Journal of Mathematical Systems, Estimation, and Control
Vol. 6, No. 2, 1996, pp. 1–34

Undiscounted Value Iteration in Stable Markov Decision Chains with Bounded Rewards*

Rolando Cavazos-Cadena†

Abstract

This work considers denumerable state Markov Decision Chains endowed with a long-run expected average reward criterion and bounded rewards. Apart from standard continuity-compactness restrictions, it is supposed that the *Lyapunov Function Condition* for bounded rewards holds true; this assumption guarantees the existence of a (possibly) unbounded solution of the optimality equation yielding optimal stationary policies. In this context, it is shown that the relative value functions and differential rewards produced by the *Value Iteration* method converge pointwise to the solution of the optimality equation, and that it is possible to obtain a sequence of stationary policies whose limit points are optimal. These results extend those in [17], where it was assumed that the 'first error function' is bounded, and in [6], where weaker convergence results were obtained assuming that under the action of an arbitrary stationary policy the state space is a communicating class.

Key words: Markov decision processes, average reward criterion, Lyapunov function condition, value iteration method, pointwise convergence

AMS Subject Classifications: 93E20, 90C40

1 Introduction

This work considers Markov Decision Processes (MDP's) with denumerable state space, discrete time parameter, bounded rewards, and endowed

*Received January 4, 1994; received in final form August 4, 1994. Summary appeared in Volume 6, Number 2, 1996.

†This research was supported by the PSF Organization under Grant No. 500-350/93-08-93, by the MAXTOR Foundation for Applied Probability and Statistics (MAXFAPS) under Grant No. 01-01-56/08-93, and by the Consejo Nacional de Ciencia y Tecnología (CONACyT) under research Grant No. 1132-E9206.

with the (lim-sup) expected *average reward criterion*. Besides standard continuity-compactness conditions, the main restriction on the structure of the model is that the *Lyapunov Function Condition* (LFC)—introduced by Hordijk in [16]—holds true, an assumption that implies the existence of a (generally *unbounded*) solution of the average reward optimality equation ($AROE$) yielding optimal stationary policies. The following questions are addressed within this framework: Is it possible to use the *value iteration* (VI) procedure (a) to approximate the solution of the $AROE$? and (b) to determine a sequence of stationary policies whose limit points are optimal?

Of course, these are classical problems that have been extensively studied, for instance, in [6,8,12,17,21, 22,24] and references therein. Therefore, it is important to begin by pointing out the main difference between the results in this work and those already available in the literature, particularly in [6] and in the paper by Hordijk, Schweitzer and Tijms [17] where, assuming that the LFC holds, the VI method was used to provide an affirmative answer to the questions posed above. First, the VI procedure has been widely analized under several variants of the Simultaneous Doeblin Condition (SDC) [23], which allow to obtain very strong convergence results. For instance, under (simultaneous) scrambling the relative value functions produced by the VI method converge at a geometric rate to the solution of the $AROE$; for this and related results see Hernández-Lerma [12] and the interior references. However, SDC imposes very hard restrictions on the recurrence structure of the model [2,3] and is substantially stronger than LFC, the basic stability condition assumed in this note; in fact, LFC in Assumption 2.2 below can be safely classified as the weakest among the conditions presently available to guarantee the existence of optimal stationary policies for *arbitrary* continuous and bounded rewards, at least for the class of MDP's in which each stationary policy has a unique ergodic set of states. Concerning results obtained under LFC, in [17] it was supposed that the reward function is (*possibly*) unbounded, but that the 'first error function' is *bounded*; this assumption is not satisfied in many interesting applications and, when applied to the bounded rewards case, it implies that the solution of the $AROE$ is itself bounded, a condition that, in general, imposes strong restrictions on the model [2,3]. On the other hand, the results in [6] refer to the case of bounded rewards and the assumption of boundedness of the first error function was avoided at the expense of assuming that every stationary policy induces a communicating Markov chain. Then it was established that the the VI method allows to approximate the solution of the optimality equation *in the Cesàro sense*. In this paper the boundedness condition on the first error function as well as the communicating assumption are avoided, and it is shown that the relative value functions and differential rewards converge *pointwise* to the solution of the $AROE$. Thus, the results presented in Theorem 3.1 below can be

seen as an improved version of those in [17] applied to the bounded rewards framework, and in [6]; details on these comments are given in Remark 3.1.

On the other hand, it should be mentioned that following the ideas in [17], Sennott [22] used the *VI* procedure to obtain convergent approximations of the solution of the *AROE*, although some of the assumptions used there seem to be very difficult to verify, particularly Assumption 5. The recent paper by Montes-de-Oca and Hernández-Lerma [18] contains an extension of the results in [22] to the case of MDP's with Borel state and action spaces, as well as interesting comments about the diverse applications of the *VI* method; see also [1,13,14].

The remainder of the paper has been organized as follows: In Section 2 the decision model and *all* the structural assumptions are introduced, and then some basic consequences are presented in the form of Lemmas 2.1 and 2.2, including uniqueness of the solution of the *AROE*. Next, in Section 3 the *VI* procedure is briefly described and the answer to the two problems considered above is stated in the form of Theorem 3.1. The necessary technical preliminaries to establish this result are contained in Sections 4 and 5, and then Theorem 3.1 is proved in Section 6. Finally, the paper concludes with some brief comments in Section 7.

Notation. As usual $\mathbb{R}$ and $\mathbb{N}$ stand for the sets of real numbers and nonnegative integers, respectively. Given an event W, the corresponding indicator function is denoted by $I[W]$. If $\mathbb{K}$ is a topological space, $\mathbb{B}(\mathbb{K})$ is the space of all continuous and bounded functions $r : \mathbb{K} \to \mathbb{R}$ endowed with the supremum norm:

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

Finally, a cartesian product of topological spaces is always endowed with the corresponding product topology and, for $a, b \in \mathbb{R}$, $a \wedge b := \min\{a, b\}$.

2 Decision Model and Basic Results

Let $(S, A, \{A(x)|x \in S\}, r, p)$ be the usual MDP where the *state space* S is a denumerable set endowed with the discrete topology, and the metric space A is the *action set*. For each $x \in S$, $A(x) \subset A$ stands for the nonempty set of admissible actions at state x, whereas the set of *admissible state–action pairs* is $\mathbb{K} := \{(x, a)|x \in S,\ a \in A(x)\}$, which is considered as a topological subspace of $S \times A$. On the other hand, $r : \mathbb{K} \to \mathbb{R}$ is the reward function and p is the transition law. The interpretation of this model is as follows: At each time $t \in \mathbb{N}$ the state of a dynamical system is observed, say $X_t = x \in S$, and an action $A_t = a \in A(x)$ 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 $X_{t+1} = y \in S$ with probability $p_{x\,y}(a)$; this is the Markov property of the decision process.

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

(ii)*For each* $x, y \in S$, *the mapping* $a \mapsto p_{x\,y}(a)$ *is continuous in* $a \in A(x)$.

(iii) $r \in \mathbb{B}(\mathbb{K})$.

Control Policies For $k \in \mathbb{N}$ the information vector up to time k is denoted by I_k and is defined by

$$I_0 := X_0, \quad \text{whereas} \quad I_k := (X_0, A_0, \cdots, X_{k-1}, A_{k-1}, X_k), \quad k > 0. \tag{2.1}$$

On the other hand, for each $t \in \mathbb{N}$, $h_t := (x_0, a_0, \cdots, x_{t-1}, a_{t-1}, x_t)$ denotes an admissible history of the process up to time t; this means that $x_k \in S$ for all $k \leq t$ and $a_k \in A(x_k)$ if $k < t$. A policy $\pi = \{\pi_t\}$ is a (possibly randomized) measurable rule for choosing actions which may depend on the current state as well as on the record of previous states and actions. If π is the policy being used and B is a Borel subset of A, the probability of the event $[A_t \in B]$ given $I_t = h_t$ is $\pi_t(B|h_t)$, where $\pi_t(A(x_t)|h_t) = 1$ is always valid; for details see, for instance, [12 pp. 1–4] or [15]. The class of all policies is denoted by $\mathbb{P}$. Given the initial state and the policy π being used the distribution of the state–action process $\{(X_t, A_t)\}$ is uniquely determined and is denoted by P_x^π, while E_x^π stands for the corresponding expectation operator. Now set $\mathbb{F} := \Pi_{x \in S} A(x)$, that is, $\mathbb{F}$ consists of all (choice) functions $f : S \to A$ satisfying that $f(x) \in A(x)$ for all $x \in S$; notice that $\mathbb{F}$ is a compact metric space in the product topology [7]. A policy π is stationary if there exists $f \in \mathbb{F}$ such that when the system is in progress under π, action $f(x)$ is applied whenever the observed state is $X_t = x$ regardless of $t \in \mathbb{N}$. Next, define $\mathbb{M} := \Pi_{n \in \mathbb{N}} \mathbb{F}$, i.e., the (compact metric) space $\mathbb{M}$ consists of all sequences $\{f_t\}$ with $f_t \in \mathbb{F}$ for all $t \in \mathbb{N}$. A policy $\pi \in \mathbb{P}$ is *Markov* if there exists $\{f_t\} \in \mathbb{M}$ for which the following occurs when the system evolves under π: for each time $t \in \mathbb{N}$, the action prescribed by π at time t is $f_t(X_t)$. The class of stationary (resp. Markov) policies is naturally identified with $\mathbb{F}$ (resp. $\mathbb{M}$) and, with these conventions, it is clear that $\mathbb{F} \subset \mathbb{M} \subset \mathbb{P}$.

Performance Index The (lim-sup expected) *average reward* at state $x \in S$ under policy π is defined by

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

whereas

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

is the optimal average reward at state x. A policy π is *average optimal* (AO) if $J(x, \pi) = J(x)$ for all $x \in S$.

Optimality Equation To establish the existence of optimal stationary policies it is necessary to complement Assumption 2.1 with an additional condition [20]. In Assumption 2.2 below an appropriate restriction on the transition structure of the model is given so that the existence of an optimal stationary policy is guaranteed. First, let $z \in S$ be a given state, *fixed* throughout the remainder of the paper, and define the first passage time T as follows:

$$T := \min\{n > 0 | X_n = z\}, \tag{2.4}$$

where the usual convention that the minimum of the empty set is ∞ is enforced.

Assumption 2.2 *(LFC for bounded rewards [16,17; 4,6,23].) There exists $l : S \to [0, \infty)$ satisfying the following conditions (i)-(iii); such a function is referred to as a Lyapunov function for bounded rewards.*

(i) $1 + \sum_{y \neq z} p_{x\,y}(a) l(y) \leq l(x), \quad x \in S,\ a \in A(x)$.

(ii) For each $x \in S$, the mapping

$$f \mapsto \sum_{y \neq z} p_{x\,y}(f(x)) l(y) = E_x^f [l(X_1) I[T > 1]]$$

is continuous in $f \in \mathbb{F}$.

(iii) For each $f \in \mathbb{F}$ and $x \in S$, $E_x^f [l(X_n) I[T > n]] \to 0$ as $n \to \infty$.

Under Assumptions 2.1 and 2.2 the *average reward optimality equation* $(AROE)$ given by (2.5) below has a solution yielding an optimal stationary policy.

Lemma 2.1 *Suppose that Assumptions 2.1 and 2.2 hold true. Then there exist $h : S \to \mathbb{R}$ and $g \in \mathbb{R}$ such that (i)-(iv) below occur.*

(i) $g = J(x)$ for each $x \in S$; see (2.2) and (2.3).

(ii) $h(z) = 0$ and $|h(x)| \leq 2\|r\| \cdot l(x), \quad x \in S$.

(iii) The AROE is satisfied by g and $h(\cdot)$, that is,

$$g + h(x) = \sup_{a \in A(x)} [r(x, a) + \sum_y p_{x\,y}(a) h(y)], \quad x \in S. \tag{2.5}$$

(iv) An optimal stationary policy exists. Furthermore, for each $x \in S$ the right hand side of (2.5)—considered as a function of $a \in A(x)$—has a maximizer $f^*(x)$, and the corresponding policy $f^* \in \mathbb{F}$ is optimal.

A proof of this result can be found in [16, chapter 5]. Notice that g in this lemma is *uniquely determined*, since it is *the* optimal average reward at every state. The function h is also unique, as established in part (iv) of the following lemma.

Lemma 2.2 *Under Assumptions 2.1 and 2.2 the following assertions (i)-(iv) are satisfied.*

(i) For each $\pi \in \mathbb{P}$, $x \in S$ and $n \in \mathbb{N}$,

$$E_x^\pi[\sum_{t=0}^{n} I[T > t] + l(X_{n+1})I[T > n+1]] \leq l(x); \tag{2.6}$$

consequently,

$$E_x^\pi[T] \leq l(x). \tag{2.7}$$

Now define

$$\Delta_n(x) := \sup_{\pi \in \mathbb{P}} E_x^\pi[l(X_n)I[T > n]], \quad x \in S, \ n \in \mathbb{N}. \tag{2.8}$$

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

$$\lim_{n \to \infty} \Delta_n(x) = 0 \tag{2.9}$$

and

$$\frac{1}{n+1} \sup_{\pi \in \mathbb{P}} E_x^\pi[l(X_{n+1})] \to 0 \quad \text{as} \quad n \to \infty. \tag{2.10}$$

(iii) Let $f \in \mathbb{F}$ and $c \in [0, \infty)$ be arbitrary but fixed.

(a) Suppose that $U : S \to \mathbb{R}$ satisfies

$$|U(x)| \leq c \cdot l(x) \quad \text{and} \quad U(x) \leq \sum_y p_{x\,y}(f(x))U(y), \quad x \in S. \tag{2.11}$$

Then $U(x) \leq U(z)$ for all $x \in S$.

Similarly,

(b) If $L : S \to \mathbb{R}$ is such that for all $x \in S$

$$|L(x)| \leq c \cdot l(x) \quad \text{and} \quad L(x) \geq \sum_y p_{x\,y}(f(x))L(y),$$

then $L(x) \geq L(z)$ *for all* $x \in S$.

(iv) Suppose that $h_1, h_2 : S \to \mathbb{R}$ *satisfy (a)-(c) below.*

(a) $h_1(z) = h_2(z) = 0$;

(b) For some $c \in [0, \infty)$, $|h_i(x)| \leq c \cdot l(x), \quad x \in S, \quad i = 1, 2,$ *and*

(c) For $i = 1, 2,$

$$g + h_i(x) = \sup_{a \in A(x)} [r(x, a) + \sum_y p_{x\,y}(a) h_i(y)], \quad x \in S. \tag{2.12}$$

Then $h_1 = h_2$.

Proof: *(i)* Notice that, by (2.1) and (2.4), for each $k \in \mathbb{N}$ the event $[T > k]$ is I_k-measurable and $I[T > k+1] = I[T > k] I[X_{k+1} \neq z]$. Then for each $x \in S$, $\pi \in \mathbb{P}$ and $k \in \mathbb{N}$

$$\begin{aligned} &E_x^\pi [I[T > k] + l(X_{k+1}) I[T > k+1] | I_k, A_k] \\ &\qquad = E_x^\pi [I[T > k](1 + l(X_{k+1}) I[X_{k+1} \neq z]) | I_k, A_k] \\ &\qquad = I[T > k](1 + \sum_{y \neq z} p_{X_k\,y}(A_k) l(y)) \\ &\qquad \leq I[T > k] l(X_k), \end{aligned}$$

where the second equality follows from the Markov property and the inequality is due to Assumption 2.2(i). Then, taking expectation with respect to P_x^π it follows that

$$E_x^\pi [I[T > k] + l(X_{k+1}) I[T > k+1]] \leq E_x^\pi [I[T > k] l(X_k)]. \tag{2.13}$$

Inequality (2.6) can now be established by induction as follows: For $n = 0$ the assertion is equivalent to Assumption 2.2(i). Now suppose that (2.6) occurs for $n = k - 1 \in \mathbb{N}$. In this case

$$\begin{aligned} &E_x^\pi [\sum_{t=0}^{k} I[T > t] + l(X_{k+1}) I[T > k+1]] \\ &\qquad = E_x^\pi [\sum_{t=0}^{k-1} I[T > t]] + E_x^\pi [I[T > k] + l(X_{k+1}) I[T > k+1]] \\ &\qquad \leq E_x^\pi [\sum_{t=0}^{k-1} I[T > t]] + E_x^\pi [l(X_k) I[T > k]] \quad \text{(by(2.13))} \end{aligned}$$

and then the induction hypothesis yields

$$E_x^\pi [\sum_{t=0}^{k} I[T > t] + l(X_{k+1}) I[T > k+1]] \leq l(x),$$

which is (2.6) with $n = k$. Finally, since $l \geq 0$, (2.6) implies

$$E_x^\pi[T] = \lim_{n\to\infty} \sum_{k=0}^{n} E_x^\pi[I[T > k]] \leq l(x).$$

(*ii*) Convergence (2.9) was obtained in [16]; see *the proof* of equation 5.7.2 in [16, pp. 43-44]. To establish (2.10) observe that

$$\begin{aligned}
&\textstyle\sum_{k=1}^{n+1} I[X_k = z, X_t \neq z, k < t \leq n+1] + I[X_t \neq z, 1 \leq t \leq n+1] \\
&\qquad = I[X_k = z \text{ for } \textit{some } t \in \{1, 2, \ldots, n+1\}] \\
&\qquad\quad + I[X_k \neq z \text{ for } \textit{all } t \in \{1, 2, \ldots, n+1\}] \\
&\qquad = 1,
\end{aligned}$$

so that

$$\begin{aligned}
E_x^\pi[l(X_{n+1})] = &\sum_{k=1}^{n+1} E_x^\pi[l(X_{n+1})I[X_k = z, X_t \neq z, k < t \leq n+1]] \\
&+ E_x^\pi[l(X_{n+1})I[X_k \neq z, 1 \leq k \leq n+1]].
\end{aligned} \tag{2.14}$$

Next, note that for each positive integer $k \leq n+1$ (see (2.1) for the definition of I_k)

$$\begin{aligned}
&E_x^\pi[l(X_{n+1})I[X_k = z, X_t \neq z, k < t \leq n+1]|I_k] \\
&\qquad = I[X_k = z]E_z^{\pi'}[l(X_{n+1-k})I[X_t \neq z, 0 < t \leq n+1-k]] \\
&\qquad = I[X_k = z]E_z^{\pi'}[l(X_{n+1-k})I[T > n+1-k]] \\
&\qquad \leq I[X_k = z]\Delta_{n+1-k}(z)
\end{aligned} \tag{2.15}$$

where the 'shifted' policy π' is determined by

$$\pi'_t(\cdot|h_t) = \pi_{t+k}(\cdot|X_0, A_0, \cdots, X_{k-1}, A_{k-1}, h_t)$$

(see [12, p. 5]) and (*a*) the first equality follows from the Markov property, (*b*) the second equality is due to the definition of T in (2.4), and (*c*) the definition of $\Delta_{n+1-k}(z)$ in (2.8) was used to obtain the inequality. Taking expectations with respect to P_x^π, (2.15) yields

$$\begin{aligned}
&E_x^\pi[l(X_{n+1})I[X_k = z, X_t \neq z, k < t \leq n+1]] \\
&\qquad \leq P_x^\pi[X_k = z]\Delta_{n+1-k}(z) \leq \Delta_{n+1-k}(z).
\end{aligned} \tag{2.16}$$

To conclude observe that, by (2.4) and (2.8),

$$E_x^\pi[l(X_{n+1})I[X_t \neq z, 1 \leq t \leq n+1]] = E_x^\pi[l(X_{n+1})I[T > n+1] \leq \Delta_{n+1}(x),$$

which combined with (2.14) and (2.16) implies

$$\frac{E_x^\pi[l(X_{n+1})]}{n+1} \leq \frac{\Delta_{n+1}(x)}{n+1} + \frac{\sum_{k=1}^{n+1} \Delta_{n+1-k}(z)}{n+1},$$

and then (2.10) follows from (2.9).

(*iii*) First consider part (*a*). Set $d(x) := U(x) - U(z)$, $x \in S$ and notice that the second equality in (2.11) implies that

$$\begin{aligned} d(x) &\leq \sum_y p_{x\,y}(f(x))d(y) \\ &= \sum_{y \neq z} p_{x\,y}(f(x))d(y) \\ &= E_x^f[d(X_1)I[T > 1]], \quad x \in S, \end{aligned}$$

where $d(z) = 0$ and the definition of T in (2.4) were used to obtain the equalities. Then a simple induction argument yields

$$d(x) \leq E_x^\pi[d(X_n)I[T > n]], \quad x \in S,\ n \in \mathbb{N}. \tag{2.17}$$

To conclude observe that $d(x) \leq |U(x)| + |U(z)| \leq c \cdot l(x) + |U(z)| \leq (c + |U(z)|)l(x)$, where $l(\cdot) \geq 1$ was used to obtain the third inequality; see Assumption 2.2(*i*). Then (2.17) implies that for all $x \in S$

$$d(x) \leq (c + |U(z)|)E_x^f[l(X_n)I[T > n]] \to 0 \ \text{ as } \ n \to \infty,$$

where the convergence follows from Assumption 2.2(*iii*) (or from (2.9)). Therefore, $U(x) - U(z) = d(x) \leq 0$ for all $x \in S$, which is the desired conclusion. Finally, (*b*) follows by setting $U \equiv -L$ and applying part (*a*).

(*iv*) Combining (2.12) with Lemma 3.3 in [15] it follows that

$$|h_1(x) - h_2(x)| \leq \sup_{a \in A(x)} [\sum_y p_{x\,y}(a)|h_1(y) - h_2(y)|], \quad x \in S. \tag{2.18}$$

Now observe that for each $x \in S$, the mapping

$$a \mapsto \sum_y p_{x\,y}(a)l(y) = \sum_{y \neq z} p_{x\,y}(a)l(y) + p_{x\,z}(a)l(z), \quad a \in A(x),$$

is continuous, by Assumptions 2.1 and 2.2, and that by condition (b) in the statement of the theorem,

$$|h_1(x) - h_2(x)| \leq 2c \cdot l(x), \quad x \in S. \tag{2.19}$$

These facts imply, via Proposition 18 in [19 p. 232], that for each $x \in S$, the mapping

$$a \mapsto \sum_y p_{x\,y}(a)|h_1(y) - h_2(y)|$$

is continuous in $a \in A(x)$. Since each set $A(x)$ is compact, there exists $f \in \mathbb{F}$ such that

$$\sum_y p_{x\,y}(f(x)|h_1(y) - h_2(y)| = \sup_{a \in A(x)} [\sum_y p_{x\,y}(a)|h_1(y) - h_2(y)|], \quad x \in S,$$

and in combination with (2.18) this equality yields

$$|h_1(x) - h_2(x)| \leq \sum_y p_{x\,y}(f(x))|h_1(y) - h_2(y)|, \quad x \in S.$$

Then, setting $U \equiv |h_1 - h_2|$, part (iii) (a) and (2.19) together imply that

$$|h_1(x) - h_2(x)| \leq |h_1(z) - h_2(z)| = 0, \quad x \in S$$

where condition (a) in the statement of the theorem was used to obtain the equality. □

As already mentioned, the main objective of this work is to apply the *VI* procedure to approximate the solution $(g, h(\cdot))$ of the *AROE*. The precise result in this direction, stated as Theorem 3.1 in the following section, requires an additional condition on the transition structure of the model which is now introduced.

Assumption 2.3 *For each* $a \in A(z)$, $p_{z\,z}(a) > 0$.

Remark 2.1 It is interesting to observe that Assumption 2.3 does *not* imply any loss of generality, since it can be obtained by making an appropriate transformation on the transition law. In fact, suppose that $M = (S, A, \{A(x)|x \in S\}, r, p)$ satisfies Assumptions 2.1 and 2.2 and define the transformed transition law p^* as follows:

$$p^*_{x\,y}(a) := (1 - \alpha)\delta_{x\,y} + \alpha \cdot p_{x\,y}(a), \quad (x, a) \in \mathbb{K},\ y \in S,$$

where $\alpha \in (0,1)$ is a given number and $\delta_{x\,y} := 1$ (resp. 0) if $x = y$ (resp. $x \neq y$). Now set $M^* := (S, A, \{A(x)|x \in S\}, r, p^*)$, which clearly satisfies Assumptions 2.1 and 2.3. Moreover, it is not difficult to see that $l^*(\cdot) := l(\cdot)/\alpha$ is a Lyapunov function for M^*, so that Assumption 2.2 is also satisfied by M^*. The models M and M^* are *equivalent*, in the following sense: Let the pair (g, h) be the solution to the *AROE* for model M and let (g^*, h^*) be the corresponding pair for model M^*. Then (a) $g^* = g$, (b) $h^* = h/\alpha$, and (c) A policy $f \in \mathbb{F}$ is optimal for model M if and only if f is optimal for M^*. The transformation $p \mapsto p^*$ was introduced by Schweitzer in [21].

Remark 2.2 Throughout the remainder Assumptions 2.1–2.3 are supposed to hold true, even without explicit reference. On the other hand, Assumption 2.3 will be used only in one place, namely, in the the proof of part (ii) of Lemma 5.1.

3 Value Iteration and Main Theorem

In this section the main result of this note is presented in the form of Theorem 3.1 below. To begin with, the necessary notions are introduced.

Definition 3.1 *(The VI Method.)*

(i) The sequence $\{V_n : S \to \mathbb{R} \,|\, n = -1, 0, 1, \cdots\}$ *of value iteration functions is recursively defined as follows:* $V_{-1} \equiv 0$ *and, for* $n \geq 0$,

$$V_n(x) = \sup_{a \in A(x)} [r(x,a) + \sum_y p_{x\,y}(a) V_{n-1}(y)], \quad x \in S.$$

(ii) The relative value functions $R_n : S \to \mathbb{R}$ *are defined by*

$$R_n(x) := V_n(x) - V_n(z), \quad x \in S, \ n = -1, 0, 1, 2, \cdots.$$

(iii) For each $x \in S$ *and* $n \in \mathbb{N}$ *define the nth differential reward at* x *by*

$$g_n(x) := V_n(x) - V_{n-1}(x).$$

It is known that for all $n \in \mathbb{N}$ there exists a policy $\pi^n \in \mathbb{M}$ such that [12,20]

$$\begin{aligned} V_n(x) &= E_x^{\pi^n}[\sum_{t=0}^{n} r(X_t, A_t)] \\ &= \sup_{\pi \in \mathbb{P}} E_x^{\pi}[\sum_{t=0}^{n} r(X_t, A_t)], \quad x \in S, \end{aligned} \tag{3.1}$$

and then it is clear that

$$|V_n(\cdot)| \leq (n+1)\|r\|. \tag{3.2}$$

Also, combining (3.1) with Lemma 3.3 in [15] it follows that

$$|V_{n+k}(x) - V_n(x)| \leq \sup_{\pi\in\mathbb{P}} |E_x^\pi[\sum_{t=n+1}^{n+k} r(X_t, A_t)]| \leq k\|r\|, \quad n, k \in \mathbb{N}, \tag{3.3}$$

which yields

$$\|g_n(\cdot)\| \leq \|r\|, \quad n \in \mathbb{N}. \tag{3.4}$$

The following theorem, showing that the *VI* method can be used to approximate the solution $(g, h(\cdot))$ of the *AROE* (2.5), is the main result of this note.

Theorem 3.1 *Under Assumptions 2.1–2.3, (i)-(iv) below occur.*

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

Moreover,

(ii) For all $x \in S$

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

(iii) For each $x \in S$,

$$\lim_{n\to\infty} R_n(x) = h(x).$$

(iv) Given $n \in \mathbb{N}$ *there exists a policy* $f_n \in \mathbb{F}$ *such that, for each* $x \in S$, $f_n(x)$ *is a maximizer of the mapping*

$$a \mapsto r(x,a) + \sum_y p_{x\,y}(a) R_n(y), \quad a \in A(x).$$

Furthermore, every limit point of $\{f_n\} \subset \mathbb{F}$ *is optimal.*

The proof of this result is contained in Section 6. Unfortunately, we have not been able neither of finding a direct way to establish this theorem nor of adapting the arguments used in [6] and [17] to the framework of Assumptions 2.1–2.3. Rather, the proof of Theorem 3.1 given below is somewhat technical and is based on the preliminaries presented in the following two sections. Before going any further, the relation of Theorem 3.1 with other results in the literature is discussed in some detail.

Remark 3.1 The main differences between Theorem 3.1 and the results in [17], applied to the bounded rewards case, and in [6], are as follows:

(*i*) In [17] the conclusions in Theorem 3.1 are shown to occur but, in addition to Assumptions 2.1–2.3, the following condition is supposed to hold true:

The first 'error function' $e(\cdot) := R_0(\cdot) - [g + h(\cdot)]$ is bounded,

that is, $\|e\| < \infty$. Since $\|R_0\| = \|V_0(\cdot) - V_0(z)\| \leq 2\|r\|$ (see (3.2)), $\|e\| < \infty$ implies that $h(\cdot)$ is a bounded function, a condition that is violated in interesting applications (see Example 3.1 below).

On the other hand, it should be emphasized that, in general, the assumptions in Section 2 do *not* guarantee that the first error function e defined above is bounded for arbitrary reward function $r \in \mathbb{B}(\mathbb{K})$. This can be seen as follows:

$$\begin{aligned} \text{For all } r \in \mathbb{B}(\mathbb{K}), \quad \|e\| < \infty \Rightarrow\ & \text{For all } r \in \mathbb{B}(\mathbb{K}), \\ & h(\cdot) \text{ in the } AROE \text{ is bounded} \\ & \text{(as shown above)} \\ \Rightarrow\ & SDC \text{ holds (see [2])} \\ \Rightarrow\ & \sup_{x \in S, f \in \mathbb{F}} E_x^f[T] < \infty \end{aligned}$$

where the last implication follows from the fact that, by Assumption 2.2, state z is positive recurrent under arbitrary $f \in \mathbb{F}$. However, under Assumption 2.2, $E_x^f[T]$ is always *finite* but, in general, *not* a bounded function of $(x, f) \in S \times \mathbb{F}$; see Example 3.1 below.

(*ii*) In [6] it was supposed that (*a*) Under the action of an arbitrary stationary policy the state process $\{X_t\}$ is a communicating chain, that is, given $f \in \mathbb{F}$ and $x, y \in S$ there exists $n \equiv n(x, y, f)$ such that $P_x^f[X_n = y] > 0$, and (*b*) Assumptions 2.1 and 2.2 hold true. Within this framework, convergences weaker than those in Theorem 3.1 were obtained, namely, it was shown in [6] that $\{g_n(z)\}$ and $\{R_n(\cdot)\}$ converge in the *Cesàro sense* to g and $h(\cdot)$, respectively, i.e.,

$$\lim_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} g_k(z) = g, \quad \text{and} \quad \lim_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} R_k(x) = h(x), \quad x \in S.$$

Example 3.1 Let A be a finite set endowed with the discrete topology and let $\{U_{n\,a}, D_{n\,a} | n \in \mathbb{N}, a \in A\}$ be a collection of independent $\mathbb{N}$-valued random variables such that

(*i*) $P[U_{n\,a} = k] = q_a(k), \quad n, k \in \mathbb{N}, a \in A$; i.e., for each $a \in A$, the random variables $\{U_{n\,a} | n \in \mathbb{N}\}$ are identically distributed with common distribution $\{q_a(k) | k \in \mathbb{N}\}$, and

(*ii*) $P[D_{n\,a} = 1] = \mu_a = 1 - P[D_{n\,a} = 0]$.

These random variables are interpreted as the arriving and service streams in a single–server queueing model with state space $S = \mathbb{N}$, action space A and $A(x) \equiv A$ for all $x \in \mathbb{N}$. Let $X_n = x \in \mathbb{N}$ be the number of customers waiting for service at the beginning of the time period $[n, n+1)$. If $x > 0$ and action $A_n = a \in A(x)$ is applied, then the number of arrivals in $[n, n+1)$ is $U_{n\,a}$, whereas the server can provide a complete service with probability μ_a; $D_{n\,a}(= 0, 1)$ is interpreted as the number of customers leaving the system after service completion in period $[n, n+1)$. These considerations can be summarized in the following evolution equation:

$$X_{n+1} = X_n + U_{n\,a} - D_{n\,a} \quad \text{if} \quad X_n > 0 \quad \text{and} \quad A_n = a; \tag{3.5}$$

when $X_n = 0$ the server stays idle in the period $[n, n+1)$ so that

$$X_{n+1} = U_{n\,a} \quad \text{if} \quad X_n = 0 \quad \text{and} \quad A_n = a.$$

From these equations the transition law is determined by

$$\begin{aligned} p_{x\,y}(a) &= (1-\mu_a)q_a(y-x) + \mu_a q_a(y-x-1), \quad x, y \in \mathbb{N}, \quad x > 0 \\ &= q_a(y), \end{aligned} \tag{3.6}$$

where $q_a(s) := 0$ for $s < 0$. If a reward function $r \in \mathbb{B}(\mathbb{N} \times A) \equiv \mathbb{B}(\mathbb{K})$ is chosen, then the finiteness of A implies that Assumption 2.1 holds. To verify the other assumptions the following additional condition is imposed:

$$\lambda_a < \mu_a \quad a \in A, \tag{3.7}$$

where $\lambda_a := \sum_{k=1}^{\infty} k q_a(k)$, $a \in A$. Notice that (3.7) yields that $\lambda_a < 1$ for all action a, and this in turn implies that

$$q_a(0) > 0, \quad a \in A. \tag{3.8}$$

Next set $z := 0$. In this case (3.6) and (3.8) together show that Assumption 2.3 holds with $z = 0$. To verify Assumption 2.2 define $C, B \in (0, \infty)$ by

$$C := \max_{a \in A}\{(\mu_a - \lambda_a)^{-1}\} \quad \text{and} \quad B := \max_{a \in A}\{q_a(0)^{-1}(1 + C\lambda_a)\},$$

and set

$$l(x) := Cx + B, \quad x \in \mathbb{N}.$$

In this case straightforward calculations using (3.6) yield that

$$1 + \sum_{y \neq z} p_{x\,y}(a) l(y) \leq l(x), \quad x \in \mathbb{N}, \tag{3.9}$$

which is the first condition of Assumption 2.2, whereas the second one follows from the finiteness of the action set. To verify part (*iii*) of Assumption 2.3, let $f \in \mathbb{F}$ be arbitrary but fixed. From the evolution equation (3.5), $X_{n+1} \geq X_n - 1$ if $X_n > 0$, which immediately yields that $T \geq x$, P_x^f–almost surely for all $x \in \mathbb{N}$, so that

$$E_x^f[T] \geq x, \quad x \in \mathbb{N}; \tag{3.10}$$

see (2.4) for the definition of T. Now observe that (3.9) implies that $E_x^f[T] \leq l(x)(< \infty)$ (see the proof of Lemma 2.2(*i*)). In particular,

$$P_x^f[T > k] \to 0 \quad \text{as} \quad k \to \infty \tag{3.11}$$

and, by the dominated convergence theorem,

$$E_x^f[(T-k)I[T > k]] = E_x^f[T - T \wedge k] \to 0 \quad \text{as} \quad k \to \infty.$$

On the other hand, the Markov property and the definition of T yield that

$$E_x^f[(T-k)I[T > k]|I_k] = I[T > k]E_{X_k}^f[T] \geq I[T > k]X_k$$

(see (2.1), (2.4) and (3.10)) and then the last displayed convergence yields

$$E_x^f[X_k I[T > k]] \leq E_x^f[(T-k)I[T > k]] \to 0 \quad \text{as} \quad k \to \infty. \tag{3.12}$$

To conclude observe that

$$E_x^f[l(X_k)I[T > k]] = C \cdot E_x^f[X_k I[T > k]] + B \cdot P_x^f[T > k] \to 0 \quad \text{as} \quad k \to \infty,$$

by (3.11) and (3.12), and this is precisely the third part of Assumption 2.2. In short, the assumptions in Section 2 are satisfied in this example and, by Theorem 3.1, the differential rewards and relative value functions produced by the *VI* method converge to the solution (g, h) of the *AROE* for arbitrary reward function $r \in \mathbb{B}(\mathbb{K})$. On the other hand, it has been shown that $x \leq E_x^f[T]$, $x \in S$, so that $E_x^f[T]$ is not a bounded function of (x, f); by the comments in Remark 3.1(*i*), this shows that for some $r \in \mathbb{B}(\mathbb{K})$ the first error function is not bounded, so that the convergences in Theorem 3.1 can not be obtained from the results in [17]. For instance, if for $a \in A$, $r(x, a) := 1$ for $x \neq 0$ and $r(0, a) := 0$, it is not difficult to see that the first error function and the function $h(\cdot)$ in the *AROE* are both unbounded.

4 Preliminaries: First Part

This section starts the journey to the proof of Theorem 3.1. To begin with, some useful notation is introduced.

Definition 4.1

(i) The functions $U, L : S \to \mathbb{R}$ are defined as follows: For each $x \in S$,

$$U(x) := \limsup_{n\to\infty} g_n(x), \quad \text{and} \quad L(x) := \liminf_{n\to\infty} g_n(x).$$

(ii) For each $k \in \mathbb{N}$ define $b_k := \sup_{\pi\in\mathbb{P}} E_z^\pi[TI[T > k]]$.

Remark 4.1 Notice that $\|U\|,\ \|L\| \leq \|r\|$, by (3.4).

The main objective in this section is to establish the following result, which is the first essential component of the proof of Theorem 3.1 presented in Section 6; see Remark 2.2.

Theorem 4.1 *There exists a policy $\pi \in \mathbb{M}$ such that*

$$L(z) + \frac{[U(z) - L(z)]}{E_z^\pi[T]} \leq g; \tag{4.1}$$

recall that g is the optimal average reward.

The proof of this theorem has been divided into several pieces given below in the form of Lemmas 4.1–4.3; the arguments in the proofs of these preliminaries are along the ideas used in [4] and [5]. The first two lemmas refer to continuity properties derived from Assumptions 2.1 and 2.2.

Lemma 4.1
(i) $\lim_{k\to\infty} b_k = 0$.
(ii) For each $k \in \mathbb{N}$ and $x \in S$, the mappings

$$\pi \mapsto E_x^\pi[r(X_k, A_k)I[T > k]], \quad \text{and} \quad \pi \mapsto P_x^\pi[T = k]$$

are continuous in $\pi \in \mathbb{M}$.

(iii) The function $\pi \mapsto E_z^\pi[T]$, $\pi \in \mathbb{M}$, is continuous.

(iv) Suppose that the sequence $\{\pi_n\} \subset \mathbb{M}$ converges to π $(\in \mathbb{M})$ as $n \to \infty$. Then

$$\lim_{n\to\infty} E_z^{\pi_n}[\sum_{t=0}^{T\wedge n-1} r(X_t, A_t)] = E_z^\pi[\sum_{t=0}^{T-1} r(X_t, A_t)].$$

Proof: (i) Let $\pi \in \mathbb{P}$ be arbitrary. Since the event $[T > k]$ is I_k-measurable, the Markov property implies that $E_x^\pi[TI[T > k]|I_k] = I[T > k]E_{X_k}^{\pi'}[T] \leq I[T > k]l(X_k)$, where (2.7) was used to obtain the inequality and the shifted policy π' is determined by

$$\pi'_t(\cdot|h_t) = \pi_{t+k}(\cdot|X_0, A_0, \ldots, X_{k-1}, A_{k-1}, h_t);$$

cf. the proof of Lemma 2.2 (*ii*). Then, it follows that

$$E_x^\pi[TI[T>k]] \le E_x^\pi[I[T>k]l(X_k)],$$

so that $0 \le b_k \le \Delta_k(z) \to 0$ as $k \to \infty$; see (2.8) and (2.9).

(*ii*) This part follows from an induction argument using Assumptions 2.1 and 2.2 and Proposition 2.18 in [19 p. 232].

(*iii*) For each positive integer k, $E_z^\pi[T \wedge k] = k + \sum_{r=0}^{k-1}(r-k)P_z^\pi[T=r]$ is a continuous function of $\pi \in \mathbb{M}$, by part (*ii*). On the other hand, using that $T - T \wedge k \le (T-k)I[T>k] \le TI[T>k]$ it follows that

$$\sup_{\pi\in\mathbb{P}} |E_z^\pi[T] - E_z^\pi[T\wedge k]| \le \sup_{\pi\in\mathbb{P}} E_z^\pi[TI[T>k]] = b_k \to 0 \quad \text{as} \quad k\to\infty,$$

by part (*i*). Therefore, being a uniform limit of continuous functions, the mapping $\pi \mapsto E_z^\pi[T],\ \pi \in \mathbb{M}$, is itself continuous.

(*iv*) Let $k \in \mathbb{N}$ be fixed, select $n \in \mathbb{N} \cup \{\infty\}$ satisfying $k < n$, and observe that for each $\delta \in \mathbb{P}$,

$$|E_z^\delta[\sum_{t=0}^{T\wedge n-1} r(X_t,A_t)] - E_z^\delta[\sum_{t=0}^{T\wedge k-1} r(X_t,A_t)]| \le E_z^\delta[\sum_{t=T\wedge k}^{T\wedge n-1} |r(X_t,A_t)|].$$

Since $0 \le T\wedge n - T\wedge k \le T - T\wedge k \le (T-k)I[T>k] \le TI[T>k]$, it follows that

(*a*)

$$|E_z^\delta[\sum_{t=0}^{T\wedge n-1} r(X_t,A_t)] - E_z^\delta[\sum_{t=0}^{T\wedge k-1} r(X_t,A_t)]| \le \|r\|\cdot E_z^\delta[T\wedge n - T\wedge k] \le \|r\|\cdot b_k.$$

In particular, setting $n = \infty$,

(*b*)
$$|E_z^\delta[\sum_{t=0}^{T-1} r(X_t,A_t)] - E_z^\delta[\sum_{t=0}^{T\wedge k-1} r(X_t,A_t)]| \le \|r\|\cdot b_k.$$

On the other hand, part (*ii*) above yields that

(*c*) $\delta \mapsto E_z^\delta[\sum_{t=0}^{T\wedge k-1} r(X_t,A_t)] = E_z^\delta[\sum_{t=0}^{k-1} r(X_t,A_t)I[T>t]], \quad \delta \in \mathbb{M}$, is a continuous mapping .

To conclude observe that (*a*) and (*b*) together with the triangle inequality yield, via straightforward calculations, that for each $k, n \in \mathbb{N}$, with $n > k$,

$$|E_z^{\pi_n}[\sum_{t=0}^{T\wedge n-1} r(X_t,A_t)] - E_z^{\pi}[\sum_{t=0}^{T-1} r(X_t,A_t)]|$$

$$\leq |E_z^{\pi_n}[\sum_{t=0}^{T\wedge k-1} r(X_t,A_t)] - E_z^{\pi}[\sum_{t=0}^{T\wedge k-1} r(X_t,A_t)]| + 2\|r\|\cdot b_k.$$

Upon taking limit superior as $n\to\infty$ in both sides of this inequality and using that $\lim_{n\to\infty}\pi_n=\pi$, property (c) above implies that

$$\limsup_{n\to\infty} |E_z^{\pi_n}[\sum_{t=0}^{T\wedge n-1} r(X_t,A_t)] - E_z^{\pi}[\sum_{t=0}^{T-1} r(X_t,A_t)]| \leq 2\|r\|b_k,$$

and combining this inequality with part (i) it follows that

$$\lim_{n\to\infty} E_z^{\pi_n}[\sum_{t=0}^{T\wedge n-1} r(X_t,A_t)] = E_z^{\pi}[\sum_{t=0}^{T-1} r(X_t,A_t)],$$

which is the desired conclusion. □

Lemma 4.2 *Let $c\in\mathbb{R}$ and $W:\mathbb{N}\times(\mathbb{N}\setminus\{0\})\to\mathbb{R}$ satisfy (i) and (ii) below.*

(i) $|W(m,k)|\leq c\cdot k, \quad m,k\in\mathbb{N},\ k\geq 1$;

(ii) For each positive integer k, $\lim_{m\to\infty}W(m,k)=:\lambda(k)$ *exists.*

Then, if $\{\pi_m\}\subset\mathbb{M}$ is a sequence converging to $\pi\in\mathbb{M}$ and $\{n(m)\}\subset\mathbb{N}$ is a sequence increasing to ∞,

$$\lim_{m\to\infty}\sum_{k=1}^{n(m)} W(m,k)P_z^{\pi_m}[T=k] = \sum_{k=1}^{\infty}\lambda(k)P_z^{\pi}[T=k]. \tag{4.2}$$

Proof: First notice that, by Lemma 4.1(ii),

(a) For each $k\in\mathbb{N}\setminus\{0\}$, $\lim_{m\to\infty}P_z^{\pi_m}[T=k]=P_z^{\pi}[T=k]$.

Next, define the sequence $\{f_m:\mathbb{N}\setminus\{0\}\to\mathbb{R}\}$ as follows: For each $m\in\mathbb{N}$,

$$f_m(k):=W(m,k) \quad \text{if} \quad 1\leq k\leq n(m), \quad f_m(k):=0, \quad \text{otherwise.} \tag{4.3}$$

From the assumptions in the statement of the lemma it follows that

(b) For $k, m \in \mathbb{N},\ k \geq 1$,

$$|f_m(k)| \leq c \cdot k, \quad k, m \in \mathbb{N},\ k \geq 1, \quad \text{and} \quad \lim_{m \to \infty} f_m(k) = \lambda(k).$$

On the other hand, by Lemma 4.1(iii),

$$(c) \quad \lim_{m \to \infty} \sum_{k=1}^{\infty} k P_z^{\pi_m}[T = k] = \lim_{m \to \infty} E_z^{\pi_m}[T] = E_z^{\pi}[T] = \sum_{k=1}^{\infty} k P_z^{\pi}[T = k].$$

Then, (a)–(c) allow to use Proposition 2.18 in [19, p.232] to conclude that

$$\lim_{m \to \infty} \sum_{k=1}^{\infty} f_m(k) P_z^{\pi_m}[T = k] = \sum_{k=1}^{\infty} \lambda(k) P_z^{\pi}[T = k],$$

which, by (4.3), is equivalent to (4.2). □

The following result follows combining (2.2) and (2.3) with equation (6.2) in [5] applied with $n = 1$, or from Theorem 4.2 in [4].

Lemma 4.3 *Let $\pi \in \mathbb{P}$ be an arbitrary policy. Then,*

$$\frac{E_z^{\pi}[\sum_{t=0}^{T-1} r(X_t, A_t)]}{E_z^{\pi}[T]} \leq g.$$

Lemmas 4.1-4.3 will be now used to establish Theorem 4.1.

Proof of Theorem 4.1: Let $n \in \mathbb{N}$ be arbitrary. As already noted, there exists $\pi^n \in \mathbb{M}$ such that

$$V_n(x) = E_x^{\pi^n}[\sum_{t=0}^{n} r(X_t, A_t)], \quad x \in S. \tag{4.4}$$

Now select a sequence $\{n(m)\} \subset \mathbb{N}$ increasing to ∞ such that

$$\lim_{m \to \infty} g_{n(m)}(z) = \lim_{m \to \infty} [V_{n(m)}(z) - V_{n(m)-1}(z)] = U(z) =: \lambda(1); \tag{4.5}$$

see Definition 4.1(i). On the other hand, since $\{\pi^n\} \subset \mathbb{M}$ and $\mathbb{M}$ is a *compact metric* space, it can be assumed—taking a subsequence if necessary—that

$$\lim_{m \to \infty} \pi^{n(m)} =: \pi \in \mathbb{M} \tag{4.6}$$

exists. To complete the proof it will be shown that this policy π satisfies (4.1). First observe that (4.4) and Bellman's optimality principle together yield that

$$V_n(z) = E_z^{\pi^n}[\sum_{t=0}^{T\wedge n-1} r(X_t, A_t) + V_{n-T\wedge n}(X_{T\wedge n})],$$

or equivalently,

$$V_n(z) - E_z^{\pi^n}[V_{n-T\wedge n}(X_{T\wedge n})] = E_z^{\pi^n}[\sum_{t=0}^{T\wedge n-1} r(X_t, A_t)]. \tag{4.7}$$

On the other hand,

$$\begin{aligned}
&V_n(z) - E_z^{\pi^n}[V_{n-T\wedge n}(X_{T\wedge n})] \\
&= V_n(z) - \sum_{k=1}^{n} V_{n-k}(z) P_z^{\pi^n}[T=k] - E_z^{\pi^n}[V_0(X_n) I[T>n]] \\
&= \sum_{k=1}^{n} (V_n(z) - V_{n-k}(z)) P_z^{\pi^n}[T=k] + E_z^{\pi^n}[(V_n(z) - V_0(X_n)) I[T>n]],
\end{aligned} \tag{4.8}$$

and (see (3.3))

$$(V_{n(m)}(z) - V_{n(m)-k}(z) | k = 2, 3, \cdots) \in \Pi_{k=2}^{\infty}[-k\|r\|, k\|r\|] =: \mathbb{E},$$

where, by convention, $V_s(z) := 0$ for $s < -1$; recall that $V_{-1} \equiv 0$, by Definition 3.1. Since $\mathbb{E}$ is a compact metric space, after taking a subsequence of $\{n(m)\}$ it can be assumed that, in addition to (4.5) and (4.6),

$$\lim_{m\to\infty} [V_{n(m)}(z) - V_{n(m)-k}(z)] =: \lambda(k)$$

exists for all $k \geq 2$. Setting $W(m,k) := V_{n(m)}(z) - V_{n(m)-k}(z)$, $m, k \in \mathbb{N}$, $k \geq 1$ and $c := \|r\|$, the last displayed equality, (4.5) and (4.6) together yield, via Lemma 4.2, that as $m \to \infty$,

$$\sum_{k=1}^{n(m)} [V_{n(m)}(z) - V_{n(m)-k}(z)] P_z^{\pi^{n(m)}}[T=k] \to \sum_{k=1}^{\infty} \lambda(k) P_z^{\pi}[T=k]. \tag{4.9}$$

To continue observe that

$$\begin{aligned}
&|E_z^{\pi^n}[(V_n(z) - V_0(X_n)) I[T>n]]| \\
&\qquad \leq E_z^{\pi^n}[((n+1)\|r\| + \|r\|) I[T>n]] \quad \text{(by (3.2))} \\
&\qquad \leq \|r\| E_z^{\pi^n}[2T I[T>n]] \leq 2\|r\| b_n \quad \text{(see Definition 4.1}(ii)\text{)}
\end{aligned}$$

so that, by Lemma 4.1(i),

$$E_z^{\pi_n}[(V_n(z) - V_0(X_n))I[T > n]] \to 0 \quad \text{as} \quad n \to \infty. \tag{4.10}$$

Therefore, replacing n by $n(m)$ in (4.8) and taking limit as $m \to \infty$, (4.9) and (4.10) together imply that

$\lim_{m\to\infty}\{V_{n(m)}(z) - E_z^{\pi^{n(m)}}[V_{n(m)-T\wedge n(m)}(X_{T\wedge n(m)})]\}$

$$= \sum_{k=1}^{\infty} \lambda(k) P_z^{\pi}[T = k]. \tag{4.11}$$

Next, combining (4.6) and Lemma 4.1(iv) it follows that

$$\lim_{m\to\infty} E_z^{\pi^{n(m)}}[\sum_{t=0}^{T\wedge n(m)-1} r(X_t, A_t)] = E_z^{\pi}[\sum_{t=0}^{T-1} r(X_t, A_t)] \tag{4.12}$$

so that, replacing n by $n(m)$ in (4.7) and taking limit as $m \to \infty$ in both sides of the resulting equality, equations (4.11) and (4.12) together imply that

$$\sum_{k=1}^{\infty} \lambda(k) P_z^{\pi}[T = k] = E_z^{\pi}[\sum_{t=0}^{T-1} r(X_t, A_t)]. \tag{4.13}$$

Finally, it will be shown that

$$\lambda(s) \geq U(z) + (s-1)L(z), \quad s = 1, 2, \cdots. \tag{4.14}$$

Assuming this inequality, the conclusion follows in this way: Notice that

$$\begin{aligned}
\sum_{s=1}^{\infty} \lambda(s) P_z^{\pi}[T = s] &\geq \sum_{s=1}^{\infty} \{U(z) + (s-1)L(z)\} P_z^{\pi}[T = s] \\
&= U(z) - L(z) + L(z) \sum_{s=1}^{\infty} s P_z^{\pi}[T = s] \\
&= U(z) - L(z) + L(z) E_z^{\pi}[T],
\end{aligned}$$

and that in combination with (4.13) this implies that

$$U(z) - L(z) + L(z) E_z^{\pi}[T] \leq E_z^{\pi}[\sum_{t=0}^{T-1} r(X_t, A_t)],$$

so that

$$L(z) + \frac{U(z) - L(z)}{E_z^{\pi}[T]} \leq \frac{E_z^{\pi}[\sum_{t=0}^{T-1} r(X_t, A_t)]}{E_z^{\pi}[T]} \leq g,$$

where the second inequality follows from Lemma 4.3. Thus, the policy $\pi \in \mathbb{M}$ given in (4.6) satisfies (4.1). To complete the proof it is sufficient to establish (4.14). With this in mind note that, by Definition 4.i(i), for each integer $k \geq 1$

$$\begin{aligned} L(z) &= \liminf_{n\to\infty}[V_n(z) - V_{n-1}(z)] \\ &\leq \liminf_{m\to\infty}[V_{n(m)-k}(z) - V_{n(m)-k-1}(z)]. \end{aligned} \tag{4.15}$$

Then, for each positive integer s,

$$\begin{aligned} \lambda(s) &= \lim_{m\to\infty}[V_{n(m)}(z) - V_{n(m)-s}(z)] \\ &= \lim_{m\to\infty}[V_{n(m)}(z) - V_{n(m)-1}(z) + \sum_{k=1}^{s-1}(V_{n(m)-k}(z) - V_{n(m)-k-1}(z))] \\ &= U(z) + \lim_{m\to\infty}\sum_{k=1}^{s-1}(V_{n(m)-k}(z) - V_{n(m)-k-1}(z))] \quad \text{(see (4.5))} \\ &\geq U(z) + \sum_{k=1}^{s-1}\liminf_{m\to\infty}(V_{n(m)-k}(z) - V_{n(m)-k-1}(z))] \end{aligned}$$

and combining the last inequality with (4.15) it follows that

$$\lambda(s) \geq U(z) + (s-1)L(z);$$

this establishes (4.14), since the integer $s \geq 1$ was arbitrary and, as already mentioned, completes the proof of Theorem 4.1. □

5 Preliminaries: Second Part

The objective of this section is to establish the second essential component of the proof of Theorem 3.1, which is the following.

Theorem 5.1 *For each $x \in S$,*

$$L(x) \geq g.$$

The proof of this theorem is somewhat technical and has been split into Lemmas 5.1-5.3 below. The idea behind the arguments used to establish part (*ii*) of the following lemma was motivated by the proof of the 'Key Renewal Theorem' as presented in Feller [10, Section 7 of Chapter 13].

Lemma 5.1 *(i) For all $x \in S$, (a) $L(x) \geq L(z)$, and (b) $U(x) \leq U(z)$.*

(ii) Let $\{n(m)\}_{m=0}^{\infty} \subset \mathbb{N} \setminus \{0\}$ *be a given sequence increasing to* ∞ *such that*

$$\lim_{m\to\infty} g_{n(m)}(z) = L(z).$$

Then

$$\lim_{m\to\infty} g_{n(m)-1}(z) = L(z). \tag{5.1}$$

Proof: (i) For each $x \in S$ select a sequence $\{n_x(m)\}_{m=0}^{\infty}$ of positive integers satisfying

$$\lim_{m\to\infty} g_{n_x(m)}(x) = L(x); \tag{5.2}$$

see Definition 4.1. Next, for each $m \in \mathbb{N}$ select $f_m \in \mathbb{F}$ such that

$$V_{n_x(m)-1}(x) = r(x, f_m(x)) + \sum_y p_{x\,y}(f_m(x)) V_{n_x(m)-2}(y), \quad x \in S;$$

the existence of such a stationary policy follows from Assumption 2.1 and the boundedness of the *VI* functions $V_k(\cdot)$ (see (3.2) and recall that $V_{-1} \equiv 0$). Now observe that, by Definition 3.1(i),

$$V_{n_x(m)}(x) \geq r(x, f_m(x)) + \sum_y p_{x\,y}(f_m(x)) V_{n_x(m)-1}(y) \quad x \in S.$$

The last two displayed relations together imply, by Definition 3.1(iii), that for each state x

$$g_{n_x(m)}(x) \geq \sum_y p_{x\,y}(f_m(x)) g_{n_x(m)-1}(y). \tag{5.3}$$

On the other hand, using that $\mathbb{F}$ is compact metric, it is possible to select a subsequence such that, in addition to (5.2),

$$f_m \to f \in \mathbb{F} \ \text{ as } \ m \to \infty. \tag{5.4}$$

Since $|g_k(\cdot)| \leq \|r\|$ (see (3.4)), inequality (5.3), Fatou's lemma and Assumption 2.1 together imply that, for all $x \in S$

$$\begin{aligned}
\liminf_{m\to\infty} \{g_{n_x(m)}(x) + \|r\|\} &\geq \sum_y \liminf_{m\to\infty} (p_{x\,y}(f_m(x))[g_{n_x(m)-1}(y) + \|r\|]) \\
&= \sum_y p_{x\,y}(f(x))(\liminf_{m\to\infty} g_{n_x(m)-1}(y) + \|r\|) \\
&\geq \sum_y p_{x\,y}(f(x))(L(y) + \|r\|)
\end{aligned}$$

where the second inequality is due to the definition of $L(y)$ as the limit inferior of the *whole* sequence $\{g_n(y)\}$. Combining this with (5.2) it follows that

$$\begin{aligned} L(x) &\geq \sum_y p_{x\,y}(f(x)) \liminf_{m\to\infty} g_{n_x(m)-1}(y) \\ &\geq \sum_y p_{x\,y}(f(x)) L(y), \quad x \in S. \end{aligned} \tag{5.5}$$

Since $|L(\cdot)| \leq \|r\| \leq \|r\| l(\cdot)$ (see Remark 4.1 and recall that the Lyapunov function $l(\cdot)$ is ≥ 1), Lemma 2.2(*iiib*) and (5.5) together yield that $L(x) \geq L(z)$, $x \in S$ establishing part (a) whereas part (b) follows along the same lines.

(ii) Let $\{n(m)\}_{m=0}^{\infty}$ be a sequence of positive integers increasing to ∞ such that $\lim_{m\to\infty} g_{n(m)}(z) = L(z)$ and let $L'(z)$ be an *arbitrary* limit point of $\{g_{n(m)-1}(z)\}_{m=0}^{\infty}$. From Definition 4.1 it is clear that $L'(z) \geq L(z)$ and that (5.1) will be established if it can be proved that

$$L'(z) = L(z). \tag{5.6}$$

With this in mind, observe that taking a subsequence—if necessary—it can be assumed that

$$\lim_{m\to\infty} g_{n(m)-1}(z) = L'(z). \tag{5.7}$$

Now, in the proof of part (i) set $n_z(m) = n(m)$ for all $m \in \mathbb{N}$ and note that, since $\mathbb{F}$ is a compact metric space, after picking an additional subsequence it can be supposed that (5.4) also holds. Then, (5.5) with $x = z$ yields

$$L(z) \geq \sum_y p_{z\,y}(f(z)) \liminf_{m\to\infty} g_{n(m)-1}(y) \geq \sum_y p_{z\,y}(f(z)) L(y) \geq L(z),$$

where part (i) was used to obtain the right-most inequality. Therefore, all inequalities in the last displayed relation are equalities, and then

$$\sum_y p_{z\,y}(f(z)) [\liminf_{m\to\infty} g_{n(m)-1}(y) - L(y)] = 0. \tag{5.8}$$

Finally, from Definition 4.1(i), $\liminf_{m\to\infty} g_{n(m)-1}(y) \geq L(y)$ for all $y \in S$, so that (5.8) implies that

$$\liminf_{m\to\infty} g_{n(m)-1}(y) = L(y) \quad \text{if} \quad p_{z\,y}(f(z)) > 0;$$

then Assumption 2.3 and (5.7) together yield that $L'(z) = L(z)$ and, as already mentioned, this completes the proof. □

Lemma 5.2 *Let $\{R_n\}$ be the sequence of relative value functions introduced in Definition 3.1. Then,*

$$|R_n(x)| \leq 3\|r\| l(x), \quad x \in S,\ n \in \mathbb{N}.$$

Proof: This result is essentially contained in [6]; by completeness, a short proof is given. Let $n \in \mathbb{N}$ and $x \in S$ be fixed and select policy π^n as in (3.1) . Using Bellman's optimality principle, it follows that

$$\begin{aligned} R_n(x) &= V_n(x) - V_n(z) \\ &= E_x^{\pi^n}[\sum_{t=0}^{T\wedge n-1} r(X_t, A_t) + V_{n-T\wedge n}(X_{T\wedge n}) - V_n(z)]. \end{aligned} \tag{5.9}$$

Since $T \leq n$ implies that $T \wedge n = T$ and $X_{T\wedge n} = X_T = z$ (by (2.4)), then

$$\begin{aligned} |V_{n-T\wedge n}(X_{T\wedge n}) - V_n(z)| I[T \leq n] &= |V_{n-T}(z) - V_n(z)| I[T \leq n] \\ &\leq \|r\| T I[T \leq n], \end{aligned} \tag{5.10}$$

where (3.3) was used to obtain the inequality. On the other hand $T \wedge n = n$ on the event $[T > n]$, so that using (3.3)

$$\begin{aligned} |V_{n-T\wedge n}(X_{T\wedge n}) - V_n(z)| I[T > n] &= |V_0(X_n) - V_n(z)| I[T > n] \\ &\leq (\|r\| + (n+1)\|r\|) I[T > n] \\ &\leq 2\|r\| T I[T > n]. \end{aligned}$$

Combining this inequality with (5.10) it follows that

$$\begin{aligned} &E_x^{\pi^n}[|V_{n-T\wedge n}(X_{T\wedge n}) - V_n(z)|] \\ &\quad = E_x^{\pi^n}[|V_{n-T}(z) - V_n(z)| I[T \leq n]]] \\ &\qquad + E_x^{\pi^n}[|V_0(X_n) - V_n(z)| I[T > n]] \\ &\quad \leq \|r\| E_x^{\pi^n}[T I[T \leq n]] + 2\|r\| E_x^{\pi^n}[T I[T > n]] \\ &\quad \leq 2\|r\| E_x^{\pi^n}[T]. \end{aligned} \tag{5.11}$$

Finally, observe that

$$|E_x^{\pi^n}[\sum_{t=0}^{T\wedge n-1} r(X_t, A_t)]| \leq \|r\| E_x^{\pi^n}[T \wedge n] \leq \|r\| E_x^{\pi^n}[T]$$

which, combined with (5.9) and (5.11), yields $|R_n(x)| \leq 3\|r\| E_x^{\pi^n}[T] \leq 3\|r\| l(x)$, where Lemma 2.2($i$) was used to obtain the second inequality. □

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

Lemma 5.3 *There exists a sequence* $\{\tilde{h}_k : S \to \mathbb{R}\}$ *satisfying* $|\tilde{h}_k(\cdot)| \leq 3\|r\|l(\cdot)$ *for all* $k \in \mathbb{N}$ *and, furthermore,*

$$L(z)+\tilde{h}_k(x) \geq r(x,a)+\sum_y p_{x\,y}(a)\tilde{h}_{k+1}(y), \quad (x,a) \in \mathbb{K}, \quad k \in \mathbb{N}. \tag{5.12}$$

Proof: Pick a sequence $\{n(m)\}$ such that $g_{n(m)}(z) \to L(z)$ as $m \to \infty$. By Lemma 5.1(ii) this implies that

$$\lim_{m\to\infty} g_{n(m)-k}(z) = L(z), \quad k \in \mathbb{N}. \tag{5.13}$$

Next set $\mathbb{D} := \Pi_{x\in S}[-3\|r\|l(x), 3\|r\|l(x)]$, and for $s < 0$, $g_s(z) := 0$ and $R_s(\cdot) := 0$. With this notation

$$W_m := (g_{n(m)-k}; R_{n(m)-k}|k \in \mathbb{N}) \in ([-\|r\|, \|r\|] \times \mathbb{D})^\infty =: \mathbb{H}. \tag{5.14}$$

Since $\mathbb{H}$ is compact metric in the product topology, taking a subsequence if necessary it can be assumed that

$$\lim_{m\to\infty} W_{n(m)} =: W \in \mathbb{H} \tag{5.15}$$

exists, and from (5.13) it follows that W is of the form

$$W = (L(z); \tilde{h}_k|k \in \mathbb{N}) \tag{5.16}$$

for certain functions $\tilde{h}_k : S \to \mathbb{R}$ belonging to $\mathbb{D}$, i.e.,

$$|\tilde{h}_k(x)| \leq 3\|r\|l(x), \quad x \in S, \ k \in \mathbb{N}.$$

To complete the proof it will be shown that the sequence $\{\tilde{h}_k\}$ satisfies (5.12). First note that (5.14)-(5.16) yield that

$$\lim_{m\to\infty} R_{n(m)-k}(x) = \tilde{h}_k(x), \quad x \in S, \ k \in \mathbb{N}. \tag{5.17}$$

On the other hand, Definition 3.1(i) implies that for all $(x,a) \in \mathbb{K}$ and $m, k \in \mathbb{N}$ with $k < n(m)$

$$V_{n(m)-k}(x) \geq r(x,a) + \sum_y p_{x\,y}(a)V_{n(m)-k-1}(y)$$

which is equivalent to

$$g_{n(m)-k}(z) + R_{n(m)-k}(x) \geq r(x,a) + \sum_y p_{x\,y}(a)R_{n(m)-k-1}(y); \tag{5.18}$$

see parts (ii) and (iii) of Definition 3.1. Since $\sum_y p_{x\,y}(a)l(y) < \infty$ (by Assumption 2.2), Lemma 5.2, (5.17) and the dominated convergence theorem together imply that

$$\lim_{m\to\infty} \sum_y p_{x\,y}(a) R_{n(m)-k-1}(y) = \sum_y p_{x\,y}(a)\tilde{h}_{k+1}(y).$$

Taking limit as $m \to \infty$ in both sides of (5.18) and using the above convergence together with (5.13) and (5.17), it follows that for all $(x,a) \in \mathbb{K}$ and $k \in \mathbb{N}$

$$L(z) + \tilde{h}_k(x) \geq r(x,a) + \sum_y p_{x\,y}(a)\tilde{h}_{k+1}(y),$$

and the proof is complete. □

Lemmas 5.1–5.3 will be now used to establish Theorem 5.1.

Proof of Theorem 5.1: Let $\{\tilde{h}_k\}$ be the sequence in Lemma 5.3. A simple induction argument using (5.12) yields that for all $x \in S$, $\pi \in \mathbb{P}$ and $n \in \mathbb{N}$

$$L(z) + \frac{\tilde{h}_0(x)}{n+1} \geq \frac{E_x^\pi[\sum_{t=0}^n r(X_t, A_t)]}{n+1} + \frac{E_x^\pi[\tilde{h}_{n+1}(X_{n+1})]}{n+1}.$$

Since $|\tilde{h}_k(\cdot)| \leq 3\|r\| l(\cdot)$, (2.10) implies that $\frac{1}{n+1} E_x^\pi[\tilde{h}_{n+1}(X_{n+1})] \to 0$ as $n \to \infty$. Thus, taking limit superior in both sides of the last displayed inequality it follows that

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

and then $L(z) \geq J(x) = g$, since $\pi \in \mathbb{P}$ was arbitrary; see (2.2), (2.3) and Lemma 2.1(i). Then Lemma 5.1(i) yields that $L(x) \geq L(z) \geq g$ for all state x and the proof is complete. □

6 Proof of Theorem 3.1

After the preliminaries in the previous sections the main result of this note can be established as follows.

Proof of Theorem 3.1: (i) Let $\pi \in \mathbb{M}$ be as in Theorem 4.1 and note that

$$L(z) \leq L(z) + \frac{U(z) - L(z)}{E_z^\pi[T]} \leq g \leq L(z),$$

where $U(z) \geq L(z)$ was used to obtain the leftmost inequality, the middle one is nothing but (4.1), and the third inequality follows from Theorem 5.1. Then $L(z) = g$ and

$$\frac{U(z) - L(z)}{E_z^\pi[T]} = 0.$$

Since $E_z^\pi[T] < \infty$, by Lemma 2.2(i), this yields that

$$\limsup_{n\to\infty} g_n(z) = U(z) = L(z) = \liminf_{n\to\infty} g_n(z),$$

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

(ii) Combining part (i) above with Lemma 5.1(i) it follows that

$$U(x) \leq U(z) = L(z) = g \leq L(x), \quad x \in S,$$

which, using that $U(\cdot) \geq L(\cdot)$, yields $U(x) = L(x) = g$, $x \in S$, i.e., $\lim_{n\to\infty} g_n(x) = g$ for each state x; see Definition 4.1(i).

(iii) Notice that, by Lemma 5.2,

$$R_n \in \mathbb{D} = \Pi_{x\in S}[-3\|r\|l(x), 3\|r\|l(x)], \quad n \in \mathbb{N}, \tag{6.1}$$

and that $\mathbb{D}$ is a compact metric space. Thus, to establish that

$$\lim_{n\to\infty} R_n = h$$

it is sufficient to verify that any limit point of $\{R_n\}$ coincides with h. With this in mind, let $Q \in \mathbb{D}$ be an arbitrary limit point of $\{R_n\}$, select a sequence $\{n(m)\}$ of positive integers such that

$$\lim_{m\to\infty} R_{n(m)}(x) = Q(x) \in [-3\|r\|l(x), 3\|r\|l(x)], \quad x \in S, \tag{6.2}$$

and observe that, by Definition 3.1 and part (ii) above, $R_k(x) - R_{k-1}(x) = g_k(x) - g_k(z) \to 0$ as $k \to \infty$. Therefore

$$\lim_{m\to\infty} R_{n(m)-1}(x) = Q(x), \quad x \in S, \tag{6.3}$$

also holds. Now let $x \in S$ be arbitrary but fixed and note that straightforward calculations using Definition 3.1 yield that, for all $x \in S$ and $m \in \mathbb{N}$,

$$g_{n(m)}(z) + R_{n(m)}(x) = \sup_{a\in A(x)} [r(x,a) + \sum_y p_{x\,y}(a) R_{n(m)-1}(y)]. \tag{6.4}$$

Next, let $G \subset S$ be a *finite* set and $\varepsilon > 0$. Using Assumptions 2.1 and 2.2, it follows that the mappings

$$a \mapsto \sum_y p_{x\,y}(a) l(y) \quad \text{and} \quad a \mapsto \sum_{y \in G} p_{x\,y}(a) l(y)$$

are continuous in $a \in A(x)$. Since $\sum_{y \in G} p_{x\,y}(a) l(y) \nearrow \sum_y p_{x\,y}(a) l(y)$ as $G \nearrow S$ and $A(x)$ is a compact subspace of A, Dini's Theorem [19, p. 162] implies that

$$\sup_{a \in A(x)} \sum_{y \notin G^*} p_{x\,y}(a) l(y) < \varepsilon \tag{6.5}$$

for some finite set $G^* \subset S$. Then, using Lemma 3.3 in [15] it follows that

$$\begin{aligned}
&|\sup_{a \in A(x)} [r(x,a) + \sum_y p_{x\,y}(a) R_{n(m)-1}(y)] - \sup_{a \in A(x)} [r(x,a) + \sum_y p_{x\,y}(a) Q(y)]| \\
&\qquad \le \sup_{a \in A(x)} \sum_y p_{x\,y}(a) |R_{n(m)-1}(y) - Q(y)| \\
&\qquad \le \sum_{y \in G^*} |R_{n(m)-1}(y) - Q(y)| \\
&\qquad\quad + \sup_{a \in A(x)} \sum_{y \notin G^*} p_{x\,y}(a) [|R_{n(m)-1}(y)| + |Q(y)|] \\
&\qquad \le \sum_{y \in G^*} |R_{n(m)-1}(y) - Q(y)| + 6\|r\|\varepsilon
\end{aligned}$$

where (6.1), the inclusion in (6.2) and (6.5) were used to obtain the last inequality. Since G^* is a *finite* subset of S and $\varepsilon > 0$ is arbitrary, it follows, via (6.3), that

$$\begin{aligned}
&|\sup_{a \in A(x)} [r(x,a) + \sum_y p_{x\,y}(a) R_{n(m)-1}(y)] - \sup_{a \in A(x)} [r(x,a) + \sum_y p_{x\,y}(a) Q(y)]| \\
&\qquad\qquad\qquad\qquad \to 0 \quad \text{as} \quad m \to \infty.
\end{aligned} \tag{6.6}$$

Finally, take limit as m goes to ∞ in both sides of (6.4). In this case, part (i), (6.2) and (6.6) together imply that

$$g + Q(x) = \sup_{a \in A(x)} [r(x,a) + \sum_y p_{x\,y}(a) Q(y)],$$

i.e., $Q(\cdot)$ is a solution of the *AROE*. To conclude note that $R_k(z) = 0$ for all k, by Definition 3.1, so that (6.2) yields that $Q(z) = 0$. Then, using the inclusion in (6.2) together with the properties of the function $h(\cdot)$ in

Lemma 2.1, the equality $Q(\cdot) = h(\cdot)$ follows from Lemma 2.2(iv) and, as already mentioned, this completes the proof of part (iii).

(iv) For each $n \in \mathbb{N}$ let the policy $f_n \in \mathbb{F}$ be such that, for all $(x, a) \in \mathbb{K}$,

$$r(x, f_n(x)) + \sum_y p_{x\,y}(f_n(x))R_n(y) \geq r(x, a) + \sum_y p_{x\,y}(a)R_n(y); \qquad (6.7)$$

the existence of such a stationary policy f_n follows from Assumption 2.1 and the boundedness of $R_n(\cdot) = V_n(\cdot) - V_n(z)$. Now let $f \in \mathbb{F}$ be an accumulation point of $\{f_n\}$ and pick a subsequence $\{f_{n(m)}\}$ such that

$$f_{n(m)} \to f \quad \text{as} \quad m \to \infty. \qquad (6.8)$$

To complete the proof it will be shown that this policy f is average optimal. As already noted, given $x \in S$, the mapping $a \mapsto \sum_y p_{x\,y}(a)l(y)$ is *finite and continuous* in $a \in A(x)$, so that using Lemma 5.2 and part (iii) above it follows that

$$\lim_{m\to\infty} \sum_y p_{x\,y}(f_{n(m)}(x))R_{n(m)}(y) = \sum_y p_{x\,y}(f(x))h(y),$$

by Proposition 2.18 in [19 p. 232], and

$$\lim_{m\to\infty} \sum_y p_{x\,y}(a)R_{n(m)}(y) = \sum_y p_{x\,y}(a)h(y),$$

by the dominated convergence theorem. Replacing n by $n(m)$ in both sides of (6.7) and taking limit as $m \to \infty$, the last two displayed equalities, (6.8) and Assumption 2.1 together imply

$$r(x, f(x)) + \sum_y p_{x\,y}(f(x))h(y) \geq r(x, a) + \sum_y p_{x\,y}(a)h(y), \quad (x, a) \in \mathbb{K},$$

and then f is optimal, by Lemma 2.1(iv). □

7 Conclusion

The value iteration procedure has been studied in the context of MDP's with denumerable state space, bounded rewards and satisfying the continuity and stability conditions in Assumptions 2.1–2.3; as already mentioned, Assumption 2.3 does not represent a real restriction, since it can be achieved after performing the Schweitzer transformation described in Remark 2.1. Within this framework, the convergence of the relative value functions and differential rewards to the solution of the *AROE* was established in Theorem 3.1, and the relation of this result with the theorems obtained in [6] and [17] was briefly discussed in Remark 3.1. On the other hand, there

are, at least, two interesting problems to be considered. First, note that the Lyapunov stability condition in Assumption 2.2 can be appropriately modified to include *unbounded* rewards [16,17,23], and that the arguments used to establish Theorem 3.1 rely heavily on the assumption that the reward function is *bounded*. Thus, it is interesting to consider the following problem:

P1: Is it possible to extend the results in Theorem 3.1 to include unbounded reward functions?

On the other hand, an important application of the ideas behind the *VI* method is to the construction of average optimal *adaptive* policies. The point here is that, frequently, the transition-reward structure of the model is *not* completely known to the controller, but depends on *unknown* parameters, so that the control task must be combined with an estimation scheme. The combination of the *VI* method with an estimation procedure was initiated by Federgruen and Schweitzer [9] and Hernández-Lerma and Marcus [11]; in the latter paper the Non-stationary Value Iteration (*NVI*) adaptive policy was introduced for discounted models, and the idea was extended to the average case in a series of papers, including [1]; an extensive discussion about this theme can be found in [12]. However, the optimality results for the *NVI* adaptive policy in the average case has been established under *SDC*-like stability restrictions. The reason behind this, is that conditions for the average optimality of the *NVI* adaptive policy are strongly linked to assumptions ensuring that the convergence results in Theorem 3.1 occur, and when the average version of the *NVI* adaptive policy was introduced, those results were available under *SDC* but not under *LFC*. Therefore, the following seems to be an interesting problem:

P2: When the transition-reward structure of the model depends on unknown parameters, combine the results in Theorem 3.1 with an estimation scheme to establish the average optimality of the *NVI* adaptive policy.

Research on these problems is presently in progress.

Acknowledgement

It is a pleasure to thank Professors R. Montes-de-Oca and O. Hernández-Lerma for showing me their paper [18] before publication. Also, the author is grateful to the unknown reviewers for their careful reading of the original manuscript and their helpful suggestions to improve the content and presentation of the paper.

References

[1] R.S. Acosta-Abreu and O. Hernández-Lerma. Iterative adaptive control of denumerable state average-cost Markov systems, *Control and Cybernetics* **14** (1985), 313-322.

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

[3] R. Cavazos-Cadena. Necessary conditions for the optimality equation in average-reward Markov decision processes, *Journal of Applied Mathematics and Optimization* **19** (1989), 97-112.

[4] R. Cavazos-Cadena and O. Hernández-Lerma. Equivalence of Lyapunov stability criteria in a class of Markov decision processes, *Journal of Applied Mathematics and Optimization* **26** (1992), 113-137.

[5] R. Cavazos-Cadena. Existence of optimal stationary policies in average reward Markov decision processes with a recurrent state, *Journal of Applied Mathematics and Optimization* **26** (1992), 171-194.

[6] R. Cavazos-Cadena. Cesàro convergence of the undiscounted value iteration method in Markov decision processes under the Lyapunov stability condition, *Boletín de la Sociedad Matemática Mexicana*, 1994, to appear.

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

[8] A. Federgruen and H.C. Tijms. The optimality equation in average cost denumerable state semi-Markov decision problems, recurrency conditions and algorithms, *Journal of Applied Probability* **15** (1978), 356-373.

[9] A. Federgruen and P.J. Schweitzer. Nonstationary Markov decision problems with converging parameters, *Journal of Optimization Theory and Applications* **34** (1981), 207-241.

[10] W. Feller. *An Introduction to Probability Theory and Applications, Vol. I.* New York: John Wiley, 1968.

[11] O. Hernández-Lerma and S.I. Marcus. Adaptive control of discounted Markov decision Chains, *Journal of Optimization Theory and Applications* **46** (1985), 227-235.

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

[13] O. Hernández-Lerma and J.B. Lasserre. Error bounds for rolling horizon policies in discrete-time Markov control processes, *IEEE Transactions on Automatic Control* **35** (1990), 1118-1124.

[14] O. Hernández-Lerma and J.B. Lasserre. Value iteration and rolling plans for Markov control processes with unbounded rewards, *Journal of Mathematical Analysis and Applications*, **177** (1993), 38-55.

[15] K. Hinderer. *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter*, Lecture Notes Oper. Res., **33**. New York: Springer-Verlag, 1970.

[16] A. Hordijk. *Dynamic Programming and Markov Potential Theory*, Mathematical Centre Tract No. 51 (1974), Mathematisch Centrum, Amsterdam.

[**17**] A. Hordijk, P.J. Schweitzer, and H.C. Tijms. The asymptotic behavior of the minimal total expected cost for the denumerable state Markov decision model, *Journal of Applied Probability* **12** (1975), 298-305.

[18] R. Montes-de-Oca and O. Hernández-Lerma. Value iteration in average cost Markov control processes on Borel spaces, Reporte interno No. 143, CINVESTAV-IPN, México D.F., 1993, submitted.

[19] H.L. Royden. *Real Analysis*, 2nd. Edition. New York: Macmillan, 1968.

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

[21] P.J. Schweitzer. Iterative solution of the functional equations for undiscounted Markov renewal programming, *Journal of Mathematical Analysis and Applications* **34** (1971), 495-501.

[22] L.I. Sennott. Value iteration in countable state average cost markov decision processes with unbounded costs, *Annals of Operations Research* **28** (1991), 261-272.

[23] L.C. Thomas. Connectedness conditions for denumerable state Markov decision processes, in *Recent Developments in Markov Decision Processes* (R. Hartley, L.C. Thomas, and D.J. White, eds.), pp. 181-204. New York: Academic Press, 1965.

[24] H.C. Tijms. On dynamic programming with arbitrary state space, compact action space and the average reward criterion, Report B/W 55/75, Mathematisch Centrum, Amsterdam, 1975.

Departamento de Estadística y Cálculo, Universidad Autónoma Agraria Antonio Narro, Buenavista, Saltillo COAH 25315, México

Communicated by Anders Lindquist