Boletín de la Sociedad Matemática Mexicana Vol. 38, 1993

CESÀRO CONVERGENCE OF THE UNDISCOUNTED VALUE ITERATION METHOD IN MARKOV DECISION PROCESSES UNDER THE LYAPUNOV STABILITY CONDITION*

BY ROLANDO CAVAZOS-CADENA

1. Introduction

This work concerns Markov Decision Processes (MDP's) with denumerable state space and discrete time parameter; the reward function is *continuous and bounded* and the performance of a control strategy is measured by the (long-run expected) *average* reward criterion. In this context the following problems are addressed:

Construct

(i) A solution of the *average reward optimality equation* (AROE),
and
(ii) A sequence of policies whose limit points are optimal.

These problems are approached via the Value Iteration (VI) procedure which has been successfully used, for instance, in the following cases: (a) For MDP's with finite state space [12], (b) when the transition law of the system satisfies strong ergodicity conditions like simultaneous Doeblin or Scrambling [3], or (c) when the initial 'error function' in the VI procedure is bounded [8]; see also [5] and inner references. Besides standard continuity-compactness assumptions, conditions (1) and (2) below are supposed to hold true:

(1) Under the action of any stationary policy, the state space is a communicating class,
and
(2) the Lyapunov Function Condition (LFC) for bounded rewards holds true [7, 13];

the latter assumption 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.

Within this framework, an answer to the problems posed above is contained in Theorem (2.1), which is the main result of this note. At this point it should be mentioned that the solution is based on convergences that are weaker than those obtained under conditions stronger than LFC. For instance, under

*This research was generuously auspiced by PSFO under Grant No. 125-350/90-04-93. Also, this work was supported in part by MAXTOR Foundation for Applied Probability and Statistics (MAXFAPS) under Grant No. 01-01-56/04-93, and by the Third World Academy of Sciences (TWAS).

(simultaneous) scrambling the relative value functions produced by the VI method converge uniformly at a geometric rate to a solution of the AROE [3] whereas, under LFC, Theorem (2.1) yields pointwise convergence in a Cesàro sense; however, just this type of result is sufficient to achieve the desired goals.

The remainder of the paper is organized as follows: Section 2 contains a brief description of the control model, the VI procedure as well as the statement of the main result in the form of Theorem (2.1), which is proved in Section 4 after the preliminaries given in Section 3. Finally, this work concludes in Section 5 with some brief comments.

2. Decision model and main result

Let (S, A, r, p) be the usual MDP where the metric space A and the denumerable set S are the action and state spaces, respectively; for each $x \in S$, $A(x) \subset A$ stands for the nonempty set of admissible actions at state x. On the other hand, r is the reward function and p is the transition law. The interpretation of this model is as follows: At each time $t \in \mathbb{N} := \{0, 1, 2, \cdots\}$ 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, (i) a reward $r(x, a)$ is obtained and (ii) regardless of the previous states and actions, the state of the system at time $t + 1$ will be $X_{t+1} = y \in S$ with probability $p(y|x, a)$; this is the Markov property of the 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 mappings $a \mapsto p(y|x, a)$, and $a \mapsto r(x, a)$ are continuous on $A(x)$.
(iii) r is a bounded function, i.e.,

$$\|r\| := \sup\{\|r(x, a)\| \mid x \in S,\ a \in A(x)\} < \infty.$$

Control Policies. For $k \in \mathbb{N}$, the information vector I_k is defined as follows:

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

whereas, 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 $k \leq t$, and $a_k \in A(x_k)$ if $k < t$. A *policy* $\pi = \{\pi_t\}$ is a (measurable, possibly randomized) 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 $\pi_t(A(x_t)|h_t) = 1$ always holds. Set $\mathbb{F} := \Pi_{x \in S} A(x)$, that is, $\mathbb{F}$ consists of all choice functions $f: S \to A$ such that $f(x) \in A(x)$, $x \in S$; notice that $\mathbb{F}$ is compact in the product topology. A policy π is stationary if there exists $f \in \mathbb{F}$ such that $1 = \pi(\{f(x_t)\}|h_t)$ is always valid and, as usual, $\mathbb{F}$ is identified with the class of stationary policies. Given the initial state $X_0 = x_0$ and the policy π

being used, the distribution of the state-action process $\{(X_t, A_t)\}$ is uniquely determined; it is denoted by $P_\pi[\cdot|X_0 = x]$, whereas $E_\pi[\cdot|X_0 = x]$ stands for the corresponding expectation operator (see, for instance, [5, Ch.1]). On the other hand, it can be shown that under the action of any stationary policy the state process $\{X_t\}$ is a Markov chain with stationary transition mechanism [5, 9].

ASSUMPTION (2.2). *Under the action of any stationary policy, the state space is a communicating class. More explicitly, for all $f \in \mathbb{F}$ and $x, y \in S$, there exists $n \equiv n(x, y, f)$ such that $P_f[X_n = y|X_0 = x] > 0$.*

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

$$J(x, \pi) := \limsup_{k\to\infty} \frac{1}{k+1} E_\pi\left[\sum_{t=0}^{k} r(X_t, A_t)|X_0 = x\right]$$

whereas

$$J(x) := \sup_\pi J(x, \pi)$$

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

Optimality Equation. Throughout the remainder $z \in S$ is a *fixed* state and the hitting time T is defined by

$$T := \min\{n > 0 \mid X_n = z\}. \tag{2.2}$$

On the other hand, for an event W, $I[W]$ denotes the corresponding indicator function. Under the following assumption the average reward optimality equation (AROE) given by (2.3) below has a solution yielding an optimal stationary policy.

ASSUMPTION (2.3). *(LFC for bounded rewards [7, 13].) There exists $l: S \to [0, \infty)$ which satisfies conditions (i)-(iii) below; such a function is referred to as a Lyapunov function for bounded rewards.*

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

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

$$f \mapsto \sum_{y\neq z} p(y|x, f(x))l(y) = E_f[l(X_1)I[T > 1]|X_0 = x]$$

is continuous on $\mathbb{F}$.

(iii) For each $f \in \mathbb{F}$ *and* $x \in S$, $E_f[l(X_n)I[T > n]|X_0 = x] \to 0$ *as* $n \to \infty$.

The main consequence of Assumptions (2.1) and (2.3) is the following [7].

LEMMA (2.1). *Under Assumptions (2.1) and (2.3), there exists* $h: S \to \mathbb{R}$ *and* $g \in R$ *such that (i)-(iv) below occur.*

(i) $g = J(x)$, $x \in S$;

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

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

$$g + h(x) = \sup_{a \in A(x)} \left[r(x,a) + \sum_y p(y|x,a)h(y) \right], \quad x \in S. \tag{2.3}$$

(iv) An optimal stationary policy exists. Furthermore, for each $x \in S$, *the right hand side of (2.3)—considered as a function of* $a \in A(x)$—*has a maximizer* $f^*(x)$, *and the corresponding policy* $f^* \in \mathbb{F}$ *is optimal.*

REMARK (2.1). The notation in Lemma (2.1) will be used consistently. Notice that g is *the* optimal average reward at every state, and then it is uniquely determined; the uniqueness of h will be established in Lemma (3.2) below.

REMARK (2.2). There are two additional consequences of Assumption (2.3) which will be important in the next sections:

(i) For any policy π, $E_\pi[T|X_0 = x] \leq l(x)$, $x \in S$, where T is as in (2.2); see, for instance [2, 7].

(ii) For each policy $f \in \mathbb{F}$, the corresponding Markov chain has a unique invariant distribution, say $\{q_f(x) \mid x \in S\}$; in this case, for any bounded reward function r' and $x \in S$,

$$\lim_{n\to\infty} \frac{1}{n+1} E_f \left[\sum_{t=0}^{n} r'(X_t, A_t)|X_t = x \right] \to \sum_y q_f(y) r'(y, f(y)) \quad \text{as} \quad n \to \infty;$$

in particular [9], $J(x,f) = \sum_y q_f(y) r(y, f(y))$, $x \in S$.

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

$$V_n(x) = \sup_{a \in A(x)} \left[r(x,a) + \sum_y p(y|x,a) V_{n-1}(y) \right], \quad x \in S. \tag{2.4}$$

It is known that for all $n \in \mathbb{N}$ there exists a policy π^n such that

$$V_n(x) = E_{\pi^n}\left[\sum_{t=0}^{n} r(X_t, A_t)|X_0 = x\right] = \sup_{\pi} E_{\pi}\left[\sum_{t=0}^{n} r(X_t, A_t)|X_0 = x\right], \quad x \in S; \tag{2.5}$$

see [1, 5, 9, $\cdots$]. On the other hand it is clear that

$$|V_n(\cdot)| \leq (n+1)\|r\|. \tag{2.6}$$

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

$$R_n(x) := V_n(x) - V_n(z), \quad x \in S,\ n = -1, 0, 1, 2, \cdots. \tag{2.7}$$

MAIN RESULT. Define $\{g_n\} \subset \mathbb{R}$ and $\{Q_n : S \to \mathbb{R}\}$ as follows: For $n \in \mathbb{N}$,

$$g_n := \frac{1}{n+1} V_n(z); \tag{2.8}$$

$$Q_n(x) := \frac{1}{n+1} \sum_{k=0}^{n} R_k(x), \quad x \in S. \tag{2.9}$$

With this notation, the following result solves the problems in Section 1.

THEOREM (2.1). *Under Assumptions (2.1)–(2.3) (i)-(iii) below occur.*

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

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

$$\lim_{n\to\infty} Q_n(x) = h(x);$$

see Lemma 3.2 below.

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

$$a \mapsto r(x, a) + \sum_{y} p(y|x, a) Q_n(y), \quad a \in A(x)$$

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

A proof of this theorem will be given in Section 4.

3. Preliminaries

This section contains some preliminaries that will be used in the proof of Theorem (2.1). Throughout this section, Assumptions (2.1) and (2.3) are supposed to hold true. To begin with, for each $n \in \mathbb{N}$ set

$$\delta(n) := \sup_{\pi} E_\pi[l(X_n)I[T > n]|X_0 = z], \tag{3.1}$$

and

$$\Delta(n) := \sup_{\pi} \frac{1}{n+1} E_\pi[l(X_n)|X_0 = z]. \tag{3.2}$$

LEMMA (3.1) *As $n \to \infty$, the following convergences hold:*

(i) $\delta(n) \to 0$, and (ii) $\Delta(n) \to 0$.

Proof (i) This part was obtained in [7]; see *the proof* of equation 5.7.2 in [7, pp.43–44].

(ii) Let $n \in \mathbb{N}$ and the policy π be arbitrary but *fixed,* and define $U := [X_k = z$ for some $k \leq n]$. It is clear that $1 = P_\pi[X_0 = z|X_0 = z] = P_\pi[U|X_0 = z]$ and then, decomposing U into disjoint sets according to the last visit to z up to time n it follows that

$$\sum_{k=0}^{n} I[X_k = z, X_t \neq z,\ k < t \leq n] = I[U] = 1, \quad P_\pi[\cdot|X_0 = z]\text{–a.s.},$$

which implies

$$E_\pi[l(X_n)|X_0 = z] = \sum_{k=0}^{n} E_\pi[l(X_n)I[X_k = z, X_t \neq z, k < t \leq n]|X_0 = z]. \tag{3.3}$$

Now observe that $I[X_k = z]$ is I_k-measurable (by (2.1)) and then an application of the Markov property yields

$$\begin{aligned} E_\pi[l(X_n)I[X_k = z, X_t \neq z, k < t \leq n]|I_k] \\ = I[X_k = z]E_{\pi'}[l(X_{n-k})I[X_t \neq z, 0 < t \leq n-k]|X_0 = z] \end{aligned}$$

where the shifted policy π is determined by ([5, p.5])

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

Hence,

$$\begin{aligned} E_\pi[l(X_n)I[X_k = z, X_t \neq z, k < t \leq n]|I_k] \\ = I[X_k = z]E_{\pi'}[l(X_{n-k})I[T > n-k]|X_0 = z] \\ \leq \delta(n-k); \end{aligned}$$

see (2.2) and (3.1). Taking expectation with respect to $P_\pi[\cdot|X_0 = z]$, the last inequality yields

$$E_\pi[l(X_n)I[X_k = z, X_t \neq z, k < t \leq n]|X_0 = z] \leq \delta(n-k)$$

which, via (3.3), implies that $E_\pi[l(X_n)|X_0 = z] \leq \sum_{k=0}^{n} \delta(n-k)$. Then, since $\pi \in \mathbb{P}$ was arbitrary, it follows that

$$\Delta(n) \leq \frac{1}{n+1} \sum_{k=0}^{n} \delta(n-k)$$

(see (3.2)) and the conclusion is reached using part (i). ∎

LEMMA (3.2). *Let g be the optimal average reward and suppose that the functions $h_j: S \to \mathbb{R}$, $j = 1, 2$ satisfy conditions (i)–(ii):*

(i) $h_j(z) = 0$;

(ii) $|h_j(x)| \leq c \cdot l(x)$, $x \in S$, *where c is a finite constant;*

(iii) $g + h_j(x) = \sup_{a \in A(x)} \big[r(x,a) + \sum_y p(y|x,a)h_j(y)\big]$, $x \in S$.

Then

$$h_1 = h_2.$$

Proof. Let $x \in S$ be arbitrary. Using Assumptions (2.1) and (2.3) together with Proposition 18 in [10, p. 232], it follows that the mapping $a \mapsto r(x,a) + \sum_y p(y|x,a)h_1(y)$ is continuous on the compact set $A(x)$, so that there exists $f_1(x) \in A(x)$ such that

$$r\big(x, f_1(x)\big) + \sum_y p\big(y|x, f_1(x)\big)h_1(y) = \sup_{a \in A(x)} \left[r(x,a) + \sum_y p(y|x,a)h_1(y)\right], \quad x \in S.$$

Hence:

$$g + h_1(x) = r\big(x, f_1(x)\big) + \sum_y p\big(y|x, f_1(x)\big)h_1(y),$$

and

$$g + h_2(x) \geq r\big(x, f_1(x)\big) + \sum_y p\big(y|x, f_1(x)\big)h_2(y);$$

see assumption (iii) in the statement of the theorem. The last two relations together yield

$$\begin{aligned} h_1(x) - h_2(x) &\leq \sum_y p\big(y|x, f_1(x)\big)\big(h_1(y) - h_2(y)\big) \\ &= \sum_{y \neq z} p\big(y|x, f_1(x)\big)\big(h_1(y) - h_2(y)\big) \\ &= E_{f_1}\big[\big(h_1(X_1) - h_2(X_1)\big)I[T > 1]|X_0 = x\big], \end{aligned}$$

where $h_1(z) = h_2(z) = 0$ was used in the first equality. Then, an induction argument yields: For all $n \in \mathbb{N}$ and $x \in S$,

$$\begin{aligned} h_1(x) - h_2(x) &\leq E_{f_1}[(h_1(X_n) - h_2(X_n))I[T > n]|X_0 = x] \\ &\leq 2cE_{f_1}[l(X_n)I[T > n]|X_0 = x], \end{aligned}$$

where assumption (ii) in the statement of the theorem was used to obtain the second inequality. Then, Assumption (2.3)(iii) implies that $h_1(\cdot) - h_2(\cdot) \leq 0$. Similarly, it can be shown that $h_1(\cdot) - h_2(\cdot) \geq 0$, and the result follows. ∎

LEMMA (3.3). *Let $\{V_n\}$ and $\{R_n\}$ be the sequences of value iteration and relative value functions in (2.4) and (2.7), respectively. Then for all $n \in \mathbb{N}$ and $x \in S$, (i) and (ii) below hold true.*

(i) $\left|(n+1)g + h(x) - V_n(x)\right| \leq 2\|r\| \sup_\pi E_\pi[l(X_{n+1})|X_0 = x]$,

and

(ii) $|R_n(x)| \leq 3\|r\|l(x)$.

Proof (i). Let $x \in S$ and $n \in \mathbb{N}$ be arbitrary. A standard induction argument using the AROE (2.3) yields

$$(n+1)g + h(x) = \sup_\pi E_\pi\left[\sum_{t=0}^{n} r(X_t, A_t) + h(X_{n+1})|X_0 = x\right].$$

This equation and (2.5) together imply, via Lemma 3.3 in [6], that

$$\begin{aligned} |(n+1)g + h(x) - V_n(x)| &\leq \sup_\pi E_\pi[|h(X_{n+1})||X_0 = x] \\ &\leq 2\|r\| \sup_\pi E_\pi[l(X_{n+1})|X_0 = x], \end{aligned}$$

where Lemma (2.1)(ii) was used to obtain the second inequality.

(ii) First observe that (2.5) implies that, for all $x \in S$, and $n, k \in \mathbb{N}$,

$$\begin{aligned} |V_n(x) - V_{n+k}(x)| &\leq |\sup_\pi E_\pi\left[\sum_{t=0}^{n} r(X_t, A_t)|X_0 = x\right] \\ &\quad - \sup_\pi E_\pi\left[\sum_{t=0}^{n+k} r(X_t, A_t)|X_0 = x\right]| \\ &\leq \sup_\pi |E_\pi\left[\sum_{t=n+1}^{n+k} r(X_t, A_t)|X_0 = x\right]| \end{aligned}$$

and then

$$|V_n(x) - V_{n+k}(x)| \leq k\|r\|. \tag{3.5}$$

Let $n \in \mathbb{N}$ and $x \in S$ be fixed, take the policy π^n as in (2.5) and set $T_n := \min\{n, T\}$; see (2.2) for the definition of T. Then

$$\begin{aligned} V_n(x) &= E_{\pi^n}\left[\sum_{t=0}^{n} r(X_t, A_t)|X_0 = x\right] \\ &= E_{\pi^n}\left[\sum_{t=0}^{T_n-1} r(X_t, A_t) + \sum_{t=T_n}^{n} r(X_t, A_t)|X_0 = x\right] \\ &= E_{\pi^n}\left[\sum_{t=0}^{T_n-1} r(X_t, A_t) + V_{n-T_n}(X_{T_n})|X_0 = x\right], \end{aligned}$$

where Bellman's optimality principle was used to obtain the last equality [5]. Thus,

$$\begin{aligned} |R_n(x)| &= |V_n(x) - V_n(z)| \\ &= |E_{\pi^n}\left[\sum_{t=0}^{T_n-1} r(X_t, A_t) + V_{n-T_n}(X_{t_n})|X_0 = x\right] - V_n(z)| \\ &\le \|r\|E_{\pi^n}[T_n|X_0 = x] + |E_{\pi^n}[V_{n-T_n}(X_{T_n})|X_0 = x] - V_n(z)| \\ &\le \|r\|E_{\pi^n}[T|X_0 = x] + E_{\pi^n}[|V_{n-T_n}(X_{T_n}) - V_n(z)||X_0 = x] \end{aligned} \tag{3.6}$$

Now observe that on the event $[T \le n]$, (a) $X_{T_n} = z$ and (b) $T = T_n$. Then (3.5) implies that

$$\begin{aligned} I[T \le n]|V_{n-T_n}(X_{T_n}) - V_n(z)| &= I[T \le n]|V_{n-T}(z) - V_n(z)| \\ &\le I[T \le n]\|r\|T. \end{aligned} \tag{3.7}$$

On the other hand, $T_n = n$ on the event $[T > n]$ and then

$$\begin{aligned} I[T > n]|V_{n-T_n}(X_{T_n}) - V_n(z)| &= I[T > n]|V_0(X_n) - V_n(z)| \\ &\le [1 + (n+1)]\|r\|I[T > n] \quad \text{(see(2.6))} \\ &\le 2(n+1)\|r\|I[T > n] \\ &\le 2T\|r\|I[T > n]. \end{aligned}$$

Combining this inequality with (3.7) it follows that $2\|r\|E_{\pi^n}[T|X_0 = x] \ge E_{\pi^n}[|V_{n-T_n}(X_{T_n}) - V_n(z)||X_0 = x]$ which, together with (3.6), yields that $3\|r\|E_{\pi^n}[T|X_0 = x] \ge |R_n(x)|$, and the result follows from Remark 2.2(i).

∎

LEMMA (3.4). *Assume that* $\{W_n : S \to \mathbb{R}\}$ *and* $W : S \to \mathbb{R}$ *satisfy the following. There exists a constant* c *such that,*

(i) for all $x \in S$ *and* $n \in \mathbb{N}$, $|W_n(x)| \le c \cdot l(x)$, *and*

(ii) $\lim_{n\to\infty} W_n(x) = W(x)$.

Let $x \in S$ be arbitrary and suppose that $\{a_n\} \subset A(x)$ converges to a. Then,

$$\lim_{n\to\infty} \sum_y p(y|x, a_n)W_n(y) = \sum_y p(y|x, a)W(y).$$

Proof From Assumptions (2.1) and (2.3) it follows that, for any $x \in S$, the mapping $a \mapsto \sum_y p(y|x, a)l(y) = \sum_{y\neq z} p(y|x, a)l(y) + p(z|x, a)l(z)$ is continuous and finite on $A(x)$. Then, an application of Proposition 18 in p. 232 of [10] yields the result. ∎

4. Proof of Theorem (2.1)

In this section a proof of Theorem (2.1) is given. Before going any further it is interesting to observe that Assumption (2.2) has not been used yet; however, it will play a central role in the argumentation below.

Proof of Theorem (2.1) (i) From (2.8), (3.2) and Lemma 3.3(i) it follows that

$$|g_n - g| = \left|\frac{1}{n+1}V_n(z) - g\right| \leq 2\|r\|\Delta(n) + \frac{1}{n+1}|h(z)| = 2\|r\|\Delta(n),$$

since $h(z) = 0$, and then Lemma (3.1)(ii) yields that $\lim_{n\to\infty} g_n = g$.

(ii) By (2.4) the following occurs: For all $k \in \mathbb{N}$, $x \in S$ and $a \in A(x)$,

$$V_k(x) \geq r(x, a) + \sum_y p(y|x, a)V_{k-1}(y)$$

and then,

$$\tilde{g}_k + R_k(x) \geq r(x, a) + \sum_y p(y|x, a)R_{k-1}(y), \tag{4.1}$$

where

$$\tilde{g}_k := V_k(z) - V_{k-1}(z);$$

notice that $\sum_{k=0}^n \tilde{g}_k = V_n(z) - V_{-1}(z) = V_n(z)$. Summing up from $k = 0$ to $n > 0$ in both sides of (4.1) it follows that

$$\begin{aligned} V_n(z) + \sum_{k=0}^n R_k(x) &\geq (n+1)r(x, a) + \sum_y p(y|x, a)\sum_{k=0}^n R_{k-1}(y) \\ &= (n+1)r(x, a) + \sum_y p(y|x, a)\sum_{k=0}^{n-1} R_k(y); \end{aligned}$$

recall that $R_{-1}(\cdot) = V_{-1}(\cdot) - V_{-1}(z) \equiv 0$. Then, by (2.8) and (2.9),

$$g_n + Q_n(x) \geq r(x,a) + \frac{n}{n+1}\sum_y p(y|x,a)Q_{n-1}(y). \tag{4.2}$$

On the other hand, (2.9) and Lemma (3.3)(ii) together yield that

$$\{Q_n(x) \mid |x \in S\} \in \Pi_{x\in S}[-3\|r\|l(x), \quad 3\|r\|l(x)] =: \mathbb{K}, \quad n \in \mathbb{N}. \tag{4.3}$$

Since $\mathbb{K}$ is compact (metric) in the product topology, it is sufficient to establish that any limit point of $\{Q_n\}$ coincides with h. With this in mind, let $Q \in \mathbb{K}$ be a limit point of $\{Q_n\}$ and select a subsequence $\{Q_{n_k}\}$ such that

$$\lim_{k\to\infty} Q_{n_k}(x) = Q(x), \quad x \in S. \tag{4.4}$$

Using that $Q_{n_k-1}(\cdot) = [(n_k+1)/n_k]Q_{n_k}(\cdot) - R_{n_k}(\cdot)/n_k$, (by (2.9)) and the pointwise boundedness of $\{R_n(\cdot)\}$ (see Lemma 3.3(ii)), it follows that

$$\lim_k Q_{n_k-1}(x) = Q(x), \quad x \in S. \tag{4.5}$$

Now, replace n by n_k in (4.2) and take limit as k goes to infinity in both sides of the resulting inequality. In this case, (4.4), (4.5) and Lemma (3.4) together yield

$$g + Q(x) \geq r(x,a) + \sum_y p(y|x,a)Q(y), \quad x \in S, \quad a \in A(x). \tag{4.6}$$

Next, let $f^* \in \mathbb{F}$ be an optimal stationary policy (cf. Lemma (2.1)(iv)) and let $w \in S$ be arbitrary but *fixed*. Define $\Phi\colon S \to \mathbb{R}$ by

$$\begin{aligned} &\Phi(x) := 0 \quad \text{if} \quad x \neq w, \\ &\Phi(w) := g + Q(w) - r\big(w, f^*(w)\big) - \sum_y p(y|w, f^*(w))Q(y), \end{aligned} \tag{4.7}$$

and notice that

(a) $\Phi(w) \geq 0$ (by (4.6)), and

(b) for all $x \in S$, $g + Q(x) \geq r\big(x, f^*(x)\big) + \Phi(x) + \sum_y p\big(y|x, f^*(x)\big)Q(y)$, with equality for $x = w$.

Then, an induction argument yields that, for all $n \in \mathbb{N}$ and $x \in S$,

$$g + \frac{1}{n+1}Q(x) \geq \frac{1}{n+1}E_{f^*}\Big[\sum_{t=0}^{n}\big(r(X_t, A_t) + \Phi(X_t)\big) + Q(X_{n+1})|X_0 = x\Big]. \tag{4.8}$$

On the other hand, (4.3) and (4.4) together with Lemma (3.1)(ii) imply that

$$\frac{1}{n+1}E_{f^*}[Q(X_{n+1})|X_0 = x] \to 0 \quad \text{as} \quad n \to \infty,$$

and taking limit as n goes to infinity in both sides of (4.8) it follows, via Remark (2.2), that

$$g \geq \sum_y q_{f^*}(y)\big(r(y, f^*(y)) + \Phi(y)\big) = g + q_{f^*}(w)\Phi(w),$$

where the equality is due to the optimality of f^* and (4.7); recall that q_{f^*} is the unique invariant distribution of f^*. Therefore,

$$0 \geq q_{f^*}(w)\Phi(w). \tag{4.9}$$

Finally, *by Assumption (2.2)*, $q_{f^*}(w) > 0$ occurs [4, pp. 39–42], and then, (4.9) implies that $\Phi(w) \leq 0$; since $\Phi \geq 0$, this yields $\Phi(w) = 0$ or, equivalently,

$$g + Q(w) = r\big(w, f^*(w)\big) + \sum_y p(y|w, f^*(w))Q(y);$$

see (4.7). Since $w \in S$ was arbitrary, this equality and (4.6) together imply that

$$g + Q(x) = \sup_{a \in A(x)} \left[r(x,a) + \sum_y p(y|x,a)Q(y) \right], \quad x \in S,$$

and an application of Lemma (3.2) yields that $Q(\cdot) = h(\cdot)$. In short, it has been established that any limit point of $\{Q_n\}$ coincides with h and, as already mentioned, this completes the proof of part (ii).

(iii) Let $f \in \mathbb{F}$ be a limit point of $\{f_n\}$, pick a subsequence $\{f_{n_k}\}$ such that

$$\lim_{k \to \infty} f_{n_k}(x) = f(x), \quad x \in S, \tag{4.10}$$

and observe that for arbitrary $k \in \mathbb{N}$, $x \in S$ and $a \in A(x)$,

$$r(x,a) + \sum_y p(y|x,a)Q_{n_k}(y) \leq r\big(x, f_{n_k}(x)\big) + \sum_y p(y|x, f_{n_k}(x))Q_{n_k}(y).$$

Taking limit as k goes to infinity in both side of this inequality it follows, combining (4.10), part (ii) and Lemma 3.4, that for all $x \in S$ and $a \in A(x)$,

$$r(x,a) + \sum_y p(y|x,a)h(y) \leq r\big(x, f(x)\big) + \sum_y p(y|x, f(x))h(y),$$

so that, by Lemma (2.1)(iv), f is optimal. ∎

5. Concluding remarks

The value iteration procedure has been studied in the context of communicating MDP's endowed with the average reward criterion and bounded rewards. The convergences in Theorem (2.1) are weaker than those obtained in other papers, but LFC in the basic Assumption (2.3) is also weaker than the conditions usually imposed to obtain the convergence of both the relative value functions $\{R_n\}$ and $\{V_n(z) - V_{n-1}(z)\}$. One exception to this remark is a recent paper by Sennott [11] where the recurrence assumptions are extremely weak, but a very special combination in the transition-reward structure of the model is required; such a condition seems to be very difficult to verify in a general context. On the other hand there are, at least, three interesting problems to be considered. First, notice that Assumption (2.2) played an important role in the proof of Theorem (2.1), and that the Lyapunov function condition can be formulated to include unbounded rewards [7, 13]. Then, it is interesting to ask if it is possible to obtain a result similar to Theorem (2.1) when (i) Assumption (2.2) is violated, or (ii) the reward fuction is unbounded. On the other, a third interesting problem consists in investigating if the results in Theorem (2.1) can be improved to obtain, under LFC, convergence of $\{R_n\}$ and $\{V_n(z) - V_{n-1}(z)\}$. Research in these directions is presently in progress.

ACKNOWLEDGEMENT. The author is grateful to three unknown reviewers for their careful reading of the original manuscript and helpful suggestions to improve the presentation of paper.

DEPARTAMENTO DE ESTADÍSTICA Y CÁLCULO
UNIVERSIDAD AUTÓNOMA AGRARIA ANTONIO NARRO
BUENAVISTA, SALTILLO, COAH 25315
MÉXICO

REFERENCES

[1] D. P. BERTSEKAS, Dynamic Programming: Deterministic and Stochastic Models, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[2] R. CAVAZOS-CADENA AND O. HERNÁNDEZ-LERMA, *Equivalence of Lyapunov stability criteria in a class of Markov decision processes,* J. Appl. Math. and Optimization, **26**, 113–137, (1992).

[3] A. FEDERGRUEN AND H. C. TIJMS, *The optimality equation in average cost denumerable state semi-markov decision problems, recurrency conditions and algorithms,* J. Appl. Probab., **15**, 356-373 (1978).

[4] M. LOÈVE, Probability Theory I, Springer, New York, 1977.

[5] O. HERNÁNDEZ-LERMA, Adaptive Markov Control Processes, Springer, New York, 1989.

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

[7] A. HORDIJK, *Dynamic Programming and Potential Theory,* Mathematical Centre tract **51**, Mathematisch Centrum, Amsterdam, 1974.

[8] A. HORDIJK AND P. J. SCHWEITZER, *The Asymptotic behaviour of the minimal total expected cost for the denumerable state Markov decision model,* J. Appl. Probab., **12**, 298–305 (1975).

[9] S. M. Ross, Applied Probability Models with Optimization Applications, Holden-Day, San Francisco, 1970.
[10] H. L. Royden, Real Analysis, Macmillan, New York, 1968.
[11] L.I. Sennott, *Value iteration in infinite state average cost Markov decision processes with unbounded costs,* Annals of Oper. Res., **28**, 261–272, (1991).
[12] P. J. Schweitzer and A. Federgruen, *The Asymptotic behavior of undiscounted value iteration in Markov decision problems,* Math. Oper. Res., **2**, 360–381, (1977).
[13] L. C. Thomas, *Conectedness conditions for denumerable state Markov decision processes,* in: R. Hartley, L. C. Thomas and D. J. White, Eds., Recent Developments in Markov Decision Processes, Academic Press, New York, 1980, 181–204.