6.1. *The following assertions hold.*

(i) *The discrepancy function Φ_R has a Lyapunov function $\tilde{\ell}_0$ satisfying $\tilde{\ell}_0(\cdot) \le \tilde{c}_0\ell(\cdot)$, where $\tilde{c}_0 = 2(1+\ell(z))$.*

(ii) *The mapping $R + \Phi_R$ has a Lyapunov function ℓ_0 satisfying*

$$\ell \le \ell_0 \le c_0\ell, \tag{6.3}$$

where $c_0 = 1 + 2(1+\ell(z))$.

(iii) *The solution of the optimality equation corresponding to $R + \Phi_R$ as described in Lemma 3.1 is the pair $(g, h(\cdot)) \equiv (g_R, h_R(\cdot))$, the same solution of the optimality equation associated to R.*

Proof. (i) Using the relations $|g| \le \ell(z)$ and $h(z) = 0$ in parts (i) and (ii) of Lemma 3.1, from (6.2) it follows that for every $(x,a) \in \mathbb{K}$,

$$h(x) \ge R(x,a) + \Phi_R(x,a) - \ell(z) + \sum_{y\in S\setminus\{z\}} \mathbb{P}_{xy}(a)h(y).$$

On the other hand, since ℓ is a Lyapunov function for R, the inequality in Definition 3.1(i) yields

$$(1+\ell(z))\ell(x) \ge |R(x,a)| + 1 + \ell(z) + \sum_{y\in S\setminus\{z\}} \mathbb{P}_{xy}(a)(1+\ell(z))\ell(y).$$

Adding the last two inequalities it follows that

$$\tilde{\ell}_0(x) \ge \Phi_R(x,a) + 1 + \sum_{y\in S\setminus\{z\}} \mathbb{P}_{xy}(a)\tilde{\ell}_0(y), \qquad (x,a) \in \mathbb{K}, \tag{6.4}$$

where

$$\tilde{\ell}_0(y) := h(y) + (1+\ell(z))\ell(y), \qquad y \in S.$$

Observe that the second inequality in Lemma 3.1(ii) yields $\tilde{\ell}_0(\cdot) \le \tilde{c}_0\ell(\cdot)$ where $\tilde{c}_0 = 2(1+\ell(z))$, as well as $\tilde{\ell}_0 \ge 0$; since Φ_R is nonnegative, (6.4) immediately implies that $\tilde{\ell}_0(x) \ge 1$, and that $\tilde{\ell}_0$ satisfies the first property characterizing a Lyapunov function for Φ_R. Concerning the verification of the second and third requirements in Definition 3.1 for the function $\tilde{\ell}_0$, via the inequality $\tilde{\ell}_0 \le \tilde{c}_0\ell$, they follow from the corresponding properties of ℓ.

(ii) By Remark 3.1, the mapping $\ell_0 = \ell + \tilde{\ell}_0$ is a Lyapunov function for $R + \Phi_R$; the conclusion follows since, using part (i), $\ell \le \ell_0 \le \ell + \tilde{c}_0\ell = (1+\tilde{c}_0)\ell$.

(iii) Note that, by Lemma 3.1(i), the pair $(g, h(\cdot)) \equiv (g_R, h_R(\cdot))$ satisfies that:

(a) $h(z) = 0$ and $|h(\cdot)| \le \alpha\ell(\cdot) \le \alpha\ell_0$ for some constant α, whereas (6.2) implies that

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

(b)
$$g + h(x) = \sup_{a \in A(x)} \left[R(x,a) + \Phi_R(x,a) + \sum_{y \in S} \mathbb{P}_{xy}(a) h(y) \right], \qquad x \in S,$$

that is, the pair $(g, h(\cdot)) \equiv (g_R, h_R(\cdot))$ satisfies the optimality equation corresponding to $R + \Phi_R$; the conclusion follows, since a pair $(g, h(\cdot))$ satisfying the above properties (a) and (b) is uniquely determined, by Lemma 3.1(i).

Next, the finite sets S_k in (5.1) will be used to truncate the mapping $R + \Phi_R$.

Definition 6.2. Let $R \in \mathcal{C}(\mathbb{K})$ be such that R has a Lyapunov function ℓ, and let Φ_R be the discrepancy function introduced in Definition 6.1.

(i) For each positive integer k the mapping $\Delta_k \in \mathcal{C}(\mathbb{K})$ is given by

$$\Delta_k(x,a) := \begin{cases} 0, & x \in S_k,\ a \in A(x), \\ |g| + 1 + |R(x,a) + \Phi_R(x,a)|, & x \in S \setminus S_k,\ a \in A(x). \end{cases}$$

(ii) For $u = 1, -1$ and $k = 1, 2, 3, \ldots$, $\Delta_{k,u} : \mathbb{K} \to \mathbb{R}$ is defined as follows:

$$\Delta_{k,u}(x,a) := u[\Delta_k(x,a) + \Phi_{\Delta_k}] + R(x,a) + \Phi_R(x,a), \qquad (x,a) \in \mathbb{K},$$

where Φ_{Δ_k} is the discrepancy function corresponding to Δ_k.

Note that Lemma 6.1 and Remark 3.1 yield that Δ_k admits a Lyapunov function, and then the functions $\Delta_{k,u}$ are well-defined. The main result of this section is concerned with properties of the solutions of the optimality equations corresponding to the functions Δ_k and $\Delta_{k,u}$.

Theorem 6.1. *Suppose that Assumption 2.1 holds, and that $R \in \mathcal{C}(\mathbb{K})$ has a Lyapunov function ℓ. In this context, the following assertions hold.*

(i) *There exists a mapping $\ell^* : S \to [1, \infty)$ such that*
$$\ell \le \ell^*(\cdot) \le c^* \ell(\cdot) \tag{6.5}$$
for some $c^* > 0$ and for each positive integer k, $\ell^*(\cdot)$ is a Lyapunov function for Δ_k.

(ii) *If $(g_k, h_k(\cdot))$ is the unique solution of the optimality equation corresponding to the reward function Δ_k as in Lemma 3.1 then*
$$\lim_{k \to \infty} g_k = 0, \tag{6.6}$$
and there exists a positive integer N and constants $b_k \in \mathbb{R}$ such that
$$h_k(\cdot) \ge b_k \quad \text{for } k \ge N. \tag{6.7}$$

(iii) *There exists $\tilde{\ell} : S \to [1, \infty)$ such that $\ell \le \tilde{\ell}(\cdot) \le \tilde{c}\ell(\cdot)$ for some $\tilde{c} > 0$ and for each $u = 1, -1$ and $k = 1, 2, 3, \ldots$, the mapping $\tilde{\ell}(\cdot)$ is a Lyapunov function for $\Delta_{k,u}$.*

(iv) *Let $(g_{ku}, h_{ku}(\cdot))$ be the unique solution of the optimality equation corresponding to the reward function $\Delta_{k,u}$ as in Lemma 3.1. With this notation, the following assertions hold:*
$$|h_{ku}(\cdot)| \le \beta \ell(\cdot) \quad \text{for some constant } \beta, \tag{6.8}$$

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

and

$$(g_{ku}, h_{ku}(\cdot)) = u \cdot (g_k, h_k(\cdot)) + (g, h(\cdot)), \qquad u = -1, 1,\ k = 1, 2, 3, \ldots, \tag{6.9}$$

where $(g, h(\cdot))$ *is the solution of the optimality equation corresponding to* R *as in Lemma 3.1. Also, there exists a positive integer* N *and constants* $b_{k,u} \in \mathbb{R}$ *such that if* $k \geq N$ *then*

$$h_{k,u}(\cdot) \geq b_{k,u} \quad \text{if } u = 1, \qquad h_{k,u}(\cdot) \leq b_{k,u} \quad \text{if } u = -1. \tag{6.10}$$

Proof. (i) By Remark 3.1(i), if ℓ_0 is the Lyapunov function for $R + \Phi_R$ in Lemma 6.1(ii) then

$$\ell^* := (|g| + 2)\ell_0$$

is a Lyapunov function for $|g| + 1 + |R + \Phi_R|$; since $|\Delta_k| \leq |g| + 1 + |R + \Phi_R|$, Remark 3.1(i) yields ℓ^* is a Lyapunov function for every mapping Δ_k, and the conclusion follows, since (6.3) implies that $\ell \leq \ell^* \leq c^*\ell$ where $c^* = c_0(|g| + 2)$.

(ii) Recalling that the sets S_k increase to the state-space S, the definition of functions Δ_k yields for each $x \in S$, $\sup_{a \in A(x)} |\Delta_k(x, a) - 0| \to 0$ as $k \to \infty$. Since ℓ^* is a Lyapunov function for each mapping Δ_k, Theorem 5.1 implies that $\lim_{k\to\infty} g_k = 0$. Now select an integer N such that

$$|g_k| \leq |g| + \tfrac{1}{2}, \qquad k \geq N, \tag{6.11}$$

and note that

$$\inf_{a \in A(x)} \Delta_k(x, a) \geq |g| + 1 > g_k, \qquad x \in S \setminus S_k,\ k \geq N,$$

so that, since S_k is finite, an application of Lemma 3.2(i) yields for some constant b_k, $h_k(\cdot) \geq b_k$ if $k \geq N$.

(iii) Given a positive integer k note that, using part (i), an application of Lemma 6.1(ii) to the reward function Δ_k yields $\Delta_k + \Phi_{\Delta_k}$ has a Lyapunov function ℓ_0^* satisfying

$$\ell^*(\cdot) \leq \ell_0^*(\cdot) \leq (1 + 2(1 + \ell^*(z)))\ell^*(\cdot),$$

and then

$$\ell \leq \ell_0^*(\cdot) \leq \beta\ell$$

for some positive constant β.

Combining this fact with Lemma 6.1(ii), it follows from Remark 3.1 that $|\Delta_k + \Phi_{\Delta_k}| + |R + \Phi_R|$ has the Lyapunov function $\tilde{\ell} = \ell_0 + \ell_0^*$ which satisfies

$$\ell \leq \tilde{\ell}(\cdot) \leq \tilde{c}\ell$$

for some constant $\tilde{c} > 0$. Observing that $|\Delta_{k,u}| \leq |\Delta_k + \Phi_{\Delta_k}| + |R + \Phi_R|$ for every positive integer k and $u = -1, 1$, Remark 3.1(i) yields $\tilde{\ell}$ is a Lyapunov function for each mapping $\Delta_{k,u}$.

(iv) The definition of the discrepancy function Φ_{Δ_k} corresponding to the reward function Δ_k yields for every positive integer k,

$$g_k + h_k(x) = \Delta_k(x, a) + \Phi_{\Delta_k}(x, a) + \sum_{y \in S} \mathbb{P}_{xy}(a) h_k(y), \qquad (x, a) \in \mathbb{K}, \tag{6.12}$$

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

where

$$h_k(z) = 0 \quad \text{and} \quad |h_k(\cdot)| \le \alpha^* \ell^*(\cdot) \tag{6.13}$$

for some positive constant α^*. Combining (6.12) with (6.2) and Definition 6.2(ii), it follows that for $u = -1, 1$, and $k = 1, 2, 3, \ldots$,

$$ug_k + g + uh_k(x) + h(x) = \Delta_{k,u}(x,a) + \sum_{y \in S} \mathbb{P}_{xy}(a)[uh_k(y) + h(y)], \qquad (x,a) \in \mathbb{K} \tag{6.14}$$

and then

$$ug_k + g + uh_k(x) + h(x) = \sup_{a \in A(x)} \left[\Delta_{k,u}(x,a) + \sum_{y \in S} \mathbb{P}_{xy}(a)[uh_k(y) + h(y)] \right], \qquad x \in S.$$

Thus, the pair $(ug_k + g, uh_k(\cdot) + h(\cdot))$ satisfies the optimality equation corresponding to $\Delta_{k,u}$. On the other hand, using that $h(z) = 0$ and $|h(\cdot)| \le \alpha \ell(\cdot)$ for some $\alpha > 0$, via (6.13) it is not difficult to see that $|uh_k(\cdot) + h(\cdot)|$ is bounded above by a multiple of $\tilde{\ell}$, the Lyapunov function of $\Delta_{k,u}$, and then assertions (6.8) and (6.9) follow from Lemma 3.1(i). To conclude, note that Definition 6.2 yields for $x \in S \setminus S_k$ and $a \in A(x)$,

$$\begin{aligned} \Delta_{k,1}(x,a) &= \Delta_k(x,a) + \Phi_{\Delta_k}(x,a) + R(x,a) + \Phi_R(x,a) \\ &= |g| + 1 + |R(x,a) + \Phi_R(x,a)| + \Phi_{\Delta_k}(x,a) + R(x,a) + \Phi_R(x,a) \\ &\ge |g| + 1 \end{aligned}$$

and

$$\begin{aligned} \Delta_{k,-1}(x,a) &= -[\Delta_k(x,a) + \Phi_{\Delta_k}(x,a)] + R(x,a) + \Phi_R(x,a) \\ &= -|g| - 1 - |R(x,a) + \Phi_R(x,a)| - \Phi_{\Delta_k}(x,a) + R(x,a) + \Phi_R(x,a) \\ &\le -|g| - 1. \end{aligned}$$

Thus, selecting the positive integer N such that (6.11) holds, it follows that

$$\inf_{a \in A(x)} \Delta_{k,1}(x,a) \ge |g| + 1 \ge g + g_k = g_{k,1}, \qquad x \in S \setminus S_k,\ k \ge N,$$

and

$$\sup_{a \in A(x)} \Delta_{k,-1}(x,a) \le -|g| - 1 \le g + g_k = g_{k,-1}, \qquad x \in S \setminus S_k,\ k \ge N.$$

Since S_k is finite, Lemma 3.2 yields that there exists constants $b_{k,u}$ such that (6.10) holds, completing the proof.

7. Innovations

The previous preliminaries are related to the optimality equation (3.1) and rely on the existence of a Lyapunov function for the reward function R. In this section a result involving the behavior of the trajectories of the state-action process $\{(X_t, A_t)\}$ will be established. Throughout, $h\colon S \to \mathbb{R}$ is a given function and it is supposed that

$$\mathbb{E}_x^\pi\{(h(X_n))^2\} < \infty, \qquad x \in S,\ \pi \in \mathcal{P},\ n = 1, 2, 3, \ldots, \tag{7.1}$$

whereas the sigma-field $\mathcal{P}_n$ is given by

$$\mathcal{P}_n := \sigma(X_t, A_t, 0 \le t \le n), \qquad n = 1, 2, 3, \ldots. \tag{7.2}$$

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

Definition 7.1. Let $h\colon S \to \mathbb{R}$ be such that (7.1) holds. The sequence of $\{Y_k, k \geq 1\}$ of innovations associated to h is given by

$$Y_n = h(X_n) - \sum_{y \in S} \mathbb{P}_{X_{n-1},y}(A_{n-1})h(y), \qquad n = 1, 2, 3, \ldots.$$

Note that this specification and (7.2) together yield that Y_n is $\mathcal{P}_n$-measurable, whereas an application of the Markov property immediately implies that for every $x \in S$ and $\pi \in \mathcal{P}$,

$$Y_n = h(X_n) - \mathbb{E}_x^\pi\{h(X_n) \mid \mathcal{P}_{n-1}\}, \qquad \mathbb{P}_x^\pi\text{-a.s.} \tag{7.3}$$

Therefore,

(a) Y_n has null expectation with respect to $\mathbb{P}_x^\pi$ and

(b) Y_n is uncorrelated with the σ-field $\mathcal{P}_{n-1}$, that is,

$$\mathbb{E}_x^\pi\{Y_n W\} = 0 \quad \text{if } W \text{ is } \mathcal{P}_{n-1}\text{-measurable and } Y_n W \text{ is } \mathbb{P}_x^\pi\text{-integrable.} \tag{7.4}$$

Since $\infty > \mathbb{E}_x^\pi\{h(X_n)^2\} \geq \mathbb{E}_x^\pi\{\mathbb{E}_x^\pi\{h(X_n)|\mathcal{P}_{n-1}\})^2\}$, (7.3) implies that $\mathbb{E}_x^\pi\{Y_n^2\} < \infty$ and then $Y_n\mathbb{E}_x^\pi\{h(X_n)|\mathcal{P}_{n-1}\}$ is integrable with respect to each measure $\mathbb{P}_x^\pi$, by the Cauchy–Schwarz inequality; thus, (7.4) leads to $\mathbb{E}_x^\pi\{Y_n\mathbb{E}_x^\pi[h(X_n)|\mathcal{P}_{n-1}]\} = 0$, and combining this relation with $h(X_n) = Y_n + \mathbb{E}_x^\pi\{h(X_n)|\mathcal{P}_{n-1}\}$ it follows that

$$\mathbb{E}_x^\pi\{h(X_n)^2\} \geq \mathbb{E}_x^\pi\{Y_n^2\} + \mathbb{E}_x^\pi\{(\mathbb{E}_x^\pi[h(X_n)|\mathcal{P}_{n-1}])^2\} \geq \mathbb{E}_x^\pi\{Y_n^2\}. \tag{7.5}$$

Now, let n and k be positive integers with $n > k$. In this case (7.5) and the Cauchy–Schwarz inequality together imply that Y_nY_k is always $\mathbb{P}_x^\pi$-integrable, whereas (7.2) and (7.3) yield that Y_k is $\mathcal{P}_{n-1}$-measurable; therefore, by (7.4),

$$\mathbb{E}_x^\pi\{Y_nY_k\} = 0, \qquad n \neq k,\ x \in S,\ \pi \in \mathcal{P},$$

an orthogonality property that leads to the following classical result by Kolmogorov.

Lemma 7.1. *If n and k are two positive integers such that $n > k$ then for every $\alpha > 0$,*

$$\mathbb{P}_x^\pi\left\{\max_{\{r\colon k \leq r \leq n\}}\left|\sum_{t=k}^{r} Y_t\right| \geq \alpha\right\} \leq \frac{1}{\alpha^2}\sum_{t=k}^{n}\mathbb{E}_x^\pi\{Y_t^2\}.$$

This conclusion is established as Theorem 22.4 of Billingsley (1995) for the case in which the Y_ns are independent; however, the same arguments used in this theorem show that the conclusion holds in the context described above. The main result of this section provides sufficient conditions to ensure that the sequence of innovations converges to 0 in the Cèsaro sense.

Theorem 7.1. *If*

$$\limsup_{k\to\infty}\frac{1}{k}\mathbb{E}_x^\pi\left\{\sum_{t=1}^{k} h^2(X_t)\right\} < \infty, \qquad x \in S,\ \pi \in \mathcal{P}, \tag{7.6}$$

then

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} Y_k = 0, \qquad \mathbb{P}_x^\pi\text{-a.s. } x \in S,\ \pi \in \mathcal{P}. \tag{7.7}$$

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

Proof. Let $x \in S$ and $\pi \in \mathcal{P}$ be arbitrary but fixed, and note that (7.6) yields

$$b := \sup_{k\geq 1} \frac{1}{k} \mathbb{E}_x^\pi \left\{ \sum_{t=1}^{k} h^2(X_t) \right\} < \infty,$$

and then, by (7.5),

$$\mathbb{E}_x^\pi \left\{ \sum_{t=1}^{n} Y_t^2 \right\} < nb, \qquad n = 1, 2, 3, \ldots.$$

This fact and Lemma 7.1 together lead to

$$\mathbb{P}_x^\pi \left\{ \max_{\{r\colon k\leq r\leq n\}} \left| \sum_{t=k}^{r} Y_t \right| > \delta \right\} \leq \frac{nb}{\delta^2}, \qquad \delta > 0,\ n, k \in \mathbb{N} \setminus \{0\},\ n > k. \tag{7.8}$$

Now, given that $\varepsilon > 0$, note that this relation with $k = 1$, $n = m^2$, and $\delta = \varepsilon m^2$ yields

$$q_m := \mathbb{P}_x^\pi \left\{ m^{-2} \left| \sum_{t=1}^{m^2} Y_t \right| > \varepsilon \right\} \leq \mathbb{P}_x^\pi \left\{ \max_{\{r\colon 1\leq r\leq m^2\}} \left| \sum_{t=1}^{r} Y_t \right| > m^2\varepsilon \right\} \leq \frac{m^2 b}{\varepsilon^2 m^4},$$

which yields $\sum_{m=1}^{\infty} q_m < \infty$. By the first Borel–Cantelli lemma,

$$\mathbb{P}_x^\pi \left\{ m^{-2} \left| \sum_{t=1}^{m^2} Y_t \right| > \varepsilon \text{ i.o.} \right\} = 0,$$

where i.o. stands for infinitely often. Since $\varepsilon > 0$ is arbitrary, it follows that

$$\lim_{m\to\infty} \frac{1}{m^2} \sum_{t=1}^{m^2} Y_t = 0, \qquad \mathbb{P}_x^\pi\text{-a.s.} \tag{7.9}$$

Next, let m be a positive integer and note that

$$\left[\max_{\{j\colon 0\leq j\leq 2m\}} \left| (m^2 + j)^{-1} \sum_{t=m^2}^{m^2+j} Y_t \right| \geq \varepsilon \right] \subset \left[\max_{\{r\colon m^2\leq r<(m+1)^2\}} \left| \sum_{t=m^2}^{r} Y_t \right| \geq m^2\varepsilon \right]$$

is an inclusion that using (7.8) leads to

$$\begin{aligned} p_m &:= \mathbb{P}_x^\pi \left\{ \max_{\{j\colon 0\leq j\leq 2m\}} \left| (m^2 + j)^{-1} \sum_{t=m^2}^{m^2+j} Y_t \right| \geq \varepsilon \right\} \\ &\leq \mathbb{P}_x^\pi \left\{ \max_{\{r\colon m^2\leq r<(m+1)^2\}} \left| \sum_{t=m^2}^{r} Y_t \right| \geq m^2\varepsilon \right\} \\ &\leq \frac{(m+1)^2 b}{\varepsilon^2 m^4}. \end{aligned}$$

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

Therefore, $\sum_{m=1}^{\infty} p_m < \infty$, and the first Borel–Cantelli lemma yields

$$\lim_{m\to\infty}\left\{\max_{\{j\,:\,0\le j\le 2m\}}\left|(m^2+j)^{-1}\sum_{t=m^2}^{m^2+j}Y_t\right|\right\}=0,\qquad \mathbb{P}_x^{\pi}\text{-a.s.}\tag{7.10}$$

To conclude, let n be a positive integer and let m be the integral part of $\sqrt{n}$, so that $n=m^2+i$ where $0\le i\le 2m$. Observe that for $i>0$,

$$\left|\frac{1}{n}\sum_{t=1}^{n}Y_t\right|\le\frac{m^2}{n}\left|\frac{1}{m^2}\sum_{t=1}^{m^2}Y_t\right|+\frac{1}{m^2+i}\left|\sum_{t=m^2+1}^{m^2+i}Y_t\right|,$$

and then

$$\left|\frac{1}{n}\sum_{t=1}^{n}Y_t\right|\le\left|\frac{1}{m^2}\sum_{t=1}^{m^2}Y_t\right|+\max_{\{j\,:\,0\le j\le 2m\}}\left\{\frac{1}{m^2+j}\left|\sum_{t=m^2+1}^{m^2+j}Y_t\right|\right\},$$

an inequality that is also valid when $i=0$, that is, if $n=m^2$. After taking the limit as n goes to ∞ in both sides of this equation, (7.9) and (7.10) together imply that (7.7) holds.

8. Proof of Theorem 4.1

In this section the proof of Theorem 4.1 is finally presented. The argument relies on the auxiliary tools in Sections 5–7 and, by convenience, the main part of the argument is stated separately in the following result using the notation in Definition 6.1 and Theorem 6.1.

Theorem 8.1. *Suppose that Assumptions 2.1 and 4.1 hold, so that the reward function R has a Lyapunov function ℓ satisfying (4.1). In this context, the following assertions hold.*

(i) *Let the positive integer N be as in (6.7). For every $x\in S$, $\pi\in\mathcal{P}$ and $k>N$,*

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n-1}[\Delta_k(X_t,A_t)+\Phi_{\Delta_k}(X_t,A_t)]\le g_k,\qquad \mathbb{P}_x^{\pi}\text{-a.s.}$$

(ii) *For every initial state $x\in S$ and $\pi\in\mathcal{P}$,*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n-1}[R(X_t,A_t)+\Phi_R(X_t,A_t)]=g,\qquad \mathbb{P}_x^{\pi}\text{-a.s.}\tag{8.1}$$

Proof. (i) Observe that the definition of the discrepancy function associated to the reward function Δ_k yields

$$\Delta_k(x,a)+\Phi_{\Delta_k}(x,a)-g_k=h_k(x)-\sum_{y\in S}\mathbb{P}_{xy}(a)h_k(y),\qquad (x,a)\in\mathbb{K},$$

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

so that for every positive integer t,

$$\begin{aligned}
&\Delta_k(X_{t-1}, A_{t-1}) + \Phi_{\Delta_k}(X_{t-1}, A_{t-1}) - g_k \\
&\quad = h_k(X_{t-1}) - \sum_{y\in S} \mathbb{P}_{X_{t-1}y}(A_{t-1})h_k(y) \\
&\quad = h_k(X_{t-1}) - h_k(X_t) + h_k(X_t) - \sum_{y\in S} \mathbb{P}_{X_{t-1}y}(A_{t-1})h_k(y) \\
&\quad = h_k(X_{t-1}) - h_k(X_t) + Y_{k,t},
\end{aligned}$$

where $\{Y_{k,n}\}_{n=1,2,3,\ldots}$ is the sequence of innovations associated to the function $h_k(\cdot)$. Therefore,

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}[\Delta_k(X_{t-1}, A_{t-1}) + \Phi_{\Delta_k}(X_{t-1}, A_{t-1})] - g_k \\
&\qquad = \frac{h_k(X_0) - h_k(X_n)}{n} + \frac{1}{n}\sum_{t=1}^{n} Y_{k,t} \\
&\qquad \le \frac{h_k(X_0) - b_k}{n} + \frac{1}{n}\sum_{t=1}^{n} Y_{k,t}, \qquad k \ge N,
\end{aligned} \tag{8.2}$$

where the inequality stems from (6.7). Now, recall that $|h_k(\cdot)|$ is bounded above by a positive multiple of the Lyapunov function ℓ^* for Δ_k, and then (6.5) yields $|h_k(\cdot)| \le \beta\ell(\cdot)$ for some constant β. From this point, the property (4.1) implies that the condition (7.6) is satisfied by $h_k(\cdot)$, and an application of Theorem 7.1 yields for every $x \in S$ and $\pi \in \mathcal{P}$,

$$\frac{1}{n}\sum_{t=1}^{n} Y_{k,t} \to 0, \qquad \mathbb{P}_x^\pi\text{-a.s.}$$

Taking the limit superior as n goes to ∞ in (8.2), this equation immediately implies that

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n}[\Delta_k(X_{t-1}, A_{t-1}) + \Phi_{\Delta_k}(X_{t-1}, A_{t-1})], \\
&\qquad \le g_k, \qquad \mathbb{P}_x^\pi\text{-a.s., } x \in S,\ \pi \in \mathcal{P},\ k > N.
\end{aligned}$$

(ii) Combining (6.9) and (6.14), it follows that for every positive integer k and $u = -1, 1$,

$$g_{k,u} + h_{k,u}(x) + h(x) = \Delta_{k,u}(x, a) + \sum_{y\in S} \mathbb{P}_{xy}(a)h_{k,u}(y), \qquad (x, a) \in \mathbb{K}$$

and proceeding as in part (i), this equality leads to

$$\frac{1}{n}\sum_{t=1}^{n} \Delta_{k,u}(X_{t-1}, A_{t-1}) - g_{k,u} = \frac{h_{k,u}(X_0) - h_{k,u}(X_n)}{n} - \frac{1}{n}\sum_{t=1}^{n} Y_{k,u,t}, \tag{8.3}$$

where $\{Y_{k,u,t}\}_{t=1,2,3,\ldots}$ is the sequence of innovations corresponding to the function $h_{k,u}$, which is bounded above by a positive multiple of the Lyapunov function $\tilde{\ell}$ (see (6.8)), and then Theorem 6.1(iii) implies that $|h_{k,u}(\cdot)|| \le c\ell(\cdot)$ for some constant c. Therefore, the property

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

(4.1) of the function ℓ implies that the condition (7.6) is satisfied by $h_{k,u}(\cdot)$, and then an application of Theorem 7.1 yields for every $x \in S$ and $\pi \in \mathcal{P}$

$$\frac{1}{n}\sum_{t=1}^{n} Y_{k,u,t} \to 0 \quad \text{as } n \to \infty,\ \mathbb{P}_x^{\pi}\text{-a.s.} \tag{8.4}$$

Now, let N be the positive integer in Theorem 6.1(iv) and note that (6.10) and (8.3) together imply that for every $k \geq N$ and $n = 1, 2, 3, \ldots$,

$$\frac{1}{n}\sum_{t=1}^{n} \Delta_{k,1}(X_{t-1}, A_{t-1}) - g_{k,1} \leq \frac{h_{k,u}(X_0) - b_{k,1}}{n} - \frac{1}{n}\sum_{t=1}^{n} Y_{k,1,t},$$

and

$$\frac{1}{n}\sum_{t=1}^{n} \Delta_{k,-1}(X_{t-1}, A_{t-1}) - g_{k,-1} \geq \frac{h_{k,-1}(X_0) - b_{k,-1}}{n} - \frac{1}{n}\sum_{t=1}^{n} Y_{k,-1,t};$$

thus, via (8.4), it follows that for every $x \in S$ and $\pi \in \mathcal{P}$,

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \Delta_{k,1}(X_{t-1}, A_{t-1}) \leq g_{k,1}, \qquad \mathbb{P}_x^{\pi}\text{-a.s., } k > N, \tag{8.5}$$

and

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \Delta_{k,-1}(X_{t-1}, A_{t-1}) \geq g_{k,-1}, \qquad \mathbb{P}_x^{\pi}\text{-a.s., } k > N. \tag{8.6}$$

Next, from Definition 6.2 and observing that $\Delta_k \geq 0$, and then, since a discrepancy function is nonnegative,

$$\Delta_{k,1} = \Delta_k + \Phi_{\Delta_k} + R + \Phi_R \geq R + \Phi_R,$$

and

$$\Delta_{k,-1} = -[\Delta_k + \Phi_{\Delta_k}] + R + \Phi_R \leq R + \Phi_R,$$

are relations that when combined with (8.5) and (8.6) yield for every $x \in S, \pi \in \mathcal{P}$, and $k > N$,

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} [R(X_{t-1}, A_{t-1}) + \Phi_R(X_{t-1}, A_{t-1})] \leq g_{k,1}, \qquad \mathbb{P}_x^{\pi}\text{-a.s.} \tag{8.7}$$

and

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} [R(X_{t-1}, A_{t-1}) + \Phi_R(X_{t-1}, A_{t-1})] \geq g_{k,-1}, \qquad \mathbb{P}_x^{\pi}\text{-a.s.} \tag{8.8}$$

Finally, using (6.6) and (6.9), note that $g_{k,u} \to g$ as $k \to \infty$ for $u = -1, 1$, so that, after taking the limit as k goes to ∞, (8.7) and (8.8) lead to

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} [R(X_{t-1}, A_{t-1}) + \Phi_R(X_{t-1}, A_{t-1})] \leq g, \qquad \mathbb{P}_x^{\pi}\text{-a.s. } x \in S,\ \pi \in \mathcal{P},$$

https://doi.org/10.1239/jap/1437658607 Published online by Cambridge University Press

and

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n}[R(X_{t-1}, A_{t-1}) + \Phi_R(X_{t-1}, A_{t-1})] \geq g, \qquad \mathbb{P}_x^\pi\text{-a.s. } x \in S,\ \pi \in \mathcal{P},$$

establishing the desired conclusion.

Proof of Theorem 4.1. Since a discrepancy function is nonnegative, from Theorem 8.1(ii) it follows that

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} R(X_{t-1}, A_{t-1}) \leq g, \qquad \mathbb{P}_x^\pi\text{-a.s. } x \in S,\ \pi \in \mathcal{P}.$$

On the other hand, because the policy $f \in \mathbb{F}$ satisfies (3.3), from Definition 6.1 it follows that $\Phi(x, f(x)) = 0$ for every state x. Thus, using that $A_t = f(X_t)$ when the system is running under the policy f it follows that for every initial state x and $t \in \mathbb{N}$, the equality $\Phi(X_t, A_t) = \Phi(X_t, f(X_t)) = 0$ holds with probability 1 with respect to $\mathbb{P}_x^f$. Therefore, from Theorem 8.1(ii) it follows that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} R(X_{t-1}, A_{t-1}) = g, \qquad \mathbb{P}_x^f\text{-a.s. } x \in S;$$

thus, f is sample-path average optimal in the sense of Definition 4.1.

Remark 8.1. Determining the sample-path average optimal stationary policy $f \in \mathbb{F}$ in (3.3) requires the knowledge of the solution $(g, h(\cdot))$ of the optimality equation (3.1). When such a solution is not available, approximation schemes can be used to obtain a sequence $\{(g_r, h_r(\cdot))\}_{r=0,1,2,\ldots}$ converging to $(g, h(\cdot))$, as well as a sequence $\{f_r\}_{r=0,1,2,\ldots} \subset \mathbb{F}$ with the following property:

$$\lim_{r\to\infty} \Phi_R(x, f_r(x)) = 0, \qquad x \in S;$$

see, for instance, Montes-de-Oca and Hernández-Lerma (1996) and the references therein. Using (8.1), it can be shown that the property in the above equation implies that a Markov policy $\{f_t\}$ is sample-path optimal in the sense of Definition 4.1 (Cavazos-Cadena and Montes-de-Oca (2012)).

Acknowledgements

This work was supported in part by the PSF Organization under Grant Number 012/300/02, and by CONACYT (México) and ASCR (Czech Republic) under Grant Number 171396.

The authors are deeply grateful to the anonymous referee for their careful reading of the original manuscript, constructive criticism, and helpful suggestions to improve the paper.

References

Arapostathis, A. *et al.* (1993). Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM J. Control Optimization* **31,** 282–344.

Ash, R. B. (1972). *Real Analysis and Probability*. Academic Press, New York.

Bäuerle, N. and Rieder, U. (2010). Markov decision processes. *Jahresber. Dtsch. Math.-Ver.* **112,** 217–243.

Bäuerle, N. and Rieder, U. (2011). *Markov Decision Processes with Applications to Finance*. Springer, Heidelberg.

Billingsley, P. (1995). *Probability and Measure*, 3rd edn. John Wiley, New York.

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

Borkar, V. S. (1991). *Topics in Controlled Markov Chains*. Longman Scientific and Technical, Harlow.

Cavazos-Cadena, R. (1988). Necessary and sufficient conditions for a bounded solution to the optimality equation in average reward Markov decision chains. *Systems Control Lett.* **10,** 71–78.

Cavazos-Cadena, R. (1989). Necessary conditions for the optimality equation in average-reward Markov decision processes. *Appl. Math. Optimization* **19,** 97–112.

Cavazos-Cadena, R. and Fernández-Gaucherand, E. (1995). Denumerable controlled Markov chains with average reward criterion: sample path optimality. *Math. Meth. Operat. Res.* **41,** 89–108.

Cavazos-Cadena, R. and Hernández-Lerma, O. (1992). Equivalence of Lyapunov stability criteria in a class of Markov decision processes. *Appl. Math. Optimization* **26,** 113–137.

Cavazos-Cadena, R. and Montes-de-Oca, R. (2012). Sample-path optimality in average Markov decision chains under a double Lyapunov function condition. In *Optimization, Control, and Applications of Stochastic Systems*. Springer, New York, pp. 31–57.

Cavazos-Cadena, R., Montes-de-Oca, R. and Sladký, K. (2014). A counterexample on sample-path optimality in stable Markov decision chains with the average reward criterion. *J. Optimization Theory Appl.* **163,** 674–684.

Dai Pra, P., Di Masi, G. B. and Trivellato, B. (1999). Almost sure optimality and optimality in probability for stochastic control problems over an infinite time horizon. *Ann. Operat. Res.* **88,** 161–171.

Foster, F. G. (1953). On the stochastic matrices associated with certain queueing processes. *Ann. Math. Statist.* **24,** 355–360.

Hernández-Lerma, O. (1989). *Adaptive Markov Control Processes*. Springer, New York.

Hernández-Lerma, O., Vega-Amaya, O. and Carrasco, G. (1999). Sample-path optimality and variance-minimization of average cost Markov control processes. *SIAM J. Control Optimization* **38,** 79–93.

Hordijk, A. (1974). *Dynamic Programming and Markov Potential Theory* (Math. Centre Tracts **51**). Mathematisch Centrum, Amsterdam.

Hunt, F. Y. (2005). Sample path optimality for a Markov optimization problems. *Stoch. Process. Appl.* **115,** 769–779.

Lasserre, J. B. (1999). Sample-path average optimality for Markov control processes. *IEEE Trans. Automatic Control* **44,** 1966–1971.

Montes-de-Oca, R. and Hernández-Lerma, O. (1996). Value iteration in average cost Markov control processes on Borel spaces. *Acta Appl. Math.* **42,** 203–222.

Puterman, M. L. (1994). *Markov Decision Processes*. John Wiley, New York.

Sennott, L. I. (1999). *Stochastic Dynamic Programming and the Control of Queueing Systems*. John Wiley, New York.

Thomas, L. C. (1980). Connectedness conditions for denumerable state Markov decision processes. In *Recent Developments in Markov Decision Processes*. Academic Press, New York, pp. 181–204.

Vega-Amaya, O. (1999). Sample path average optimality of Markov control processes with strictly unbounded cost. *Appl. Math. (Warsaw)* **26,** 363–381.

Zhu, Q. and Guo, X. (2006). Another set of conditions for Markov decision processes with average sample-path costs. *J. Math. Anal. Appl.* **322,** 1199–1214.