Morfismos, Vol. 3, No. 1, 1999, pp. 9–33

A DIRECT APPROACH TO BLACKWELL OPTIMALITY *

ROLANDO CAVAZOS–CADENA [1] JEAN B. LASSERRE [2]

Abstract

This work concerns discrete-time Markov decision processes (MDP's) with denumerable state space and bounded rewards. The main objective is to show that the problem of establishing the existence of Blackwell optimal polices can be approached via well-known techniques in the theory of MDP's endowed with the average reward criterion. The main result can be summarized as follows: Assuming that the Simultaneous Doeblin Condition and mild continuity conditions are satisfied, it is shown that a policy π^* is Blackwell optimal if, and only if, the actions prescribed by π^* maximize the right–hand side of the average reward optimality equations associated to a suitably defined sequence of MDP's. In contrast with the usual approach, this result is obtained by using standard techniques and does not involve Laurent series expansions.

1991 Mathematics Subject Classification: 90C40, 93E20.
Keywords and phrases: Discrete-time Markov decision chains, Simultaneous Doeblin Condition, Blackwell optimality, Average reward criterion, Average reward optimality equation.

1 Introduction

This work concerns Markov decision processes (MDP's) with denumerable state space, discrete time parameter and bounded rewards. When

*Invited Article. This work was supported by Grant No. E120.3336 from the Consejo Nacional de Ciencia y Tecnología (CONACyT), México.

[1]Departamento de Estadística y Cálculo, Universidad Autónoma Agraria Antonio Narro, Buenavista, Saltillo COAH 25315, MÉXICO

[2]Laboratoire d'Automatique et d'Analyse des Systèmes, 7, Avenue du Colonel Roche, 31077 Toulouse Cedex, FRANCE

the decision horizon is infinite, there are two criteria that are widely used for measuring the performance of a control policy, namely, the total expected *discounted* reward and the long-run expected *average* reward criteria. In a certain sense, the former emphasizes the behavior of the policy in the early stages, since, due to the presence of the discount factor, the contribution to the criterion coming from the rewards obtained in large decision epochs decays geometrically to zero. On the other hand, the average criterion depends only on the asymptotic behavior of the rewards, but not on those obtained 'during the first milenium' [9]. To balance this situation, other criteria considering both the early and the asymptotic behavior of a control policy have been introduced as, for instance, strong average optimality and what is presently known as Blackwell optimality; the latter notion was introduced by Blackwell in [1] under the name of 1-optimality and it is the main object of interest in this note. These and other 'sensitive' criteria have been recently considered in the literature; see, for instance, [2, 6, 8–10, 13–15] and the references therein. On the other hand, the Blackwell optimality criterion has been studied—mainly—by using Laurent series expansions for α-discounted rewards around $\alpha = 1$. This approach was firstly used by Veinott [21] and Miller and Veinott [18] for *finite* MDP's and has been extended to more general frameworks; see Dekker and Hordijk [6], Yushkevich [22], and their references. Also, a different way to study Blackwell optimality via functional analysis techniques was introduced in [15].

The main objective of this note is to present a simple approach to establish the existence of Blackwell optimal polices, which is based upon familiar techniques employed in the study of MDP's endowed with the average reward criterion. In fact, the main result in this note, stated below in Theorem 3.1, can be roughly described as follows: A policy π^* is Blackwell optimal if and only if the actions prescribed by π^* maximize the right-hand side of the average reward optimality equations associated to a suitably defined sequence of MDP's. Therefore, the construction of Blackwell optimal polices reduces to solve a sequence of MDP's endowed with the average reward criterion. In addition, a simple condition guaranting the *uniqueness* of a Blackwell optimal policy is presented in Theorem 5.1. These results are obtained by assuming that the decision model satisfies the Simultaneous Doeblin Condition as well as standard continuity requirements. The proofs extend the ideas used in [2] to study strong 1-optimal policies.

The organization of the paper is as follows: In Section 2 the decision model is formally described and the discounted, average and Blackwell optimality criteria are introduced. Next, in Section 3 the main result is stated in the form of Theorem 3.1, which is proved in Section 5 after the necessary preliminaries presented in Section 4. Finally, the paper concludes in Section 6 with some brief comments.

Notation. As usual $\mathbb{R}$ stands for the set of the real numbers while $\mathbb{N} := \{0, 1, 2, \cdots\}$. Given a topological space $\mathbb{K}$, the space $\mathbb{B}(\mathbb{K})$ consists all functions $r : \mathbb{K} \to \mathbb{R}$ which are continuous and bounded, i.e.,

$$\|r\| := \sup_{k \in \mathbb{K}} |r(k)| < \infty.$$

On the other hand, a cartesian product of topological spaces is always endowed with the corresponding product topology.

2 The Model

Let $M := (S, C, \{C(x)\}, r, p)$ be the usual MDP where the *state space* S is a denumerable set endowed with the discrete topology, and the *metric space* C is the control set. For each $x \in S$, $C(x) \subset C$ is the nonempty and compact set of *admissible actions* at state x, while the set of admissible state-action pairs is given by $\mathbb{K} := \{(x, a) | x \in S, a \in C(x)\}$, which is considered as a topological subspace of $S \times C$. On the other hand, $r \in \mathbb{B}(\mathbb{K})$ is the reward function and p is the transition law. The interpretation of the model is as follows: At each time $t \in \mathbb{N}$ the state of the system is observed, say $X_t = x \in S$, and an action $A_t = a \in C(x)$ is chosen. Then, a reward $r(x, a)$ is obtained and, regardless of the previous states and actions, the state of the system at time $t+1$ will be $X_{t+1} = y$ with probability $p_{xy}(a)$, where $\sum_y p_{xy}(a) = 1$; this is the Markov property of the decision process.

Assumption 2.1 *For each $x, y \in S$, the mapping $a \mapsto p_{xy}(a)$, $a \in C(x)$, is continuous.*

Policies. A policy is a (possibly randomized) rule for choosing actions which may depend on the current state and on the record of previous states and actions; see [11, pp. 1–4] for a detailed description. The class of all policies is denoted by $\mathbb{P}$ and, given the initial state $X_0 = x$ and the policy $\pi \in \mathbb{P}$ being used, the distribution of the state-action

process $\{(X_t, A_t)\}$ is uniquely determined; it is denoted by P_x^π, whereas E_x^π stands for the corresponding expectation operator. Next, set $\mathbb{F} := \Pi_{x\in S} C(x)$, that is, $\mathbb{F}$ is the set of all (choice) functions $f : S \to C$ such that $f(x) \in C(x)$ for all $x \in S$. A policy $\pi \in \mathbb{P}$ is stationary if there exists $f \in \mathbb{F}$ such that, when the system is in progress under π, action $f(x)$ is applied when $X_t = x$ regardless of the time $t \in \mathbb{N}$; the class of all stationary policies is naturally identified with $\mathbb{F}$. Finally, observe that under the action of any stationary policy $f \in \mathbb{F}$, the state process $\{X_t\}$ is a Markov chain with stationary transition mechanism [11, 19].

Performance Criteria. Let $x \in S$ and $\pi \in \mathbb{P}$ be arbitrary.

(a) For each $\alpha \in (0, 1)$, the total expected α-discounted reward associated to the reward function r at state x under policy π is defined by

$$V_\alpha(\pi; r; x) := E_x^\pi[\sum_{t=0}^{\infty} \alpha^t r(X_t, A_t)]; \tag{2.1}$$

notice that

$$\|V_\alpha(\pi; r; \cdot)\| \leq \|r\|/(1 - \alpha). \tag{2.2}$$

(b) A policy π^* is *Blackwell optimal at* $x \in S$ if for each $\pi \in \mathbb{P}$ there exists $\alpha(\pi^*, \pi; x) \in (0, 1)$ such that

$$V_\alpha(\pi^*; r; x) - V_\alpha(\pi; r; x) \geq 0 \quad \text{for all} \quad \alpha \in (\alpha(\pi^*, \pi; x), 1). \tag{2.3}$$

The policy π^* is *Blackwell optimal* (BO) if it is Blackwell optimal at each $x \in S$.

Observe that the notion of Blackwell optimal policy is expressed in terms of the behavior of the (expected) α-discounted rewards for α close to 1. On the other hand, it is known ([1–6, 8–11, . . .]) that the limiting behavior of the α-discounted rewards as α increases to 1 is closely related to the average reward criterion (introduced below). Therefore, it is interesting to investigate the possibility of studying Blackwell optimality via the familiar techniques employed in the analysis of the average case. As already mentioned in Section 1, the main objective of this note is to show that BO policies can be determined by finding the optimizing actions in each of the average reward optimality equations ($AROE$'s) associated to an appropiate sequence of MDP's ; this result is stated

precisely as Theorem 3.1 in Section 3. First, the necessary notions and assumptions are introduced.

(c) The (lim sup expected) average reward at state $x \in S$ under policy $\pi \in \mathbb{P}$ corresponding to the reward function $r \in \mathbb{B}(\mathbb{K})$ is given by

$$J(\pi; r; x) := \limsup_{n\to\infty} \frac{1}{n+1} E_x^\pi [\sum_{t=0}^{n} r(X_t, A_t)], \tag{2.4}$$

whereas

$$J(r; x) := \sup_{\pi\in\mathbb{P}} J(\pi; r; x) \tag{2.5}$$

is the optimal average reward at x associated to r. A policy $\tilde{\pi}$ is average optimal (AO) for the reward function r if $J(\tilde{\pi}; r; x) = J(r; x)$ for all $x \in S$.

Assumption 2.2 *There exists $K \in (0, \infty)$ with the following property: For each $f \in \mathbb{F}$ there exists $z(f) \in S$ satisfying*

$$E_x^f [T_{z(f)}] \le K \quad \text{for all} \quad x \in S,$$

where $T_{z(f)} := \min\{t > 0 | X_t = z(f)\}$, and the (usual) convention that the minimum of the empty set is ∞ is enforced.

This assumption is a version of the Simultaneous Doeblin Condition [20] and, under certain conditions [3–5], it is necessary and sufficient for the existence of bounded solutions to the $AROE$ for arbitrary $r \in \mathbb{B}(\mathbb{K})$; see (2.6) below.

Lemma 2.1 *Suppose that Assumptions 2.1 and 2.2 hold. Then, for each $r \in \mathbb{B}(\mathbb{K})$, there exists $g_r \in \mathbb{R}$ and $h_r \in \mathbb{B}(S)$ such that (i)–(iv) below occur:*

(i) $g_r = J(r; x)$ for all $x \in S$.

(ii) $\|h_r\| \le B\|r\|$, *where $B := 2K$ and K is as in Assumption 2.2.*

(iii) g_r *and* h_r *satisfy the AROE corresponding to* $M = (S, C, \{C(x)\}, r, p)$, *i.e.,*

$$g_r - h_r(x) = \sup_{a\in C(x)} [r(x, a) - \sum_y p_{xy}(a) h_r(y)], \quad x \in S. \tag{2.6}$$

(iv) For each $x \in S$ the term within brackets in the right–hand side of (2.6)—considered as a function of $a \in C(x)$—has a maximizer $f(x) \in C(x)$; moreover, the corresponding policy $f \in \mathbb{F}$ is AO for the reward function r.

For a proof see, for instance, [5,11,19].

Remark 2.1 (i) The notation in (2.6) differs slightly from the usual one: $-h_r$ in the previous lemma is usually written as h_r.

(ii) Let $M' = (S, C, \{C'(x)\}, r', p)$ be an MDP such that $C'(x) \subset C(x)$ and $C'(x)$ is nonempty and compact for each $x \in S$. The set $\mathbb{F}' := \Pi_{x\in S} C'(x)$ of stationary policies associated to M' is clearly contained in $\mathbb{F}$, so that Assumption 2.2 is satisfied when $\mathbb{F}$ is replaced by $\mathbb{F}'$. This implies that Lemma 2.1 remains valid for the model M'; of course, it is assumed that $r' \in \mathbb{B}(\mathbb{K}')$, where $\mathbb{K}' := \{(x, a) | x \in S, a \in C'(x)\}$ is a topological subspace of $S \times C$.

The following lemma refers to the ergodic properties of the Markov chains induced by stationary policies; part (iv) will be used in the proof of Theorem 4.1 of Section 4.

Lemma 2.2 *Under Assumptions 2.1 and 2.2, (i)–(iv) below occur.*

(i) For each stationary policy $f \in \mathbb{F}$ the Markov chain induced by f has an invariant distribution q_f , that is, $q_f : S \to [0, 1]$ satisfies

$$q_f(y) = \sum_x q_f(x) p_{xy}(f(x)), \quad y \in S, \quad \text{and} \quad \sum_y q_f(y) = 1.$$

(ii) For each $r \in \mathbb{B}(\mathbb{K}), f \in \mathbb{F}$ and $x \in S$,

$$\lim_{n\to\infty} \frac{1}{n+1} E_x^f [\sum_{t=0}^{n} r(X_t, A_t)] = \sum_y q_f(y) r(y, f(y)) = J(f; r; x).$$

(iii) The collection $\{q_f | f \in \mathbb{F}\}$ is tight, that is, given $\varepsilon > 0$ there exists a finite set $G \subset S$ such that

$$\sum_{y\in S\setminus G} q_f(y) < \varepsilon.$$

(iv) For each finite set $F \subset S$ let I_F be the indicator function of F, i.e., $I_F(x) := 1$ (resp. 0) if $x \in F$ (resp. $x \in S \setminus F$); notice that $I_F \in \mathbb{B}(S)$ and that $\mathbb{B}(S)$ can be (naturally) considered as a subspace of $\mathbb{B}(\mathbb{K})$. Now let $G \subset S$ be a finite set. Then, for each $x \in S$,

$$J(I_{S\setminus G}; x) \to 0 \ \text{ as } \ G \nearrow S;$$

see (2.4) and (2.5).

Part (i) of this lemma follows from [16, Ch. 3], while part (ii) can be obtained from the theory of renewal-reward processes in [19, Ch. 3]. A proof of part (iii) can be found, for instance, in [5]. Finally, part (iv) follows combining part (iii) with the fact that, by Lemma 2.1(iv), each function $I_{S\setminus G}$ has associated an AO stationary policy.

3 The Result

In this section Blackwell optimal policies are characterized in terms of the *AROE*; see Theorem 3.1 below. To start with, a sequence $\{M_n\}$ of MDP's is introduced in Definition 3.1 and then some classes of policies are introduced in Definition 3.2. In words, the sequence $\{M_n\}$ is (recursively) constructed as follows: (i) M_0 is the MDP introduced in Section 2, and (ii) given M_n, the model M_{n+1} is determined in the following way. The first step consists in finding a solution of the *AROE* associated to M_n, which is denoted by $(g_n, -h_{n+1})$. Next, the optimizers in the right-hand side of the *AROE* for M_n determine the admissible actions for model M_{n+1}, while h_{n+1} is the corresponding reward function. This construction is now described in a precise manner and the opportunity is taken to introduce some notation which will be used in the sequel.

Definition 3.1 *Let $M = (S, C, \{C(x)\}, r, p)$ be the MDP introduced in Section 2. The sequence of MDP's $\{M_n = (S, C, \{C_n(x)\}, h_n, p)\}$ is recursively defined as follows:*

(i) $M_0 := M$.
Thus, M_0 is just the original MDP, so that $h_0 = r$ and $C_0(x) = C(x)$, $x \in S$; set $\mathbb{K}_0 := \mathbb{K}$.

(ii) Let $g_0 \in \mathbb{R}$ and $h_1 \in \mathbb{B}(S)$ satisfy the AROE associated to M_0, that is,

$$g_0 - h_1(x) = \sup_{a \in C_0(x)} [r(x,a) - \sum_y p_{xy}(a) h_1(y)], \quad x \in S,$$

where without loss of generality it is assumed that

$$\|h_1\| \leq B\|r\|;$$

see Lemma 2.1(ii). Define Mandl's discrepancy function Φ_0 *by [17]*

$$\Phi_0(x,a) := g_0 - h_1(x) - [r(x,a) - \sum_y p_{xy}(a)h_1(y)], \quad (x,a) \in \mathbb{K}_0. \tag{3.1}$$

Using that r *and* h_1 *are continuous and bounded in their respective domains together with Assumption 2.1, it is not difficult to see that* $\Phi_0 \in \mathbb{B}(\mathbb{K})$*; furthermore, (2.6) implies that* $\Phi \geq 0$*. Also, observe that Lemma 2.1(iv) yields that, for each* $x \in S$, $\min_{a \in C_0(x)} \Phi_0(x,a) = 0$*, and then the set*

$$C_1(x) := \{a \in C_0(x) \,|\, \Phi_0(x,a) = 0\} \tag{3.2}$$

is a nonempty and closed subset of $C_0(x)$*. Since the latter is a compact set, so is* $C_1(x)$*. Finally, define*

$$M_1 := (S, C, \{C_1(x)\}, h_1, p).$$

(iii) Suppose that $M_n := (S, C, \{C_n(x)\}, h_n, p)$ *is given, where* $n \geq 1$, $h_n \in \mathbb{B}(S)$*, and each* $C_n(x)$ *is a nonempty and compact subset of* C*. Let* $g_n \in \mathbb{R}$ *and* $h_{n+1} \in \mathbb{B}(S)$ *satisfy the AROE associated to* M_n*, i.e.,*

$$g_n - h_{n+1}(x) = \sup_{a \in C_n(x)} [h_n(x) - \sum_y p_{xy}(a)h_{n+1}(y)], \quad x \in S,$$

where h_{n+1} *is selected in such a way that*

$$\|h_{n+1}\| \leq B\|h_n\|$$

(see Remark 2.1 and Lemma 2.1(ii)), and set $\mathbb{K}_n := \{(x,a) | \, x \in S, a \in C_n(x)\}$*. Now define the discrepancy function* Φ_n *associated to* M_n *by*

$$\Phi_n(x,a) := g_n - h_{n+1}(x) - [h_n(x) - \sum_y p_{xy}(a)h_{n+1}(y)], \quad (x,a) \in \mathbb{K}_n. \tag{3.3}$$

As before it is not difficult to see that (a) $\Phi_n \geq 0$*, (b)* $\Phi_n \in \mathbb{B}(\mathbb{K}_n)$*, and (c) for each* $x \in S$, $\Phi_n(x,a) = 0$ *for some* $a \in C_n(x)$*. Combining this fact with the compactness of* $C_n(x)$*, it follows that the sets*

$$C_{n+1}(x) := \{a \in C_n(x) | \Phi_n(x,a) = 0\}, \quad x \in S, \tag{3.4}$$

are nonempty and compact. The model M_{n+1} *is then defined by*

$$M_{n+1} := (S, C, \{C_{n+1}(x)\}, h_{n+1}, p).$$

Remark 3.1 (*i*) The MDP's M_n are 'nested', in the sense that the sets of admissible actions satisfy $C_{n+1}(x) \subset C_n(x)$ for all $x \in S,\ n \in \mathbb{N}$; see (3.2) and (3.4).

(*ii*) Since $\|h_{n+1}\| \leq B\|h_n\|$ it follows that

$$\|h_n\| \leq B^n\|h_0\| = B^n\|r\|, \quad n \in \mathbb{N}. \tag{3.5}$$

(*iii*) Since $\mathbb{K}_n = \{(x,a) | \, a \in C_n(x), x \in S\}$, (3.1)–(3.4) yield that $\mathbb{K}_{n+1} = \{(x,a) \in \mathbb{K}_n | \, \Phi_n(x,a) = 0\}$.

To continue, some classes of policies are introduced.

Definition 3.2 *(i) For each $x \in S$ and $n \in \mathbb{N}$ define $\mathbb{P}_n(x) \subset \mathbb{P}$ as follows:*

$$\mathbb{P}_n(x) := \{\pi \in \mathbb{P} \mid P_x^\pi[A_t \in C_n(X_t)] = 1 \text{ for all } t \in \mathbb{N}\}.$$

(ii) For each $x \in S$ set

$$\mathbb{P}_\infty(x) := \bigcap_{n=0}^{\infty} \mathbb{P}_n(x),$$

whereas

$$\mathbb{P}_\infty := \bigcap_{x \in S} \mathbb{P}_\infty(x).$$

(iii) Given $x \in S$, define $C_\infty(x) := \bigcap_{n=0}^\infty C_n(x)$, while $\mathbb{F}_\infty := \Pi_{x \in S} C_\infty(x)$.

Remark 3.2 *(i) Since the sets $C_n(x)$ are nonempty and compact and $C_{n+1}(x) \subset C_n(x)$, it follows that $C_\infty(x)$ is a nonempty compact set and, consequently, so is $\mathbb{F}_\infty$ [7, pp. 223–224]. On the other hand, it is useful to observe that*

$$\mathbb{P}_\infty \cap \mathbb{F} = \mathbb{F}_\infty. \tag{3.6}$$

(ii) In words, a policy π belongs to $\mathbb{P}_n(x)$ if and only if the following occurs P_x^π–almost surely: each A_t is an admissible action at X_t with respect to model M_n.

(iii) Let $n \in \mathbb{N}$, $x \in S$ and $\pi \in \mathbb{P}_{n+1}(x)$ be given. In this case $1 = P_x^\pi[A_t \in C_{n+1}(X_t),\ t \in \mathbb{N}] = P_x^\pi[(X_t, A_t) \in \mathbb{K}_{n+1},\ t \in \mathbb{N}]$. Then, Remark 3.1(iii) yields that $P_x^\pi[\Phi_n(X_t, A_t) = 0 \text{ for all } t \in \mathbb{N}] = 1$. Using

that $\Phi_n \geq 0$*, this is equivalent to* $E_x^\pi[\Phi_n(X_t, A_t)] = 0$, $t \in \mathbb{N}$*, and (2.1) allows to state the following:*

$$\text{If} \quad \pi \in \mathbb{P}_{n+1}(x), \quad \text{then,} \quad V_\alpha(\pi; \Phi_n; x) = 0, \quad \alpha \in (0,1).$$

Consequently,

(iv) If $\pi^* \in \mathbb{P}_\infty(x)\,(\subset \mathbb{P}_{n+1}(x))$*, then* $V_\alpha(\pi^*; \Phi_n; x) = 0$ *for all* $n \in \mathbb{N}$ *and* $\alpha \in (0,1)$*.*

The next theorem, which provides a characterization of Blackwell optimal policies, is the main result in this note.

Theorem 3.1 *(i) A policy* $\pi \in \mathbb{P}$ *is BO at* x *if and only if* $\pi \in \mathbb{P}_\infty(x)$*.*

(ii) Policy π *is BO if and only if* $\pi \in \mathbb{P}_\infty$*.*

(iii) A stationary policy $f \in \mathbb{F}$ *is BO if and only if* $f \in \mathbb{F}_\infty$*.*

According to Theorem 3.1(i), a policy π is *BO* at x if and only if the following occurs P_x^π-almost surely: for each $t \in \mathbb{N}$, the action A_t is admissible at X_t for each one of the models M_n in Definition 3.2. Thus, the problem of determining the *BO* policies reduces to determining the sets $C_n(x)$, that is, to solving the *AROE*'s corresponding to each one of the models M_n. A proof of Theorem 3.1 will be presented in Section 5. For the moment, notice that part (ii) follows from part (i) in combination with the definition of *BO* policy, and that part (ii) and (3.6) together imply part (iii).

4 Preliminaries

This section contains the technical tools that will be used in the proof of Theorem 3.1 , which are given below in the form of Lemmas 4.1–4.4 and Theorem 4.1. Throughout the remainder, Assumptions 2.1 and 2.2 are supposed to hold. On the other hand, before going any further it is convenient to introduce some auxiliary notions.

Definition 4.1 *(i) For each* $\alpha \in (0,1)$*, the corresponding interest rate* $\rho(\alpha)$ *is given by*

$$\rho(\alpha) := \frac{1-\alpha}{\alpha}. \tag{4.1}$$

(ii) Given $h \in \mathbb{B}(S)$, the function $\tilde{h} : \mathbb{K} \to \mathbb{R}$ is defined by

$$\tilde{h}(w,a) := \sum_y p_{wy}(a)h(y), \quad (w,a) \in \mathbb{K}. \tag{4.2}$$

Remark 4.1 *Notice that $\|\tilde{h}\| \leq \|h\|$. Moreover, using Assumption 2.1 it is not difficult to see that $\tilde{h} \in \mathbb{B}(\mathbb{K})$.*

The following simple result is the starting point in the way to the proof of Theorem 3.1.

Lemma 4.1 *Let $x \in S$ and $n \in \mathbb{N}$ be fixed.*

(i) For each $\alpha \in (0,1)$, $\pi \in \mathbb{P}$ and $h \in \mathbb{B}(S)$,

$$V_\alpha(\pi; \tilde{h}; x) = [V_\alpha(\pi; h; x) - h(x)]/\alpha;$$

see (2.1) and (4.2).

(ii) For each $\pi \in \mathbb{P}_{n+1}(x)$

$$(1-\alpha)V_\alpha(\pi; h_n; x) \to g_n \text{ as } \alpha \nearrow 1;$$

recall that, by Definition 3.1, h_n and g_n are the reward function and the optimal average reward associated to the model M_n, respectively.

Proof: (i) By the Markov property, (4.2) is equivalent to

$$E_x^\pi[h(X_{t+1})|X_t = w, A_t = a] = \tilde{h}(w,a), \quad (w,a) \in \mathbb{K},\ t \in \mathbb{N},$$

so that

$$E_x^\pi[h(X_{t+1})] = E_x^\pi[\tilde{h}(X_t, A_t)], \quad t \in \mathbb{N}.$$

Then, (2.1) yields that for each $\alpha \in (0,1)$, $V_\alpha(\pi; \tilde{h}; x) = \sum_{t=0}^\infty \alpha^t E_x^\pi[\tilde{h}(X_t, A_t)] = \sum_{t=0}^\infty \alpha^t E_x^\pi[h(X_{t+1})]$, and using that $E_x^\pi[h(X_0)] = h(x)$, a simple change of variable in the index of the last summation yields

$$V_\alpha(\pi; \tilde{h}; x) = [\sum_{t=0}^\infty \alpha^t E_x^\pi[h(X_t, A_t)] - h(x)]/\alpha,$$

and the conclusion follows from (2.1).

(ii) First, consider the case $n = 0$. In this situation, (3.1) yields (recall that $h_0 = r$)

$$g_0 - h_1(w) = h_0(w,a) + \Phi_0(w,a) - \sum_y p_{wy}(a)h_1(y), \quad (w,a) \in \mathbb{K}_0 (\equiv \mathbb{K}).$$

or, equivalently,

$$-h_1(w) = H_{0\alpha}(w,a) - \alpha \sum_y p_{wy}(a) h_1(y), \quad (w,a) \in \mathbb{K}_0, \tag{4.3}$$

where

$$H_{0\alpha}(w,a) := h_0(w,a) + \Phi_0(w,a) - g_0 - (1-\alpha)\tilde{h}_1(w,a), \quad (w,a) \in \mathbb{K}_0;$$

see (4.2).

Now observe that $H_{0\alpha} \in \mathbb{B}(\mathbb{K}_0) \equiv \mathbb{B}(\mathbb{K})$, so that (4.3) yields (see Theorem 2.2(a) in [11])

$$\begin{aligned} -h_1(x) &= V_\alpha(\pi; H_{0\alpha}; x) \\ &= V_\alpha(\pi; h_0; x) + V_\alpha(\pi; \Phi_0; x) - g_0/(1-\alpha) \\ &\quad - (1-\alpha) V_\alpha(\pi; \tilde{h}_1; x), \quad x \in S, \end{aligned}$$

where the second equality follows from the linearity of the mapping $r \mapsto V_\alpha(\pi; r; x)$; see (2.1). Thus, since $V_\alpha(\pi; \Phi_0; x) = 0$ (recall that $\pi \in \mathbb{P}_{0+1} = \mathbb{P}_1$ and see Remark 3.2(iii)), the last displayed equation implies

$$\begin{aligned} |(1-\alpha)V_\alpha(\pi; h_0; x) - g_0| &= |-h_1(x) + (1-\alpha)V_\alpha(\pi; \tilde{h}_1; x)|(1-\alpha) \\ &\le (\|h_1\| + \|\tilde{h}_1\|)(1-\alpha) \to 0 \text{ as } \alpha \nearrow 1; \end{aligned}$$

see (2.2) for the inequality. To conclude, consider the case $n \ge 1$ and observe that (3.3) can be written as

$$\begin{aligned} -h_{n+1}(w) &= h_n(w) + \Phi_n(w,a) - g_n - (1-\alpha)\tilde{h}_{n+1}(w,a) \\ &\quad -\alpha \sum_y p_{wy}(a) h_{n+1}(y), \quad (w,a) \in \mathbb{K}_n. \end{aligned} \tag{4.4}$$

Now pick $\pi \in \mathbb{P}_{n+1}(x)$ and notice that $P_x^\pi[(X_t, A_t) \in \mathbb{K}_{n+1}] = 1$, $t \in \mathbb{N}$; since $\mathbb{K}_{n+1} \subset \mathbb{K}_n$, it follows that $P_x^\pi[(X_t, A_t) \in \mathbb{K}_n \text{ for all } t] = 1$. Then (4.4) yields, via the same arguments used in the case $n = 0$, that

$$\begin{aligned} -h_{n+1}(x) &= V_\alpha(\pi; h_n; x) + V_\alpha(\pi; \Phi_n; x) - g_n/(1-\alpha) \\ &\quad - (1-\alpha) V_\alpha(\pi; \tilde{h}_{n+1}; x) \\ &= V_\alpha(\pi; h_n; x) - g_n/(1-\alpha) - (1-\alpha) V_\alpha(\pi; \tilde{h}_{n+1}; x), \end{aligned}$$

where Remark 3.2(iii) was used to obtain the second equality. Therefore,

$$\begin{aligned} |(1-\alpha)V_\alpha(\pi; h_n; x) - g_n| = \\ &= |-h_{n+1}(x) + (1-\alpha)V_\alpha(\pi; \tilde{h}_{n+1}; x)|(1-\alpha) \\ &\le (\|h_{n+1}\| + \|\tilde{h}_{n+1}\|)(1-\alpha) \to 0 \text{ as } \alpha \nearrow 1, \end{aligned}$$

and the proof is complete. ■

Part (i) of Lemma 2.1 will now be used in the proof of the following result.

Lemma 4.2 *Let $x \in S$ and $\pi^* \in \mathbb{IP}_\infty$ be arbitrary but fixed.*

(i) For each $\pi \in \mathbb{IP}(\equiv \mathbb{IP}_0(x))$

$$V_\alpha(\pi^*; r; x) - V_\alpha(\pi; r; x) = V_\alpha(\pi; \Phi_0; x) + \rho(\alpha)[V_\alpha(\pi^*; h_1; x) - V_\alpha(\pi; h_1; x)];$$

see (4.1) and Definition 3.2.

(ii) If $n \geq 1$ and $\pi \in \mathbb{IP}_n(x)$, then

$$V_\alpha(\pi^*; h_n; x) - V_\alpha(\pi; h_n; x) = V_\alpha(\pi; \Phi_n; x) + \rho(\alpha)[V_\alpha(\pi^*; h_{n+1}; x) - V_\alpha(\pi; h_{n+1}; x)].$$

Proof: (i) Notice that for each $(w, a) \in \mathbb{IK}(\equiv \mathbb{IK}_0)$ and $\alpha \in (0, 1)$, (3.1) can be written as

$$-h_1(w) = r_1(w, a) - \alpha \sum_y p_{wy}(a) h_1(y), \quad (w, a) \in \mathbb{IK}, \tag{4.5}$$

where

$$r_1(w, a) := r(w, a) + \Phi_0(w, a) - g_0 - (1-\alpha)\tilde{h}_1(w, a), \quad (w, a) \in \mathbb{IK}; \tag{4.6}$$

see (4.2) for the definition of $\tilde{h}_1$. Since $r_1 \in \mathbb{IB}(\mathbb{IK})$, (4.5) yields ([11, theorem 2.2(a)]) that $V_\alpha(\pi; r_1; x) = -h_1(x)$ and then (see (4.6) and (2.1)) it follows that

$$\begin{aligned} -h_1(x) &= V_\alpha(\pi; r_1; x) \\ &= V_\alpha(\pi; r; x) + V_\alpha(\pi; \Phi_0; x) - g_0/(1-\alpha) \\ &\quad - (1-\alpha) V_\alpha(\pi; \tilde{h}_1; x). \end{aligned}$$

Combining Lemma 4.1(i) with Definition 4.1(i), the last equality yields, via straightforward calculations, that

$$-h_1(x)/\alpha = V_\alpha(\pi; r; x) + V_\alpha(\pi; \Phi_0; x) - g_0/(1-\alpha) - \rho(\alpha) V_\alpha(\pi; h_1; x). \tag{4.7}$$

Replacing π by π^* in this equation and recalling that $V_\alpha(\pi^*; \Phi_0; x) = 0$ (by Remark 3.1 (iv)), it follows that

$$-h_1(x)/\alpha = V_\alpha(\pi^*; r; x) - g_0/(1-\alpha) - \rho(\alpha) V_\alpha(\pi^*; h_1; x),$$

and the conclusion follows combining this equality with (4.7).

(ii) Let $\pi \in \mathbb{P}_n(x)$ be arbitrary but fixed and observe that (4.4) is equivalent to

$$\begin{aligned} g_n - h_{n+1}(w) &= h_n(w) + \Phi_n(w,a) - (1-\alpha)\tilde{h}_{n+1}(w,a) \\ &\quad - \alpha \sum_y p_{wy}(a) h_{n+1}(y), \quad (w,a) \in \mathbb{K}_n; \end{aligned}$$

see (4.2). Since $P_x^\pi[(X_t, A_t) \in \mathbb{K}_n \text{ for all } t \in \mathbb{N}] = 1$, this equation yields, via the arguments used in the proof of part (i), that

$$\begin{aligned} &-h_{n+1}(x)/\alpha = \\ &\quad V_\alpha(\pi; h_n; x) + V_\alpha(\pi; \Phi_n; x) - g_n/(1-\alpha) - \rho(\alpha) V_\alpha(\pi; h_{n+1}; x), \end{aligned}$$

and replacing π by π^* it follows that

$$-h_{n+1}(x)/\alpha = V_\alpha(\pi^*; h_n; x) - g_n/(1-\alpha) - \rho(\alpha) V_\alpha(\pi^*; h_{n+1}; x),$$

where $V_\alpha(\pi^*; \Phi_n; x) = 0$ was used; see Remark 3.2(iv). Then the conclusion is obtained by substracting the last two displayed equations. ■

The next result is an extension of Lemma 4.2(i).

Lemma 4.3 *Let $x \in S$ and $n \in \mathbb{N}$ be fixed. Then, for $\pi^* \in \mathbb{P}_\infty$, $\pi \in \mathbb{P}_n(x)$ and $\alpha \in (0,1)$,*

$$\begin{aligned} &V_\alpha(\pi^*; r; x) - V_\alpha(\pi; r; x) = \\ &\quad \rho(\alpha)^n [V_\alpha(\pi; \Phi_n; x) + \rho(\alpha)(V_\alpha(\pi^*; h_{n+1}; x) - V_\alpha(\pi; h_{n+1}; x))]. \quad (4.8) \end{aligned}$$

Proof: (By induction.) For $n = 0$ the assertion reduces to that in Lemma 4.2(i). Suppose that the result is true for some $n \in \mathbb{N}$ and that $\pi \in \mathbb{P}_{n+1}(x) \subset \mathbb{P}_n(x)$. In this case, by Remark 3.2(iii), $V_\alpha(\pi; \Phi_n; x) = 0$ for all $\alpha \in (0,1)$. Therefore, (4.8) and the induction hypothesis together yield

$$V_\alpha(\pi^*; r; x) - V_\alpha(\pi; r; x) = \rho(\alpha)^{n+1}[V_\alpha(\pi^*; h_{n+1}; x) - V_\alpha(\pi; h_{n+1}; x)],$$

and applying Lemma 4.2(ii) with $n+1$ instead of n it follows that

$$\begin{aligned} &V_\alpha(\pi^*; r; x) - V_\alpha(\pi; r; x) = \\ &\quad \rho(\alpha)^{n+1}[V_\alpha(\pi; \Phi_{n+1}; x) + \rho(\alpha)(V_\alpha(\pi^*; h_{n+2}; x) - V_\alpha(\pi; h_{n+2}; x))], \end{aligned}$$

which is (4.8) with $n+1$ replacing n. This completes the proof. ■

The following lemma is the last preliminary before Theorem 4.1, which together with the property in Lemma 2.1(*ii*) are the essential components of the proof of Theorem 3.1 presented in the next section.

Lemma 4.4 *Let $x \in S$ and $n \in \mathbb{N}$ be arbitrary but fixed. Then (i)–(ii) below occur.*

(i) Assume that $\{a_k\} \subset C_n(x)$ is such that

$$\Phi_n(x, a_k) \to 0 \text{ as } \ k \to \infty,$$

and that

$$\lim_{k\to\infty} a_k =: a$$

exists. Then, $a \in C_{n+1}(x)$.

(ii) Let $\pi \in \mathbb{P}_n(x)$ be arbitrary. Then,

$$\lim_{\alpha\nearrow 1} V_\alpha(\pi; \Phi_n; x) = E_x^\pi[\sum_{t=0}^{\infty} \Phi_n(X_t, A_t)].$$

Moreover, if $\pi \notin \mathbb{P}_{n+1}(x)$, then

$$E_x^\pi[\sum_{t=0}^{\infty} \Phi_n(X_t, A_t)] > 0.$$

Proof: (i) Since $\{a_k\} \subset C_n(x)$ and $C_n(x)$ is compact, $\lim_{k\to\infty} a_k = a$ implies that $a \in C_n(x)$. Using that Φ_n is continuous, it follows that $\Phi_n(x, a) = \lim_{k\to\infty} \Phi(x, a_k) = 0$, so that $a \in C_{n+1}(x)$; see (3.4).

(ii) Let $\pi \in \mathbb{P}_n(x)$. By (2.1), $V_\alpha(\pi; \Phi_n; x) = E_x^\pi[\sum_{t=0}^{\infty} \alpha^t \Phi_n(X_t, A_t)]$ and, recalling that $\Phi_n \geq 0$, the monotone convergence theorem yields

$$\lim_{\alpha\nearrow 1} V_\alpha(\pi; \Phi_n; x) = E_x^\pi[\sum_{t=0}^{\infty} \Phi_n(X_t, A_t)].$$

To conclude observe that, since Φ_n is nonegative, $E_x^\pi[\sum_{t=0}^{\infty} \Phi_n(X_t, A_t)] = 0$ implies that $\Phi_n(X_t, A_t) = 0$ P_x^π–almost surely for all $t \in \mathbb{N}$, i.e., $P_x^\pi[A_t \in C_{n+1}(X_t),\ t \in \mathbb{N}] = 1$, so that $\pi \in \mathbb{P}_{n+1}(x)$. Therefore, $\pi \notin \mathbb{P}_{n+1}(x)$ implies that $E_x^\pi[\sum_{t=0}^{\infty} \Phi_n(X_t, A_t)] > 0$. ■

Before stating Theorem 4.1, it is convenient to introduce some auxiliary functions.

Definition 4.2 *For each $n \in \mathbb{N}$, define $\hat{\Phi}_{n+1} : \mathbb{K}_n \to \mathbb{R}$ by*

$$\hat{\Phi}_{n+1}(w,a) := g_{n+1} - h_{n+2}(w) - [h_{n+1}(w) - \sum_y p_{wy}(a) h_{n+2}(y)], \quad (w,a) \in \mathbb{K}_n. \tag{4.9}$$

From Assumption (2.1) and the boundedness of h_{n+1} and h_{n+2} it follows that $\hat{\Phi}_{n+1} \in \mathbb{B}(\mathbb{K})$. Also, $\hat{\Phi}_{n+1}$ is an extension of Φ_{n+1}; the domain of the latter is $\mathbb{K}_{n+1} \subset \mathbb{K}_n$. Thus, $\hat{\Phi}_{n+1}(x,a) = \Phi_{n+1}(x,a) \geq 0$ whenever $(x,a) \in \mathbb{K}_{n+1}$ but, in general, it cannot be asserted that $\hat{\Phi}_{n+1} \geq 0$ in its whole domain.

Theorem 4.1 *Let $x \in S$ and $n \in \mathbb{N}$ be fixed and suppose that $\pi \in \mathbb{P}_n(x)$ satisfies*

$$E_x^\pi[\sum_{t=0}^{\infty} \Phi_n(X_t, A_t)] < \infty.$$

Then (i)–(v) below occur.

(i) For each $y \in S$,

$$\liminf_{t\to\infty} \hat{\Phi}_{n+1}(X_t, A_t) I[X_t = y] \geq 0 \quad P_x^\pi\text{−almost surely.}$$

(ii) For each finite set $G \subset S$,

$$\liminf_{t\to\infty} \hat{\Phi}_{n+1}(X_t, A_t) I[X_t \in G] \geq 0 \quad P_x^\pi\text{−almost surely.}$$

$$(iii) \qquad \liminf_{n\to\infty} \frac{1}{n+1} E_x^\pi[\sum_{t=0}^{n} \hat{\Phi}_{n+1}(X_t, A_t)] \geq 0$$

$$(iv) \qquad \liminf_{\alpha\nearrow 1} (1-\alpha) V_\alpha(\pi; \hat{\Phi}_{n+1}; x) \geq 0.$$

$$(v) \qquad \limsup_{\alpha\nearrow 1} (1-\alpha) V_\alpha(\pi; h_{n+1}; x) \leq g_{n+1}.$$

Proof: To begin with, notice that the following occurs P_x^π-almost surely:

$$\Phi_n(X_t, A_t) \to 0 \text{ as } t \to \infty \text{ and } A_t \in C_n(X_t), \quad t \in \mathbb{N}. \tag{4.10}$$

This assertion follows from $E_x^\pi[\sum_{t=0}^\infty \Phi_n(X_t, A_t)] < \infty$ and the fact that $\pi \in \mathbb{P}_n(x)$; see Definition 3.2.

(i) Let $y \in S$ be fixed and select a sample path $\{(X_t, A_t)\}$ such that (4.10) holds. It will be shown that for such a trajectory,

$$\liminf_{t\to\infty} \hat{\Phi}_{n+1}(X_t, A_t) I[X_t = y] \geq 0. \tag{4.11}$$

Notice that this yields the conclusion, since (4.10) occurs P_x^π-almost surely. To verify (4.11) pick a sequence $\{t_k\} \subset \mathbb{N}$ satisfying

$$\lim_{k\to\infty} \hat{\Phi}_{n+1}(X_{t_k}, A_{t_k}) I[X_{t_k} = y] = \liminf_{t\to\infty} \hat{\Phi}_{n+1}(X_t, A_t) I[X_t = y]. \tag{4.12}$$

and consider the following two cases:

Case 1: $X_{t_k} \neq y$ for infinitely many k's.

In this situation, taking a subsequence if necessary, it can be assumed that $X_{t_k} \neq y$ for *all* $k \in \mathbb{N}$. Then $\hat{\Phi}_{n+1}(X_{t_k}, A_{t_k}) I[X_{t_k} = y] = 0$ for all k and (4.12) implies that $\liminf_{t\to\infty} \hat{\Phi}_{n+1}(X_t, A_t) I[X_t = y] = 0$, so that (4.11) holds.

Case 2: $X_{t_k} = y$ except for at most finitely many k's.

In this case (after taking a subsequence if necessary) it can be assumed that $X_{t_k} = y$ and then $A_{t_k} \in C_n(X_{t_k}) = C_n(y)$ for *all* k; recall that the trajectory $\{(X_t, A_t)\}$ satisfies (4.10). Since $C_n(y)$ is compact, selecting an appropriate subsequence it can be supposed that $\lim_{k\to\infty} A_{t_k} =: A$ exists. On the other hand, (4.10) implies that

$$\Phi_n(y, A_{t_k}) = \Phi_n(X_{t_k}, A_{t_k}) \to 0 \quad \text{as} \quad k \to \infty,$$

and then, an application of Lemma 4.4 (i) yields that $A \in C_{n+1}(y)$, that is $(y, A) \in \mathbb{K}_{n+1}$. Also observe that

$$\lim_{k\to\infty} \hat{\Phi}_{n+1}(X_{t_k}, A_{t_k}) I[X_{t_k} = y] = \lim_{k\to\infty} \hat{\Phi}_{n+1}(y, A_{t_k}) = \hat{\Phi}_{n+1}(y, A); \tag{4.13}$$

recall that $X_{t_k} = y$ for all k and that $\hat{\Phi}_{n+1}$ is continuous. Using that $(y, A) \in \mathbb{K}_{n+1}$ it follows that $\hat{\Phi}_{n+1}(y, A) = \Phi_{n+1}(y, A) \geq 0$ and, together with (4.12) and (4.13), this inequality yields (4.11). To conclude

notice that, since a given trajectory $\{(X_t, A_t)\}$ necessarily falls in either Case 1 or Case 2, the above discussion can be summarized as follows: For an arbitrary sample path $\{(X_t, A_t)\}$ satisfying (4.10), the inequality (4.11) is valid. This completes the proof of part (i) since, as already mentioned, (4.10) occurs P_x^π-almost surely.

(ii) The conclusion follows form part (i) after observing that

$$\hat{\Phi}_{n+1}(X_t, A_t)I[X_t \in G\,] = \sum_{y\in G} \hat{\Phi}_{n+1}(X_t, A_t)I[X_t = y\,].$$

(iii) Notice that part (ii) implies that for each finite set $G \subset S$, $\liminf_{n\to\infty} \frac{1}{n+1}\sum_{t=0}^n \hat{\Phi}_{n+1}(X_t, A_t)I[X_t \in G\,] \geq 0 \quad P_x^\pi$-almost surely. Since $\hat{\Phi}_{n+1}$ is bounded, Fatou's lemma yields

$$\liminf_{n\to\infty} \frac{1}{n+1} E_x^\pi[\sum_{t=0}^n \hat{\Phi}_{n+1}(X_t, A_t)I[X_t \in G\,]\,] \geq 0. \tag{4.14}$$

On the other hand, observe that

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n+1} E_x^\pi[\sum_{t=0}^n |\hat{\Phi}_{n+1}(X_t, A_t)|I[X_t \in S\setminus G\,]\,] \\ &\leq \|\hat{\Phi}_{n+1}\| J(\pi; I_{S\setminus G}; x) \\ &\leq \|\hat{\Phi}_{n+1}\| J(I_{S\setminus G}; x);\end{aligned}$$

see (2.4) and (2.5). Using this inequality it not difficult to see that

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{n+1} E_x^\pi[\sum_{t=0}^n \hat{\Phi}_{n+1}(X_t, A_t)] \geq \\ &\liminf_{n\to\infty} \frac{1}{n+1} E_x^\pi[\sum_{t=0}^n \hat{\Phi}_{n+1}(X_t, A_t)I[X_t \in G\,]\,] \\ &- \|\hat{\Phi}_{n+1}\| J(I_{S\setminus G}; x);\end{aligned}$$

combining this with (4.14) it follows that

$$\liminf_{n\to\infty} \frac{1}{n+1} E_x^\pi[\sum_{t=0}^n \hat{\Phi}_{n+1}(X_t, A_t)] \geq -\|\hat{\Phi}_{n+1}\| J(I_{S\setminus G}; x),$$

and the conclusion is obtained by letting G increase to S and applying Lemma 2.2(iv).

(iv) This part follows from part (iii) via a Tauberian theorem; see, for instance, Proposition 4-7 in [12].

(v) Notice that equation (4.9) defining $\hat{\Phi}_{n+1}$ is equivalent to

$$\begin{aligned} -h_{n+2}(w) &= [h_{n+1}(w) + \hat{\Phi}_{n+1}(w,a) - (1-\alpha)\tilde{h}_{n+2}(w,a) - g_{n+1}] \\ &\quad - \alpha \sum_y p_{wy}(a) h_{n+2}(y), \quad (w,a) \in \mathbb{K}_n, \end{aligned} \tag{4.15}$$

where $\alpha \in (0,1)$; see (4.2) for the definition of $\tilde{h}_{n+2}$. Since the function within brackets in the right hand side of (4.15) belongs to $\mathbb{B}(\mathbb{K}_n)$ and $P_x^\pi[(X_t, A_t) \in \mathbb{K}_n] = 1$, it follows [11, Theorem 2.2(a)] that

$$\begin{aligned} -h_{n+2}(x) &= V_\alpha(\pi; h_{n+1}; x) + V_\alpha(\pi; \hat{\Phi}_{n+1}; x) \\ &\quad - (1-\alpha) V_\alpha(\pi; \tilde{h}_{n+2}; x) - g_{n+1}/(1-\alpha) \end{aligned}$$

which yields,

$$\begin{aligned} (1-\alpha)[V_\alpha(\pi; h_{n+1}; x) + V_\alpha(\pi; \hat{\Phi}_{n+1}; x)] - g_{n+1} = \\ (1-\alpha)[-h_{n+2}(x) + (1-\alpha) V_\alpha(\pi; \tilde{h}_{n+2}; x)]. \end{aligned}$$

Using that $|-h_{n+2}(x) + (1-\alpha)V_\alpha(\pi; \tilde{h}_{n+2}; x)| \leq \|h_{n+2}\| + \|\tilde{h}_{n+2}\|$, the last equation implies that

$$\lim_{\alpha \nearrow 1} (1-\alpha)[V_\alpha(\pi; h_{n+1}; x) + V_\alpha(\pi; \hat{\Phi}_{n+1}; x)] = g_{n+1},$$

and then, from part (iv) it follows that

$$\limsup_{\alpha \nearrow 1} (1-\alpha) V_\alpha(\pi; h_{n+1}; x) \leq g_{n+1}.$$

This completes the proof. ■

5 Proof of Theorem 3.1

In this section Theorem 3.1 is established. The proof given below uses Remark 3.2(iv), Lemmas 4.1(ii), 4.3 and 4.4(ii), as well as Theorem 4.1(v).

Proof of Theorem 3.1 Let $x \in S$ be fixed.

(i) Pick $\pi^* \in \mathbb{P}_\infty$. First, it will be shown that π^* is Blackwell optimal at x. With this in mind, select $\pi \in \mathbb{P}$ and consider the following two cases.

Case 1. $\pi \notin \mathbb{P}_\infty(x)$.

In this case there exists $n \in \mathbb{N}$ such that $\pi \in \mathbb{P}_n(x)$ but $\pi \notin \mathbb{P}_{n+1}(x)$; notice that $\pi \in \mathbb{P}_0(x) = \mathbb{P}$. By Lemma 4.3 and (4.1), for all $\alpha \in (0,1)$

$$[V_\alpha(\pi^*; r; x) - V_\alpha(\pi; r; x)]\rho(\alpha)^{-n} = V_\alpha(\pi; \Phi_n; x) + \frac{1-\alpha}{\alpha}[V_\alpha(\pi^*; h_{n+1}; x) - V_\alpha(\pi; h_{n+1}; x)]. \tag{5.1}$$

It will be shown that

$$\liminf_{\alpha \nearrow 1} [V_\alpha(\pi^*; r; x) - V_\alpha(\pi; r; x)]\rho(\alpha)^{-n} > 0. \tag{5.2}$$

To verify this inequality notice that policy π satisfies one of the following conditions (a) or (b); see Lemma 4.4(ii).

(a) $\lim_{\alpha \nearrow 1} V_\alpha(\pi; \Phi_n; x) = E_x^\pi[\sum_{t=0}^\infty \Phi_n(X_t, A_t)] = \infty$.

In this situation, observing that $\|(1-\alpha)V_\alpha(\cdot; h_{n+1}; \cdot)\| \leq \|h_{n+1}\|$ (by (2.2)), equation (5.1) yields that

$$\lim_{\alpha \nearrow 1} [V_\alpha(\pi^*; r; x) - V_\alpha(\pi; r; x)]\rho(\alpha)^{-n} = E_x^\pi[\sum_{t=0}^\infty \Phi_n(X_t, A_t)] = \infty,$$

so that (5.2) certainly occurs.

(b) $\lim_{\alpha \nearrow 1} V_\alpha(\pi; \Phi_n; x) = E_x^\pi[\sum_{t=0}^\infty \Phi_n(X_t, A_t)] < \infty$.

In this case Theorem 4.1(v) implies that

$$\limsup_{\alpha \nearrow 1} (1-\alpha)V_\alpha(\pi; h_{n+1}; x) \leq g_{n+1}.$$

On the other hand, since $\pi^* \in \mathbb{P}_\infty \subset \mathbb{P}_{n+2}(x)$, Lemma 4.1$(ii)$ yields that

$$\lim_{\alpha \nearrow 1} (1-\alpha)V_\alpha(\pi; h_{n+1}; x) = g_{n+1}.$$

Next, combining the last two convergences the following is obtained:

$$\liminf_{\alpha \nearrow 1} (1-\alpha)[V_\alpha(\pi^*; h_{n+1}; x) - V_\alpha(\pi; h_{n+1}; x)] \geq g_{n+1} - g_{n+1} = 0,$$

which, together with (5.1) yields

$$\liminf_{\alpha \nearrow 1} [V_\alpha(\pi^*; r; x) - V_\alpha(\pi; r; x)]\rho(\alpha)^{-n} \geq E_x^\pi[\sum_{t=0}^\infty \Phi_n(X_t, A_t)] > 0,$$

where the strict inequality is due to the fact that $\pi \in \mathbb{IP}_n(x)$ but $\pi \notin \mathbb{IP}_{n+1}(x)$; (see Lemma 4.4(*ii*)) so that, again, (5.2) occurs. To conclude, observe that, since $\rho(\alpha) > 0$ for all $\alpha \in (0,1)$, (5.2) implies that there exists $\alpha(\pi^*;\pi;x) \in (0,1)$ such that $[V_\alpha(\pi^*;r;x) - V_\alpha(\pi;r;x)] > 0$ for $\alpha \in (\alpha(\pi^*;\pi;x),1)$, so that (2.3) in the definition of Blackwell optimality at x is satisfied.

Case 2. $\pi \in \mathbb{IP}_\infty(x)$.

In this case (5.1) still holds and, in addition, $V_\alpha(\pi;\Phi_n;x) = 0$ for all $\alpha \in (0,1),\ n \in \mathbb{N}$; see Remark 3.1(*iv*). Therefore,

$$\begin{aligned} |V_\alpha(\pi^*;r;x) - V_\alpha(\pi;r;x)| & \\ &\leq 2\rho(\alpha)^n \|h_{n+1}\|/\alpha \\ &\leq 2[\rho(\alpha)B]^n \|r\|[B/\alpha], \quad \alpha \in (0,1), \quad n \in \mathbb{N}; \end{aligned} \tag{5.3}$$

see (2.2) and (3.5) for these inequalities. Next select $\alpha \in (B/(B+1),1)$ and notice that in this case $0 < \rho(\alpha)B < 1$. Therefore, after taking limit as $n \to \infty$ in (5.3) it follows that

$$V_\alpha(\pi^*;r;x) = V_\alpha(\pi;r;x), \quad \alpha \in (B/(B+1),1),$$

and then (2.3) is satisfied with $\alpha(\pi^*;\pi;x) = B/(B+1)$. In fact, using that two power series coinciding in an interval necessarily coincide in its whole domain, (2.1) implies that $V_\alpha(\pi^*;r;x) = V_\alpha(\pi;r;x)$ for all $\alpha \in (0,1)$, so that in this case $\alpha(\pi^*;\pi;x)$ in (2.3) can be taken equal to 0.

Now observe that a given policy π necessarily falls into Case 1 or Case 2, so that the previous discussion can be summarized as follows:

A policy $\pi^* \in \mathbb{IP}_\infty(x)$ is Blackwell optimal at x.

Conversely, if $\pi \notin \mathbb{IP}_\infty(x)$ the analysis of Case 1 above shows that for any $\pi^* \in \mathbb{IP}_\infty(x)$ the inequality $V_\alpha(\pi;r;x) - V_\alpha(\pi^*;r;x) < 0$ occurs for all α sufficiently close to 1, so that π is not Blackwell optimal at x. Thus, the conclusion is that a policy π^* is Blackwell optimal at x if and only if $\pi^* \in \mathbb{IP}_\infty(x)$, establishing part (i). The other parts of Theorem 3.1 can be obtained as outlined at the end of Section 3. ■

This section concludes with the following result giving a simple (and verifiable) condition to guarantee the uniqueness of a Blackwell optimal policy; this is a consequence of the analysis of Case 2 in the proof of Theorem 3.1.

Theorem 5.1 *(i) Let $x \in S$ be fixed and suppose that the mapping*

$$a \mapsto r(x,a), \quad a \in C(x) \text{ is one–to–one.} \tag{5.4}$$

Then, $C_\infty(x)$ is a singleton.

(ii) If (5.4) is satisfied for all $x \in S$, then

(a) There exists exactly one BO stationary policy, say $f^* \in \mathbb{F}_\infty$.

Moreover,

(b) $\mathbb{P}_\infty$ consists, essentially, of policy f^* in the following sense:

$$\pi^* \in \mathbb{P}_\infty \text{ if and only if } P_x^\pi[A_t = f^*(X_t)] = 1, \quad x \in S, \quad t \in \mathbb{N}.$$

Proof: (i) Let $a, a' \in C_\infty(x)$ and select $f, f' \in \mathbb{F}_\infty$ such that $f(x) = a$ and $f'(x) = a'$. In this case $f, f' \in \mathbb{P}_\infty \subset \mathbb{P}_\infty(x)$ so that f and f' are BO at x. Then, the analysis of Case 2 in the proof of Theorem 3.1 yields that $V_\alpha(f; r; x) = V_\alpha(f'; r; x)$ for $0 < \alpha < 1$, i.e.,

$$\sum_{t=0}^{\infty} \alpha^t E_x^f[r(X_t, A_t)] = \sum_{t=0}^{\infty} \alpha^t E_x^{f'}[r(X_t, A_t)], \quad \alpha \in (0,1),$$

and letting α decrease to 0 in both sides of this equality it follows that

$$\begin{aligned} r(x,a) &= r(x, f(x)) = E_x^f[r(X_0, A_0)] \\ &= E_x^{f'}[r(X_0, A_0)] = r(x, f'(x)) = r(x, a') \end{aligned}$$

and then (5.4) implies that $a = a'$, so that $C_\infty(x)$ is a singleton.

(ii) If (5.4) occurs for all $x \in S$, each set $C_\infty(x)$ is a singleton, say $C_\infty(x) = \{f^*(x)\}$, and then the corresponding policy f^* is the only member of $\mathbb{F}_\infty$; this establishes (a) and then part (b) follows from Definition 3.2, which yields the following:

$$\begin{aligned} \pi \in \mathbb{P}_\infty &\iff \text{for all } x \in S, \quad P_x^\pi[A_t \in C_\infty(X_t),\, t \in \mathbb{N}] = 1 \\ &\iff \text{for all } x \in S, \quad P_x^\pi[A_t = f^*(X_t),\, t \in \mathbb{N}] = 1, \end{aligned}$$

and the proof is complete. ■

6 Conclusion

This work has considered the problem of establishing the existence of Blackwell optimal policies in Markov decision processes satisfying the Simultaneous Doeblin Condition and endowed with a continuous and bounded reward function. According to Theorem 3.1, the Blackwell optimal policies are characterized by the property that they always prescribe actions maximizing the right hand side of each of the *AROE*'s corresponding to the sequence of 'nested' MDP's introduced in Definition 3.1; this result was obtained by using only standard techniques in the theory of MDP's with the average reward criterion. On the other hand, it would be interesting to extend the approach in this note to more general frameworks, for instance MDP's with Borel state space or satisfying conditions less restrictive than Simultaneous Doeblin (see [22]). Research in these directions is presently in progress.

Acknowledgement

The authors are grateful to Professor O. Hernández-Lerma for helpful comments.

Rolando Cavazos–Cadena
Departamento de Estadística y Cálculo
Universidad Autónoma Agraria
Antonio Narro
Buenavista, Saltillo COAH 25315
MÉXICO
rcavazos@uaaan.mx

Jean B. Lasserre
Laboratoire d'Automatique
et d'Analyse des Systèmes 7
Avenue du Colonel Roche
31077 Toulouse Cedex,
FRANCE
lasserre@laas.fr

References

[1] D. Blackwell, *Discrete Dynamic programming*, Ann. Math. Statist., **33** (1962), 719–726.

[2] R. Cavazos-Cadena and J. B. Lasserre, *Strong 1-optimal stationary policies in denumerable Markov decision processes*, Syst. Control Lett., **11** (1988), 65–71.

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

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

[5] R. Cavazos-Cadena, *Recent results on conditions for the existence of average optimal stationary policies*, Ann. Oper. Res., **26** (1991), 171–194.

[6] R. Deckker and A. Hordijk, *Average, sensitive and Blackwell optimal policies in denumerable Markov decision chains with unbounded rewards*, Math. Oper. Res., **13** (1988), 395–421.

[7] J. Dugundji, *Topology*, Allyn and Bacon, Boston 1966.

[8] E. Fernández-Gaucherand, M. K. Ghosh and S. I. Marcus, *Controlled Markov processes on the infinite planning horizont: weighted and overtaking cost criteria*, Zeit. Oper. Res., **39** (1994), 131–155.

[9] J. Flynn, *Conditions for the equivalence of optimaly criteria in dynamic programming*, Ann. Statist., **4** (1976), 936–953.

[10] M. K. Ghosh and S. I. Marcus, *On strong average optimality of Markov decision processes with unbounded costs*, Oper. Res. Lett., **11** (1992), 99–104.

[11] O. Hernández-Lerma, *Adaptive Markov Control Processes*, Springer-Verlag, New York, 1989.

[12] D.P. Heymann and M. J. Sobel, *Stochastic Models in Operations Research, Vol II: Stochastic Optimization*, McGraw-Hill, New York 1984.

[13] A. Hordijk and R. Dekker, *Denumerable Markov decision chains: sensitive optimality criteria*, Oper. Res. Proc., Springer-Verlag, Berlin, 1982.

[14] A. Hordijk and K. Sladky, *Sensitive optimality criteria in countable state dynamic programming*, Math. Oper. Res., **2** (1977), 1–14.

[15] J. B. Lasserre, *Conditions for existence of average and Blackwell optimal stationary policies in denumerable Markov decision processes*, J. Math. Anal. Appl., **136** (1988), 479–490.

[16] M. Loeve, *Probability Theory I*, Springer-Verlag, New York, 1977.

[17] P. Mandl, *Estimation and control in Markov chains*, Adv. Appl. Probab., **6** (1974), 40–60.

[18] B. L. Miller and A. F. Veinott Jr., *Discrete dynamic programming with a small interest rate*, Ann. Math. Statist., **40** (1969), 366–370.

[19] S. M. Ross, *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, 1970.

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

[21] A. F. Veinott Jr., *On discrete dynamic programming with sensitive discount optimality criteria*, Ann. Math. Statist., **40** (1969), 1635–1660.

[22] A.A. Yushkevich, *Blackwell optimality in Borelian continuous–in–action Markov decision processes*, SIAM J. Control Optim., **35** (1997), 2157–2182.