Control
and Cybernetics
VOL. 16 (1987) No. 1

Finite-state approximations and adaptive control of discounted Markov decision processes with unbounded rewards*

by

ROLANDO CAVAZOS-CADENA
Departamento de Mathematicas
Centro de Investigacion del I.P.N.
Apartado Postal 14-740
07000, Mexico, D.F., Mexico

In this paper we consider discounted, unbounded rewards, denumerable state Markov decision processes which depend on unknown parameters. Following the approach of Hernández-Lerma and Marcus [6], we combine the nonstationary value iteration scheme of Federgruen and Schweitzer [4] with a finite-state procedure introduced in [1], to obtain a finite-state iterative method, which in the presence of a strongly consistent method of estimation, is used to find the optimal total expected discounted reward corresponding to the true parameter value. Also, an adaptive policy with asymptotic optimality properties is proposed.

1. Introduction and summary

In this paper we deal with denumerable state, discounted, unbounded rewards Markov decision processes which depend on unknown parameters. We consider the problems of determining: (i) a finite-state iterative method to find the optimal total expected discounted reward corresponding to the true parameter value, and (ii) adaptive policies with asymptotic optimality properties. We follow the approach in [6] where these problems were solved for the finite-state, bounded rewards case.

Let (S, A, p, r, β) be the usual discounted Markov decision process (MDP) where S is the state space, assumed to be an arbitrary non-empty denumerable set endowed with the discrete topology, A is the action (or control)

* This research was supported in part by the Consejo del Sistema Nacional de Educación Tecnológica (COSNET) under Grant 178/84 and in part by the Universidad Autónoma Agraria "Antonio Narro".

set, assumed to be a metric space endowed with the Borel σ-field; for $x \in S$, $A(x) \subset A$ is the measurable (non-empty) set of admissible actions at state x. Now, let K be defined by $K := \{(x, a) | a \in A(x), x \in S\}$. $r: K \to R$ is the (measurable) reward function and p is the transition law; that is, if the present state is x and an action $a \in A(x)$ is selected, an immediate (expected) reward $r(x, a)$ is obtained and the next state will be y with probability $p(x, y, a)$. β is the discount factor and we suppose that $0 < \beta < 1$.

Let Θ be a metric space and suppose that, for each $\theta \in \Theta$, we have a MDP $(S, A, p(\theta), r(\theta), \beta)$ with transition probabilities $p(x, y, a, \theta)$ and rewards $r(x, a, \theta)$ depending on θ. For $n = 0, 1, 2, ..., X_n$ and A_n denote (respectively) the state and the action at stage n, while $I_n = (X_0, A_0, ..., X_{n-1}, A_{n-1}, X_n)$ stands for the history of the process -or information vector- up to stage n. I_n is a random vector taking values in H_n, where $H_0 := S$ and for $n \geq 1$, $H_n := K \times H_{n-1}$. Let $(x_n, a_n, n = 0, 1, 2, ...) \in H_\infty$ be a given realization of the state and action sequences. In this case, we write $i_n = (x_0, a_0, ..., x_{n-1}, a_{n-1}, x_n)$ for the corresponding history up to time n.

A (randomized) **policy** $D = \{D_n | n = 0, 1, 2, ...\}$ is a sequence such that, for each n, D_n is a function defined on H_n and taking values in the set of probability measures defined on the Borel σ-field of A, in such a way that, for each $i_n \in H_n$, $D_n(\cdot | i_n)$ is concentrated on $A(x_n)$. A policy D is said to be **deterministic** if, for every n and $i_n \in H_n$, $D_n(\cdot | i_n)$ is concentrated on a single point of $A(x_n)$. The class of all deterministic policies will be denoted by $\mathscr{D}$. A policy $D \in \mathscr{D}$ is said to be a **Markov** — or memoryless — policy if, for every n and $i_n \in H_n$, $D_n(\cdot | i_n)$ depends only on the present state x_n; that is, for a Markov policy there exists a sequence $\{f_n: S \to A | f_n(x) \in A(x), x \in S, n = 0, 1, 2, ...\}$ such that $D_n(\cdot | i_n)$ is concentrated on $f_n(x_n)$ and, if we have that $f_0 = f_1 = f_2 = ...$, the Markov policy is said to be a **stationary** policy. It is clear that the class of stationary policies can be naturally identified with the cartesian product $\prod_{x \in S} A(x)$.

For every policy D, $x \in S$ and $\theta \in \Theta$, let

$$v(D, \theta)(x) := \sum_{n=0}^{\infty} \beta^n E_x^{D,\theta} r(X_n, A_n, \theta) \tag{1.1}$$

be the total expected discounted reward when policy D is employed, x is the initial state and θ is the parameter value, and let

$$v^*(\theta)(x) := \sup_D v(D, \theta)(x) \tag{1.2}$$

be the optimal expected discounted reward when θ is the parameter value and the initial state is x. (The supremum in (1.2) is taken over all policies). If $r(\cdot, \cdot, \theta)$ is bounded, there is no difficulty to see that the series defining $v(D, \theta)$ is well defined and, in this case, $v(D, \theta)$ is uniformly bounded on S and so is $v^*(\theta)$; moreover, $v^*(\theta)(x)$ is the supremum of $v(D, \theta)(x)$ over all

stationary policies and $v^*(\theta)$ can be obtained by the well known method of succesive approximations ([11], chapter 6). On the other hand, if $r(\cdot,\cdot,\theta)$ is **not** bounded, conditions must be imposed in order to have $v(D,\theta)(x)$ well defined. Several sets of sufficient conditions have been proposed; see, for instance, [1], [5], [8] and [14]. The conditions in [1] and some results obtained there are briefly sketched in section 2.

DEFINITION 1.1. (i) A policy D is **discount optimal** (when θ is the parameter value) if

$$v(D,\theta)(x) = v^*(\theta)(x) \quad \text{for every } x \in S.$$

(ii) A policy D is **asymptotically discount optimal in the sense of Schäl** (ADOS) (when θ is the parameter value) if, for each $x \in S$,

$$v_N(D,\theta)(x) - E_x^{D,\theta} v^*(\theta)(X_N) \to 0 \quad \text{as } N \to \infty,$$

where, for $x \in S$, $N = 0, 1, 2, \ldots$,

$$v_N(D,\theta)(x) := \sum_{n=N}^{\infty} \beta^{n-N} E_x^{D,\theta} r(X_n, A_n, \theta)$$

is the total expected reward from stage N onwards discounted from stage N.

Now, we can pose our problems as follows: Given that the true parameter value, say $\theta^* \in \Theta$, is fixed but **unknown**, determine

(A) a finite-state iterative scheme to find $v^*(\theta^*)$, and

(B) adaptive policies with asymptotic optimality properties, for instance, ADOS.

Our approach parallels the work of Hernández-Lerma and Marcus in [6], where these problems were solved for the **bounded** rewards, **finite** state case. Their solution is a straightforward application of the Non-stationary Value Iteration (NVI) scheme of Federgruen and Schweitzer [4], which is a variant of the method of succesive approximations ([11], chapter 6). In the same way, under Assumptions 4.1–4.3, we obtain a solution to problems (A) and (B) as a by-product of the Truncated Non-stationary Value Iteration (TNVI) scheme introduced in section 3.

The TNVI scheme is a combination of the finite-state approximation method proposed in [1] and the NVI scheme. In order to have that the TNVI scheme is useful to solve our problems, we need a sequence converging to the true parameter value. So, we suppose that, as the system is in progress, the controller can use the registered history to obtain such a sequence (cf. Assumption 4.1). Finally, under conditions regarding continuous dependence on θ of $r(\cdot,\cdot,\theta)$ and $p(\cdot,\cdot,\cdot,\theta)$ (cf. Assumptions 4.2 and 4.3), we obtain a solution to our problems using the results on the TNVI scheme obtained in section 3.

2. Preliminaries

Because we deal with discounted MDP's with **unbounded** rewards, we need to impose conditions to warrant that $v(D, \theta)(x)$ is well defined. In this section we sketch briefly the conditions proposed in [1] and some results obtained there which will be useful later.

DEFINITION 2.1. Let $\theta \in \Theta$.
(i) For each nonnegative extended function $u: S \to \bar{\mathbf{R}}^+$, define $H_\theta u: S \to \bar{\mathbf{R}}^+$ as follows:

$$H_\theta u(x) := \sup_{a \in A(x)} \sum p(x, y, a, \theta) u(y), \qquad x \in S,$$

where the summation is over all $y \in S$; see Remark 2.1 below. We write $H_\theta^0 u := u$ and for $n = 1, 2, \ldots$

$$H_\theta^n u := H_\theta (H_\theta^{n-1} u)$$

(ii) $R_\theta: S \to \bar{\mathbf{R}}^+$ is defined by

$$R_\theta(x) := \sup_{a \in A(x)} |r(x, a, \theta)|, \qquad x \in S.$$

(iii) $\mathscr{R}_\theta: S \to \bar{\mathbf{R}}^+$ is defined as follows:

$$\mathscr{R}_\theta(x) := \sum_{n=0}^{\infty} \beta^n H^n R_\theta(x), \qquad x \in S.$$

REMARK 2.1. A convention about the symbol $\sum$ will be used consistently throughout the following: For $(x, a) \in K$, $\theta \in \Theta$ and $V: S \to \bar{\mathbf{R}}$

$$\sum p(x, y, a, \theta) V(y) := \sum_{y \in S} p(x, y, a, \theta) V(y)$$

whenever the right-hand side is well defined, that is, whenever the series in the right-hand side is absolutely convergent or V has constant sign.

The condition imposed in [1] to have that $v(D, \theta)(x)$ is well defined is that $\mathscr{R}_\theta(x) < \infty$ for every $x \in S$.

THEOREM 2.1. *Let $\theta \in \Theta$ and suppose that $\mathscr{R}_\theta(x) < \infty$ for every $x \in S$. Then, for every policy D and $x \in S$,*

(i) $E_x^{D,\theta} |r(X_n, A_n, \theta)| \leqslant H_\theta^n R_\theta(x)$ for $n = 0, 1, 2, \ldots$

(ii) $|v(D, \theta)(x)| \leqslant \mathscr{R}_\theta(x)$, *where $v(D, \theta)(x)$ is defined by* (1.1) *and the series appearing there is absolutely convergent.*

(iii) $|v^*(\theta)(x)| \leqslant \mathscr{R}_\theta(x)$, *where $v^*(\theta)(x)$ is defined by* (1.2). *Define $\mathscr{L}_\theta$ as follows:*

$$\mathscr{L}_\theta := \{u: S \to \mathbf{R} \mid |u(x)| \leqslant \mathscr{R}_\theta(x), x \in S\}.$$

(iv) *For every $u \in \mathscr{L}_\theta$ and $(x, a) \in K$,*

(a) $\sum p(x, y, a, \theta)\, u(y)$ *converges absolutely*
(b) $|r(x, a, \theta)+\beta \sum p(x, y, a, \theta)\, u(y)| \leqslant \mathscr{R}_\theta(x)$

Define $T_\theta: \mathscr{L}_\theta \to \mathscr{L}_\theta$ *as follows: for every* $u \in \mathscr{L}_\theta$ *and* $x \in S$,

$$T_\theta u(x) := \sup_{a \in A(x)} [r(x, a, \theta)+\beta \sum p(x, y, a, \theta)\, u(y)] \tag{2.1}$$

(v) $T_\theta v^*(\theta) = v^*(\theta)$. (2.2)

Moreover, $v^*(\theta)$ is the unique fixed point of T_θ and it can be obtained by succesive approximations, that is, for every $u \in \mathscr{L}_\theta$ and every $x \in S$,

$$T_\theta^n u(x) \to v^*(\theta)(x) \qquad \text{as } n \to \infty.$$

In short, Theorem 2.1 asserts that $\mathscr{R}_\theta(x) < \infty$ for every $x \in S$, implies that $v(D, \theta)$ is well defined, $v^*(\theta)$ is the unique fixed point of the operator T_θ defined by (2.1), and $v^*(\theta)$ can be obtained by succesive approximations. Equation (2.2) is known as the **optimality equation**. A proof of this theorem can be found in [1], section 2; indeed, part (i)–(iii) are Theorem 2.1, (iv) is part (b) in Corollary 2.1 and (v) is Theorem 2.2 in [1]. Observe that T_θ is **monotone**, that is, $u, v \in \mathscr{L}_\theta$ and $u \leqslant v$ together imply $T_\theta u \leqslant T_\theta v$. Finally, although this will be not used later, we note that under the assumption in Theorem 2.1, the supremum in the definition of $v^*(\theta)$ can be taken only over all **stationary** policies (cf. [1], Theorem 2.3).

We are going to deal with several operators T_θ simultaneously. To handle this situation we introduce the following definition:

DEFINITION 2.2. Let M be a non-empty subset of Θ.

(i) $R_M: S \to \overline{\mathbf{R}}^+$ is defined by

$$R_M(x) := \sup_{\theta \in M} R_\theta(x), \qquad x \in S.$$

(ii) For each $u: S \to \overline{\mathbf{R}}^+$, $H_M u: S \to \overline{\mathbf{R}}^+$ is defined

$$H_M u(x) := \sup_{\theta \in M} H_\theta u(x), \qquad x \in S.$$

$$H_M^0 u := u \qquad \text{and for} \qquad n = 1, 2, \ldots,$$

$$H_M^n u := H_M(H_M^{n-1} u)$$

(iii) $\mathscr{R}_M: S \to \overline{\mathbf{R}}^+$ is defined by

$$\mathscr{R}_M(x) := \sum_{n=0}^{\infty} \beta^n H_M^n R_M(x), \qquad x \in S.$$

(iv) $\mathscr{L}_M$ is defined as follows:

$$\mathscr{L}_M := \{u: S \to \mathbf{R} \mid |u(x)| \leqslant \mathscr{R}_M(x), x \in S\}.$$

We need the following properties of H_M:

LEMMA 2.1. *Let* $u, w: S \to \overline{\mathbf{R}}^+$, M *a non-empty subset of* Θ, *and* $c > 0$. *Then, the following hold:*

(*i*) $H_M(cu) = cH_M(u)$ (*Homogeneity*)

(*ii*) $H_M u \leqslant H_M w$ *if* $u \leqslant w$ (*Monotonicity*).

(*iii*) *If* $H_M u(x) < \infty$ *for every* $x \in S$ *and* $v: S \to \mathbf{R}$ *satisfies that* $|v(x)| \leqslant u(x)$ *for every* $x \in S$, *then*

a) For every $\theta \in M$ *and* $(x, a) \in K$

$\sum p(x, y, a, \theta) v(y)$ *is absolutely convergent.*

b) For every $x \in S$

$$\sup_{a \in A(x), \theta \in M} \left| \sum p(x, y, a, \theta) v(y) \right| \leqslant H_M u(x).$$

(*iv*) *Let* $u_n: S \to \overline{\mathbf{R}}^+$, $n = 0, 1, 2, \ldots$. *Then, for* $k = 0, 1, 2, \ldots$,

$$H_M^k \left(\sum_{n=0}^{\infty} u_n \right) \leqslant \sum_{n=0}^{\infty} H_M^k u_n \qquad \text{(Subadditivity)}.$$

(*v*) *For* $k = 0, 1, 2, \ldots$,

$$c^k H_M^k \left(\sum_{n=0}^{\infty} c^n H_M^n u \right) \leqslant \sum_{n=k}^{\infty} c^n H_M^n u.$$

For the case in which M is a singleton, Lemma 2.1 is proposition 2.1 in [1]. The proof given there still applies in our present case if every time $\sup_{a \in A(x)}$ appears, we substitute it by $\sup_{a \in A(x),\, \theta \in M}$. It is clear that $\mathscr{R}_\theta(x) \leqslant \mathscr{R}_M(x)$ for every $x \in S$ and $\theta \in M$ and then, if $\mathscr{R}_M(x)$ is always finite, the conclusions of Theorem 2.1 are valid for every $\theta \in M$ and, in this case, the next theorem shows that T_θ can be extended to $\mathscr{L}_M$ and $v^*(\theta)$ is still the unique fixed point of T_θ.

THEOREM 2.2. *Suppose that* M *is a non-empty subset of* Θ *and* $\mathscr{R}_M(x) < \infty$ *for every* $x \in S$. *Then, for each* $\theta \in M$,

(i) *Using* (2.1), $T_\theta u$ *can be defined for every* $u \in \mathscr{L}_M$ *and, in this case,*

$$T_\theta u \in \mathscr{L}_M \qquad \text{for} \qquad u \in \mathscr{L}_M.$$

(ii) *For every* $x \in S$,

$$\beta^k H_M^k \mathscr{R}_M(x) \to 0 \qquad \text{as} \qquad k \to \infty.$$

(iii) $v^*(\theta)$ *is the unique fixed point of* T_θ *in* $\mathscr{L}_M$. *Moreover,* $v^*(\theta)$ *can be obtained by successive approximations, that is, for every* $u \in \mathscr{L}_M$ *and* $x \in S$,

$$T_\theta^n u(x) \to v^*(\theta)(x) \qquad \text{as} \qquad n \to \infty.$$

(iv) T_θ *is monotone on* $\mathscr{L}_M$; *that is,* $u, v \in \mathscr{L}_M$ *and* $u \leqslant v$ *imply*

$$T_\theta u \leqslant T_\theta v.$$

Proof. (i) By part (v) of Lemma 2.1 with $c = \beta$, $u = R_M$ and $k = 1$, we have that, for every $x \in S$,

$$\beta H_M \mathscr{R}_M(x) \leqslant \mathscr{R}_M(x) - R_M(x) < \infty$$

Then, by part (iii) of the same Lemma, the sum in the right hand side of (2.1) is well defined as soon as we have $|u(x)| \leqslant \mathscr{R}_M(x)$ for every $x \in S$, that is, as soon as $u \in \mathscr{L}_M$ and, in this situation, we obtain the following inequalities where $x \in S$ and sup is taken over $a \in A(x)$:

$$\begin{aligned} |\sup [r(x,a,\theta) + \beta \Sigma p(x,y,a,\theta)\, u(y)]| &\leqslant \sup |r(x,a,\theta)| + \\ &+ \beta \sup \Sigma p(x,y,a,\theta)\, |u(y)| \leqslant \\ &\leqslant R_\theta(x) + \beta H_\theta |u|(x) \leqslant \\ &\leqslant R_M(x) + \beta H_M \mathscr{R}_M(x) \leqslant \\ &\leqslant R_M(x) + \mathscr{R}_M(x) - R_M(x) = \mathscr{R}_M(x). \end{aligned}$$

So, $T_\theta u(x)$ can be defined by (2.1) for $u \in \mathscr{L}_M$ and in this case, we have that $T_\theta u \in \mathscr{L}_M$.

(ii) By part (v) of Lemma 2.1 with $c = \beta$ and $u = R_M$, we have that, for every $x \in S$ and $k = 0, 1, 2, \ldots$,

$$\beta^k H_M^k \mathscr{R}_M(x) \leqslant \sum_{n=k}^{\infty} \beta^n H_M^n R_M(x)$$

Now, the result follows from the convergence of the series defining $\mathscr{R}_M(x)$.

(iii) We will use the following fact:

If w and z are real valued functions bounded from above on a set A, then,

$$|\sup_{a \in A} w(a) - \sup_{a \in A} z(a)| \leqslant \sup_{a \in A} |w(a) - z(a)| \tag{2.3}$$

Let $u, v \in \mathscr{L}_M$ and $x \in S$. Using (2.3) we obtain the following inequalities, where sup is taken over $a \in A(x)$:

$$\begin{aligned} |T_\theta u(x) - T_\theta v(x)| &\leqslant \sup \left|\beta \Sigma p(x,y,a,\theta)\,(u(y) - v(y))\right| \leqslant \\ &\leqslant \beta \sup \Sigma p(x,y,a,\theta)\, |u(y) - v(y)| = \beta H_\theta |u - v|(x). \end{aligned}$$

Then, because $x \in S$ is arbitrary, we have

$$|T_\theta u - T_\theta v| \leqslant \beta H_\theta |u - v|, \qquad u, v \in \mathscr{L}_M, \tag{2.4}$$

and an introduction argument gives

$$|T_\theta^n u - T_\theta^n v| \leqslant \beta^n H_\theta^n |u - v| \qquad \text{for} \qquad n = 1, 2, \ldots.$$

Now, using the fact that $|u - v| \leqslant 2\mathscr{R}_M$ we obtain

$$|T_\theta^n u - T_\theta^n| \leqslant 2\beta^n H_\theta^n \mathscr{R}_M \leqslant 2\beta^n H_M^n \mathscr{R}_M \qquad \text{for} \qquad n = 1, 2, \ldots,$$

and then, by part (ii) of this Theorem,

$$\lim_{n \to \infty} |T_\theta^n u(x) - T_\theta^n v(x)| = 0, \qquad x \in S. \tag{2.5}$$

Take $v = v^*(\theta)$. Because $T_\theta^n v^*(\theta) = v^*(\theta)$, $n = 1, 2, \ldots$, we get from (2.5) that,

$$\lim_{n\to\infty} T_\theta^n u(x) = v^*(\theta)(x), \qquad u \in \mathscr{L}_M, \qquad x \in S. \tag{2.6}$$

Then, $v^*(\theta)$ can be obtained by succesive approximations and is the unique fixed point of T_θ in $\mathscr{L}_M$. Indeed, if $u \in \mathscr{L}_M$ and $T_\theta u = u$, we have that $T_\theta^n u = u$ for $n = 1, 2, \ldots$, in which case (2.6) implies that $u = v^*(\theta)$. This completes the proof, since (iv) is clear. ■

We need to estimate the difference between two operators T_θ and T_τ for parameter values $\theta, \tau \in \Theta$. To handle this situation, we introduce the following definition.

DEFINITION 2.3. Let M be a non-empty subset of Θ and suppose that $\mathscr{R}_M(x) < \infty$ for every $x \in S$.

(i) For $x \in S$, $\tau, \theta \in M$,

$$E(x, M, \tau, \theta) := \sup_{a \in A(x)} \big[|r(x, a, \tau) - r(x, a, \theta)| + \\ + \beta\Sigma\, |p(x, y, a, \tau) - p(x, y, a, \theta)|\, \mathscr{R}_M(y)\big],$$

where the summation is over $y \in S$.

(ii) For $F \subset S$, $\tau, \theta \in M$,

$$E(F, M, \tau, \theta) := \sup_{x \in F} E(x, M, \tau, \theta).$$

THEOREM 2.3. *Let M be a non-empty subset of Θ and suppose that $\mathscr{R}_M(x) < \infty$ for every $x \in S$. Then, for every $u \in \alpha_M$, $\tau, \theta \in M$ and $x \in S$,*

(i) $|T_\tau u(x) - T_\theta u(x)| \leqslant E(x, M, \tau, \theta)$.

(ii) $\sup_{x \in F} |T_\tau u(x) - T_\theta u(x)| \leqslant E(F, M, \tau, \theta)$.

Proof. Part (ii) follows immediately from (i). To prove (i), we use (2.3) to obtain the following inequalities, where the summation is over $y \in S$ and sup is taken over $a \in A(x)$:

$$|T_\tau u(x) - T_\theta u(x)| \leqslant \sup \big|r(x, a, \tau) - r(x, a, \theta) + \beta\Sigma \big(p(x, y, a, \tau) - \\ - p(x, y, a, \theta)\big) u(y)\big| \leqslant \sup \big[|r(x, a, \tau) - r(x, a, \theta)| + \\ + \beta\Sigma\, |p(x, y, a, \tau) - p(x, y, a, \theta)|\, \mathscr{R}_M(y)\big] = E(x, M, \tau, \theta).$$ ■

3. The truncated non-stationary value iteration scheme

Throughout this section $F_0, F_1, F_2, \ldots$, is a (fixed) sequence of subsets of S. We suppose that

(i) $F_n \subset F_{n+1}$, $\quad n = 0, 1, 2, \ldots$

(ii) $\bigcup_{n=0}^{\infty} F_n = S$.

DEFINITION 3.1. (The Truncated Non-stationary Value Iteration scheme). Let $\{\theta_n | n = 0, 1, 2, ...\}$ be a convergent sequence in Θ. Define M by

$$M := \{\theta, \theta_0, \theta_1, \theta_2, ...\} \quad \text{where} \quad \theta = \lim_{n\to\infty} \theta_n$$

Suppose that $\mathscr{R}_M(x) < \infty$ for every $x \in S$. Finally, let $u \in \mathscr{L}_M$. The sequence $\{v_n : S \to \mathbf{R} | n = -1, 0, 1, 2, ...\}$ is defined as follows:

$$v_{-1} := u, \quad \text{and for} \quad n \geqslant 0,$$

$$v_n(x) := \sup_{a\in A(x)} [r(x, a, \theta_n) + \beta \Sigma p(x, y, a, \theta_n) v_{n-1}(y)] \quad \text{if} \quad x \in F_n,$$

$$v_n(x) := u(x) \quad \text{if} \quad x \notin F_n$$

REMARK 3.1. The iterative scheme in Definition 3.1 will be called the TNVI scheme. In the case when M is a singleton, the TNVI scheme is nothing but White's extended scheme introduced in [1]. On the other hand, if S is a finite set, $u \equiv 0$ and $F_n = S$ for every n, we obtain the NVI scheme of Federgruen and Schweitzer introduced in [4], which for the case of **bounded** rewards was used by Hernández-Lerma and Marcus in [6] to solve problems (A) and (B) posed in section 1. Now, using the fact that $u \in \mathscr{L}_M$, a simple induction argument gives that $v_n \in \mathscr{L}_M$ for $n = 0, 1, 2, ...$ and then, the sums appearing in Definition 3.1 are well defined. The function u will be referred to as the **seed** of the scheme and will always belong to $\mathscr{L}_M$. On the other hand, from a computational viewpoint, it seems desirable to take $u \equiv 0$ but, although this will be done in section 4, we prefer to maintain u arbitrary at this moment and study the relevance of u in relation to the convergence of $\{v_n\}$ to $v^*(\theta)$. Finally, note that once we have selected the sets F_n, $n = 0, 1, 2, ...$ and the seed u that are going to be used in the TNVI scheme, $v_n(x)$ depends only on $(\theta_0, ..., \theta_n)$. To emphasize this dependence, we sometimes write $v_n(\theta_0^n)(x)$ instead of $v_n(x)$ where $\theta_0^n := (\theta_0, ..., \theta_n)$ (cf. section 4). However, in this section we consider a **fixed** convergent sequence $\{\theta_n\}$ and then we simply write $v_n(x)$.

The idea in the TNVI scheme is to produce approximations to $v^*(\theta)$. Our first result concerning the limit points of $\{v_n\}$ is the following.

THEOREM 3.1. *Suppose that*

$$E(x, M, \theta_n, \theta) \to 0 \quad \text{as} \quad n \to \infty \quad \text{for every} \quad x \in S. \tag{3.1}$$

Then, for every seed u, and every $x \in S$,

$$\liminf v_n(x) \geqslant v^*(\theta)(x). \tag{3.2}$$

Proof. Let $\mathbf{l}(x) := \liminf v_n(x)$, $x \in S$. We note that $|v_n(x)| \leqslant \mathscr{R}_M(x)$ for every $x \in S$, implies that $|\mathbf{l}(x)| \leqslant \mathscr{R}_M(x)$ for every $x \in S$ and then, $\mathbf{l} \in \mathscr{L}_M$. Now, let $x \in S$ and select m such that $x \in F_m$. Then, for $n \geqslant m$ and $a \in A(x)$, we have, from the definition of v_n, that

$$v_n(x) \geq r(x, a, \theta_n) + \beta \Sigma p(x, y, a, \theta_{n-1})\, v_{n-1}(y) =$$
$$= [r(x, a, \theta_n) - r(x, a, \theta) + \beta\Sigma (p(x, y, a, \theta_n) -$$
$$- p(x, y, a, \theta))\, v_{n-1}(y)] + r(x, a, \theta) + \beta \Sigma p(x, y, a, \theta)\, v_{n-1}(y).$$

Using the fact that $|v_{n-1}| \leq \mathscr{R}_M$ and the definition of $E(x, M, \theta_n, \theta)$, we see that, for $n \geq m$,

$$v_n(x) \geq -E(x, M, \theta_n, \theta) + r(x, a, \theta) + \beta \Sigma p(x, y, a, \theta)\, v_{n-1}(y).$$

Taking lim inf as $n \to \infty$ in both sides of the above inequality we have, using (3.1) that,

$$\mathbf{l}(x) \geq r(x, a, \theta) + \beta \liminf \Sigma p(x, y, a, \theta)\, v_{n-1}(y). \tag{3.3}$$

Next, we note that $\Sigma p(x, y, a, \theta)\, \mathscr{R}_M(y) \leq H_\theta\, \mathscr{R}_M(x) \leq H_M\, \mathscr{R}_M(x) < \infty$ and $|v_{n-1}| \leq \mathscr{R}_M$. These facts allow us to use Fatou's Lemma, in which case we conclude that

$$\liminf \Sigma p(x, y, a, \theta)\, v_{n-1}(y) \geq \Sigma p(x, y, a, \theta)\, \mathbf{l}(y),$$

and this inequality, in combination with (3.3), gives

$$\mathbf{l}(x) \geq r(x, a, \theta) + \Sigma p(x, y, a, \theta)\, \mathbf{l}(y).$$

Now, taking the supremum over $a \in A(x)$, we obtain

$$\mathbf{l}(x) \geq T_\theta \mathbf{l}(x).$$

Using the arbitraryness of x and the monotonicity property of T_θ, it is easily seen that

$$\mathbf{l} \geq T_\theta^n \mathbf{l} \qquad n = 0, 1, 2, \ldots$$

and the result follows from Theorem 2.2 (iii). ■

REMARK 3.2. In our approach, a continuity requirement like (3.1) seems unavoidable if we are going to have $v_n(x) \to v^*(\theta)(x)$ for every $x \in S$. When (3.1) as well as some restrictions on the tails of the transition probabilities hold, we can show that, for every seed u, $\{v_n\}$ converges pointwise to $v^*(\theta)$ (cf. Theorem 3.3). On the other hand, suppose that M is a singleton. In this case, it was shown in [1] that if the seed u satisfies $u \leq T_\theta u$, then $\{v_n\}$ converges pointwise to $v^*(\theta)$ and then it might be thought that the same occurs if (3.1) holds and the seed u is appropriately chosen. The example below shows that is not the case.

EXAMPLE 3.1. Suppose the following:

(i) $\Theta = [0, 1]$
(ii) $S = \mathbf{N} := \{1, 2, \ldots\}$
(iii) $A(x) = [0, 1]$, $x \in \mathbf{N}$
(iv) $r(x, a, \theta) = \theta^{1/x}$, $x \in \mathbf{N}$, $a \in [0, 1]$
(v) $p(x, y, a, \theta) \equiv p(x, y, a)$ satisfies that,

given $x \in \mathbf{N}$ and $n = 0, 1, 2, \ldots$, there exists $a^* \in A(x)$ such that

$$\sum_{n/2 \leqslant y \leqslant n+1} p(x, y, a^*) = 1 \tag{3.4}$$

Now, take $F_n = \{1, \ldots, n+1\}$, $n = 0, 1, 2, \ldots$ and let u be an **arbitrary** seed (of course, in this case u is bounded).
Finally, let $\{\theta_n\}$ be a sequence in $(0, 1]$ tending to 0.

In this situation we are going to show that, if $\{\theta_n\}$ goes to 0 slowly enough, we have **strict** inequality in (3.2) for **every** seed u. A simple induction argument, using the fact that $r \geqslant 0$ and condition (v) with $n = 0$, shows that

$$v_n(x) \geqslant \beta^{n+1} u(1) \quad \text{for} \quad n = 0, 1, 2, \ldots \quad \text{and} \quad x \in F_n.$$

Then, for $n = 0, 1, 2, \ldots$ and $x \in F_{n+1}$,

$$v_{n+1}(x) \geqslant r(x, a^*, \theta_{n+1}) + \beta \Sigma p(x, y, a^*)\, v_n(y) \geqslant \theta_{n+1}^{1/x} + \beta^{n+2} u(1)$$

where a^* satisfies (3.4), and then, for $n = 1, 2, \ldots$ and $x \in F_{n+1}$

$$v_{n+1}(x) \geqslant \beta \Sigma p(x, y, a^*) \left(\theta_n^{1/x} + \beta^{n+1} u(1)\right) \geqslant \beta (\theta_n)^{2/n} + \beta^{n+2} u(1).$$

where a^* is like above. We conclude that

$$\liminf v_n(x) \geqslant \beta \liminf (\theta_n)^{2/n}, \qquad x \in S.$$

Thus, if $\theta_n \to 0$ slowly enough, we have:

$$\liminf v_n(x) > 0.$$

(For instance, if $\theta_n n^{\alpha} \to C \neq 0$ for some $\alpha > 0$, we have $\liminf v_n(x) \geqslant \beta$). Now, since $r(\cdot, \cdot, 0) \equiv 0$, we have $v^*(0) \equiv 0$ and therefore, the strict inequality holds in (3.2), **whatever** the seed u is. Observe that in this example, we have, for every $x \in S$:

$$E(x, M, \theta_n 0) = r(x, \theta_n) = \theta_n^{1/x} \to 0 \quad \text{as} \quad n \to \infty, \quad \text{but}$$

$$E(S, M, \theta_n, 0) = 1 \quad \text{for} \quad n = 0, 1, 2, \ldots.$$

The next theorem shows that a strengthened version of (3.1) and an appropriate selection of the seed u, are enough to have pointwise convergence of $\{v_n\}$ to $v^*(\theta)$.

THEOREM 3.2. *Suppose that*

$$E(S, M, \theta_n, \theta) \to 0 \quad \text{as} \quad n \to \infty \tag{3.6}$$

and

$$u \leqslant v^*(\theta). \tag{3.7}$$

Then, for every $x \in S$,

$$v_n(x) \to v^*(\theta)(x) \quad \text{as} \quad n \to \infty.$$

Proof. Let $\varepsilon > 0$ and let n be a positive integer. Let $x \in F_n$ and select $a \in A(x)$ such that

$$v_n(x) \leqslant r(x, a, \theta_n) + \beta \Sigma p(x, y, a, \theta_n)\, v_{n-1}(y) + \varepsilon.$$

Now, by the optimality equation,

$$v^*(x) \geqslant r(x, a, \theta) + \beta \Sigma p(x, y, a, \theta)\, v^*(y),$$

where we write v^* instead of $v^*(\theta)$. Then,

$$v_n(x) - v^*(x) \leqslant \big[r(x, a, \theta_n) - r(x, a, \theta) + \beta \Sigma \big(p(x, y, a, \theta_n) - p(x, y, a, \theta)\big)\, v_{n-1}(y)\big] + \beta \Sigma p(x, y, a, \theta)\big(v_{n-1}(y) - v^*(y)\big) + \varepsilon. \tag{3.8}$$

In the right-hand side of (3.8), the term in brackets is less than or equal to $E(x, M, \theta_n, \theta)$ (because $|v_{n-1}| \leqslant \mathscr{R}_M$). On the other hand,

$$\sum p(x, y, a, \theta)\big(v_{n-1}(y) - v^*(y)\big) = \sum_{y \in F_{n-1}} p(x, y, a, \theta) \times \big(v_{n-1}(y) - v^*(y)\big) + \sum_{y \notin F_{n-1}} p(x, y, a, \theta)\big(u(y) - v^*(y)\big)$$

Using (3.7) we see that the second term in the right-hand side of the above equality is $\leqslant 0$. Then, from (3.8) we obtain

$$v_n(x) - v^*(x) \leqslant E(x, M, \theta_n, \theta) + \beta \sum_{y \in F_{n-1}} p(x, y, a, \theta_n) \times \big(v_{n-1}(y) - v^*(y)\big) + \varepsilon \tag{3.9}$$

Now, for $k = 0, 1, 2, \ldots$, define $d_k^+ : S \to \mathbf{R}$ as follows:

$$d_k^+(x) := v_k(x) - v^*(x) \quad \text{if} \quad v_k(x) \geqslant v^*(x) \quad \text{and} \quad x \in F_k$$
$$d_k^+(x) := 0 \quad \text{if} \quad v_k(x) < v^*(x) \quad \text{or} \quad x \notin F_k.$$

It is clear that

$$\sum_{y \in F_{n-1}} p(x, y, a, \theta)\big(v_{n-1}(y) - v^*(y)\big) \leqslant \sum p(x, y, a, \theta)\, d_{n-1}^+(y) \leqslant H_\theta\, d_{n-1}^+(x).$$

From this and (3.9) we conclude the following.

$$d_n^+(x) \leqslant E(x, M, \theta_n, \theta) + \beta H_\theta\, d_{n-1}^+(x) + \varepsilon$$

and, from the arbitraryness of $\varepsilon > 0$, we obtain

$$d_n^+(x) \leqslant E(x, M, \theta_n, \theta) + \beta H_\theta\, d_{n-1}^+(x).$$

Finally, because x is an arbitrary element of F_n, we get, since $E(F_n, \cdot, \cdot, \cdot) \leqslant E(S, \cdot, \cdot, \cdot)$:

$$d_n^+ \leqslant E(S, M, \theta_n, \theta) + \beta H_\theta\, d_{n-1}^+.$$

Let $\delta > 0$ and select $m \in \mathbf{N}$ such that $n \geqslant m$ implies $E(S, M, \theta_n, \theta) \leqslant \delta$. Then, for $n \geqslant m$ we have

$$d_n^+ \leqslant \delta + \beta H_\theta\, d_{n-1}^+,$$

and an induction argument gives that, for $k = 0, 1, 2, \ldots$

$$d^+_{m+k} \leqslant \delta\,(1-\beta^{k+1})/(1-\beta)+\beta^{k+1} H^{k+1}_\theta\, d^+_{m-1} \tag{3.10}$$

Now, observe that $d^+_{m-1} \leqslant 2\mathscr{R}_M$ and then, for every $x \in S$,

$$\beta^k H^k d^+_{m-1}(x) \to 0 \qquad \text{as} \qquad k \to \infty.$$

This result and (3.10) imply, since $\delta > 0$ is arbitrary, that, for $x \in S$,

$$d^+_n(x) \to 0 \qquad \text{as} \qquad n \to \infty. \tag{3.11}$$

Let $x \in S$ and observe that

$$v_n(x) = v^*(x)+v_n(x)-v^*(x) \leqslant v^*(x)+d^+_n(x), \qquad \text{and then,}$$

$$\limsup v_n(x) \leqslant v^*(x)+\limsup d^+_n(x) = v^*(x), \qquad x \in S. \tag{3.12}$$

where we have used (3.11) in the last equality. Now, (3.12) and Theorem 3.1 show that, for every $x \in S$

$$v_n(x) \to v^*(\theta)(x) \qquad \text{as} \qquad n \to \infty. \qquad \blacksquare$$

REMARK 3.3. In the case when M is a singleton, we have already mentioned that the TNVI scheme becomes White's extended scheme ([1]). In this circumstance, if the seed satisfies $u \leqslant v^*(\theta)$ (and $\mathscr{R}_\theta(x) < \infty$ for $x \in S$), we obtain, from Theorem 3.2, that $\{v_n\}$ converges pointwise to $v^*(\theta)$. This is a slight improvement with respect to Corollary 3.1 in [1] where it was proved that, if the seed u satisfies $u \leqslant T_\theta u$, then, $\{v_n\}$ converges pointwise to $v^*(\theta)$.

Concerning the estimation of $|v_n(x)-v^*(\theta)(x)|$, our main result is Lemma 3.1 below. Before establishing it, we introduce some notation.

DEFINITIONS and COMMENTS 3.2. Let the sequence $\{\theta_n | n = 0, 1, 2, \ldots\}$, $\theta \in \Theta$, the set M and the sequence $\{v_n | n = -1, 0, 1, \ldots\}$ be as in Definition 3.1. For $x \in S$ and $\varepsilon > 0$ (ε can be ∞), we define

$$C(x, \varepsilon) := \Big\{F \subset S \Big| \sum_{y \notin F} p(x, y, a, \theta)\, \mathscr{R}_M(y) \leqslant \varepsilon \qquad \text{for every} \qquad a \in A(x)\Big\}.$$

Observe that, for every $F \in C(x, \varepsilon)$ and $n = 0, 1, 2, \ldots$,

$$\sum_{y \notin F} p(x, y, a, \theta)\, |v_{n-1}(y)-v^*(\theta)(y)| \leqslant 2\varepsilon \qquad \text{for every} \qquad a \in A(x). \tag{3.12}$$

This is so, because both v_{n-1} and $v^*(\theta)$ belong to $\mathscr{L}_M$. Also, we always have $S \in C(x, \varepsilon)$. Now, let $f(\cdot, \varepsilon)$ be a choice function, that is,

$$f(\cdot, \varepsilon): S \to U\,[C(x, \varepsilon) | x \in S]$$

satisfies

$$f(x, \varepsilon) \in C(x, \varepsilon) \qquad \text{for every} \qquad x \in S. \tag{3.13}$$

For $n = 0, 1, 2, \ldots$, $\varepsilon > 0$ and $x \in S$, we define

$$f^0(x,\varepsilon):=\{x\}, \qquad \text{and for} \qquad n\geqslant 1$$
$$f^n(x,\varepsilon):=U\,[f(y,\varepsilon)|y\in f^{n-1}(x,\varepsilon)].$$

We write v^* instead of $v^*(\theta)$ and from now on, d_n stands for $|v_n-v^*(\theta)|$, that is,

$$v^*:=v^*(\theta)$$
$$d_n:=|v_n-v^*(\theta)|\equiv|v_n-v^*|.$$

For each $V:S\to\bar{\mathbf{R}}$, $\|V\|$ stands for the supremum norm of V;

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

and, for $F\subset S$, $V|F:S\to\bar{\mathbf{R}}$ is defined by

$$V|F(x):=\begin{cases}V(x) & x\in F\\ 0 & x\notin F.\end{cases}$$

We write $\|V\|_n$ instead of $\|V|F_n\|$.

LEMMA 3.1. *Let $\varepsilon>0$, $x\in S$, and let k be a positive integer. If*

$$f^s(x,\varepsilon)\subset F_{n-s}, \qquad s=0,1,\ldots,k-1, \qquad \textit{then,}$$

(i) $$d_n(x)\leqslant\sum_{s=0}^{k-1}\beta^s E\left(f^s(x,\varepsilon),M,\theta_{n-s},\theta\right)+\beta^k H^k\left(d_{n-k}|f^k(x,\varepsilon)\right)(x)+$$
$$+2\varepsilon\beta\,(1-\beta^k)/(1-\beta)$$

(ii) $$d_n(x)\leqslant\sum_{s=0}^{k-1}\beta^s E\left(f^s(x,\varepsilon),M,\theta_{n-s},\theta\right)+2\beta^k H^k\left(\mathscr{R}_M|f^k(x,\varepsilon)\right)(x)+$$
$$+2\varepsilon\beta/(1-\beta).$$

Proof. Observing that $d_{n-k}\leqslant 2\mathscr{R}_M$, (ii) follows immediately from (i). We prove (i) by induction.
Let $x\in F_n$. Then,

$$d_n(x)=|v_n(x)-v^*(x)|=|T_{\theta_n}v_{n-1}(x)-T_\theta v^*(x)|\leqslant$$
$$\leqslant|T_{\theta_n}v_{n-1}(x)-T_\theta v_{n-1}(x)|+|T_\theta v_{n-1}(x)-T_\theta v^*(x)|.$$

In the right-hand side of the last inequality, the first term is bounded above by $E(x,M,\theta_n,\theta)$ (Theorem 2.3 (i)) and the second one is less than or equal to $\beta H_\theta d_{n-1}(x)$ (see (2.4)). Then,

$$d_n(x)\leqslant E(x,M,\theta_n,\theta)+\beta H_\theta d_{n-1}(x) \qquad \text{for} \qquad x\in F_n. \tag{3.14}$$

Now, observe that

$$H_\theta d_{n-1}(x)=\sup\sum p(x,y,a,\theta)\,d_{n-1}(y)\leqslant$$
$$\leqslant\sup\sum_{y\in f(x,\varepsilon)}p(x,y,a,\theta)\,d_{n-1}(y)+\sup\sum_{y\notin f(x,\varepsilon)}p(x,y,a,\theta)\,d_{n-1}(y),$$

where sup is taken over $a\in A(x)$.

In the last inequality, the first term in the right-hand side is $H_\theta(d_{n-1}|f(x,\varepsilon))$ and the second one is $\leqslant 2\varepsilon$ (see (3.12) and (3.13)). Then, (3.14) implies, for $x\in F_n$,

$$d_n(x) \leqslant E(x, M, \theta_n, \theta)+\beta H_\theta(d_{n-1}|f(x,\varepsilon))(x)+2\varepsilon\beta.$$

This proves (i) for $k=1$.

Suppose that (i) holds for $k=r$ and that

$$f^s(x,\varepsilon)\subset F_{n-s}, \qquad s=0,1,\ldots,r.$$

Now, take $y\in f^r(x,\varepsilon)\subset F_{n-r}$. Using the case $k=1$ that we have just proved, we get

$$d_{n-r}(y) \leqslant E(y, M, \theta_{n-r}, \theta)+\beta H_\theta(d_{n-r-1}|f(y,\varepsilon))(y)+2\varepsilon\beta \leqslant$$
$$\leqslant E(f^r(x,\varepsilon), M, \theta_{n-r}, \theta)+\beta H_\theta(d_{n-r-1}|f^{r+1}(x,\varepsilon))(y)+2\varepsilon\beta.$$

Observing that $y\in f^r(x,\varepsilon)$ is arbitrary, we conclude that

$$d_{n-r}|f^r(x,\varepsilon) \leqslant E(f^r(x,\varepsilon), M, \theta_{n-r}, \theta)+\beta H_\theta(d_{n-r-1}|f^{r+1}(x,\varepsilon))+2\varepsilon\beta.$$

From this inequality and inequality (i) with $k=r$, the corresponding inequality with $k=r+1$ follows easily. ■

Now, we study some consequences of Lemma 3.1.

THEOREM 3.3. *Suppose that*

(i) $E(x, M, \theta_n, \theta)\to 0$ *as* $n\to\infty$ *for every* $x\in S$.

(ii) *Given* $x\in S$ *and* $\varepsilon>0$, *there exists a finite set* F *such that for every* $a\in A(x)$,

$$\sum_{y\notin F} p(x,y,a,\theta)\,\mathscr{R}_M(y) \leqslant \varepsilon.$$

Then, for every $x\in S$ *and every seed* u,

$$d_n(x)\to 0 \quad \text{as} \quad n\to\infty, \quad \textit{that is,}$$
$$\{v_n\} \textit{ converges pointwise to } v^*.$$

Proof. Assumption (ii) asserts that $C(x,\varepsilon)$ contains finite sets for every $x\in S$ and $\varepsilon>0$. Then, we can assume that $f(x,\varepsilon)$ is always a finite set and then, so is $f^s(x,\varepsilon)$, $s=0,1,2,\ldots$. Let $\delta>0$ and take $\varepsilon=\delta(1-\beta)/6\beta$.

Now, let $x\in S$. Select $k\geqslant 1$ such that $\beta^k H_\theta^k \mathscr{R}_M(x) < \delta/6$ (Theorem 2.2 (ii)). Finally, select m such that

(a) $\bigcup_{s=0}^{k-1} f^s(x,\varepsilon)\subset F_{m-k+1}$ and,

(b) for $n\geqslant m$

$$E(f^s(x,\varepsilon), M, \theta_{n-s}, \theta) < \delta(1-\beta)/3, \qquad s=0,1,\ldots,k-1.$$

The selection of m is possible by assumption (i) and because $f^s(x,\varepsilon)$ is always a finite set. Then, for $n\geqslant m$, Lemma 3.1 (ii) implies that

$$d_n(x) \leqslant \sum_{s=0}^{k-1} \beta^s \delta (1-\beta)/3 + \delta/3 + \delta/3 < \delta. \qquad \blacksquare$$

The following Theorem will play an important role in the study of the adaptive policies constructed in section 4. To establish it, we need a subsequent definition.

DEFINITION 3.3. Let B a non-empty subset of Θ. For each $\theta \in \Theta$ and $r = 0, 1, 2, \ldots$, we define $\varepsilon(r, B, \theta)$ as follows:

$$\varepsilon(r, B, \theta) := \sup_{(x,a)\in K} \sum_{y \notin F_r} p(x, y, a, \theta)\, \mathscr{R}_B(y).$$

THEOREM 3.4. *Let n, k, r be nonnegative integers, $k \geqslant 1$ and let $u \in \mathscr{L}_M$ be an arbitrary seed. Then*

(i) *$x \in F_n$ and $n \geqslant r+k-1$ together imply that,*

(a) $$d_n(x) \leqslant \sum_{s=1}^{k-1} E(F_r, M, \theta_{n-s}, \theta)\, \beta^s + E(x, M, \theta_n, \theta) + 2\beta^k H_\theta^k (\mathscr{R}_M | F_r)(x) + 2\varepsilon(r, M, \theta)\, \beta/(1-\beta).$$

(b) $$\|d_n\|_n \leqslant \sum_{s=0}^{k-1} E(S, M, \theta_{n-s}, \theta)\, \beta^s + 2 \left\| \beta^k H_\theta^k (\mathscr{R}_M | F_r) \right\|_n + 2\varepsilon(r, M, \theta)\, \beta/(1-\beta).$$

(ii) *If*

(a) *$\varepsilon(r, M, \theta) \to 0$ as $r \to \infty$,*

(b) *$E(S, M, \theta_n, \theta) \to 0$ as $n \to \infty$, and*

(c) *$\mathscr{R}_M$ is bounded on every set F_r,*

Then

$$\|d_n\|_n + \|H_\theta\, d_n\| \to 0 \qquad \text{as} \qquad n \to \infty.$$

Proof. (i) Let r be a nonnegative integer and take $\varepsilon = \varepsilon(r, M, \theta)$. Then, it follows that $f(x, \varepsilon) = F_r$ for $x \in S$, determines a choice function (cf. Definitions and Comments 3.2) and, for $s = 1, 2, \ldots$, we always have $f^s(x, \varepsilon) = F_r$. Then, (a) follows from Lemma 3.1 (ii) and (b) follows immediately from (a) by taking sup over $x \in F_n$.

(ii) Let $\delta > 0$. Select r such that

$$2\varepsilon(r, M, \theta)\, \beta/(1-\beta) < \delta/3 \qquad \text{(assumption (a))}$$

Now, select $k \geqslant 1$ such that

$$2\beta^k \|H_\theta^k (\mathscr{R}_M | F_r)\| \leqslant 2\beta^k \|\mathscr{R}_M\|_r < \delta/3 \qquad \text{(assumption (c))}.$$

Finally, let m be a positive integer such that, for $n \geqslant m$

$$\sum_{s=0}^{k-1} E(S, M, \theta_{n-s}, \theta)\, \beta^s < \delta/3 \qquad \text{(assumption (b))}.$$

Then, for $n > \max[m, k+r-1]$, inequality (b) in part (i) implies $\|d_n\|_n < \delta$ and we conclude that

$$\|d_n\|_n \to 0 \qquad \text{as} \qquad n \to \infty \tag{3.15}$$

To finish the proof, observe that

$$\|H_\theta d_n\| \leqslant \|d_n\|_n + 2\varepsilon(n, M, \theta), \tag{3.16}$$

and the result follows from (3.15) and assumption (a). ■

4. Solution to problems (A) and (B)

In this section we solve problems (A) and (B) posed in section 1. The solution we give is a straightforward application of the results on the TNVI scheme introduced in section 3. Suppose that θ is the true parameter value and that the controller wants to find $v^*(\theta)$. The difficulty is that he does not have a priori knowledge about θ. However the TNVI scheme allows him to find $v^*(\theta)(x)$, $x \in S$, as soon as he has a sequence $\{\theta_n\}$ converging to θ and some restrictions are satisfied by the model. To obtain such a sequence, the controller must observe the system while it is in progress and then he can use the registered history to obtain the approximating sequence. Specifically, we suppose that the controller has at his disposal, a strongly consistent sequence of estimators of θ.

ASSUMPTION 4.1. For any $\theta \in \Theta$, any $x \in S$ and any policy D, there exists a sequence $\{\hat{\theta}_n : H_n \to S\}$ of measurable functions, such that

$$\hat{\theta}_n \to \theta \quad P_x^{D,\theta} \text{— almost surely as } n \to \infty.$$

The sequence $\{\hat{\theta}_n\}$ is said to be a sequence of **strongly consistent** (SC) **estimators** of θ. (cf. [3], [7], [9] and [13]).

REMARK 4.1. Throughout the following, $\{\hat{\theta}_n\}$ stands for a sequence of SC estimators.

Now, the controller can decide to employ an arbitrary policy D and, at the stage n, once he has observed i_n (the information vector up to time n), he can evaluate $\theta_n := \hat{\theta}_n(i_n)$ and then obtain v_n in the TNVI scheme. Because of Assumption 4.1, the controller can be sure that the sequence $\{\theta_n\}$ he is obtaining is going toward the true parameter value. Since we are interested in finite-state methods, we suppose from now on that the TNVI scheme has been defined in such a way that

(i) The seed u is identically zero

(ii) F_n is a finite set for $n = 0, 1, 2, \ldots$

It is time to introduce some continuity requirements on the structure of the decision model.

ASSUMPTION 4.2. (i) Θ and A are metric spaces, $A(x)$ is **compact** for every $x \in S$. $|\cdot,\cdot|$ stands for the metric on Θ.

(ii) For each $x \in S$ and each $y \in S$, the maps

$$(a, \theta) \to r(x, a, \theta) \quad \text{and} \quad (a, \theta) \to p(x, y, a, \theta)$$

are continuous on $A(x) \times \Theta$.

As an immediate consequence, we obtain the following result.

LEMMA 4.1. *Suppose that assumption* 4.2 *hold. Then,*

(i) *For each* $\theta \in \Theta$, $x \in S$, $y \in S$

(a) $\sup_{a \in A(x)} |r(x, a, \tau) - r(x, a, \theta)| \to 0$ *as* $\tau \to \theta$

(b) $\sup_{a \in A(x)} |p(x, y, a, \tau) - p(x, y, a, \theta)| \to 0$ *as* $\tau \to \theta$.

(ii) *For each* $x \in S$, $n = 0, 1, 2, \ldots$,

$v_n(\theta_0^n)(x)$ *is a continuous function of* $\theta_0^n := (\theta_0, \ldots, \theta_n) \in \Theta^{n+1}$

(cf. Remark 3.1 and remember that the seed u is identically zero and the sets F_n are **finite**).

Proof. The proof is a straightforward application of the well known **Tube Lemma** ([10], Lemma 5.8) which, for our present purpose, can be stated as follows:

Let X and Y be topological spaces, X compact and, let $g: X \times Y \to \mathbf{R}$ be a continuous function.

Then, for every $y \in Y$ and $\varepsilon > 0$, there exists a neighborhood V of y, such that

$$\sup_{x \in X} |g(x, w) - g(x, y)| \leqslant \varepsilon \quad \text{for} \quad w \in V. \tag{4.1}$$

Thus, using (2.3) and (4.1) we obtain

$$|\sup_{x \in X} g(x, w) - \sup_{x \in X} g(x, y)| \leqslant \varepsilon \quad \text{for} \quad w \in V,$$

that is,

$$\sup_{x \in X} g(x, \cdot) \quad \text{is continuous} \tag{4.2}$$

(i) Taking $X = A(x)$, $Y = \Theta$ in the Tube Lemma, we obtain (a) and (b) using (4.1) with $g(\cdot,\cdot) = r(x, \cdot, \cdot)$ and $g(\cdot,\cdot) = p(x, y, \cdot, \cdot)$ respectively.

(ii) The proof is by induction.

Using (4.2) with $g(\cdot,\cdot) = r(x, \cdot, \cdot)$ we obtain that, for $x \in F_0$,

$$v_0(\theta_0)(x) = \sup_{a \in A(x)} r(x, a, \theta_0)$$

is a continuous function of $\theta_0 \in \Theta$. Now, suppose that, for each $x \in F_n$, $v_n(\theta_0^n)(x)$ is a continuous function of $\theta_0^n \in \Theta^{n+1}$ and let $x \in F_{n+1}$. Then,

$$g(x, a, \theta_0, ..., \theta_{n+1}) := r(x, a, \theta_{n+1}) + \beta \Sigma p(x, y, a, \theta_{n+1})\, v_n(\theta_0, ..., \theta_n)(y)$$

is a continuous function on $A(x) \times \Theta^{n+2}$ (by the induction hypothesis, Assumption 4.2 and by the finiteness of F_n). Then, from (4.2) we have that

$$v_{n+1}(\theta_0^{n+1})(x) := \sup_{a \in A(x)} g(x, a, \theta_0^{n+1})$$

is a continuous function on θ^{n+2}. The result follows, since $v_n(\theta_0^n)(x) = 0$ for $x \notin F_n$ and $n = 0, 1, 2, ...$ ■

In order to apply the estimations obtained in section 3, we need Assumption 4.3 below. For $t \in \Theta$ and $\delta > 0$, let $B(\delta, t)$ be defined by

$$B(\delta, t) := \{\theta \in \Theta \mid \quad |\theta, t| \leqslant \delta\} \qquad \text{and}$$

define $\Delta(\delta, t)$ and $\varepsilon(\delta, t)$ as follows:

$$\Delta(\delta, t) := \sup_{(x,a) \in K, \theta \in B(\delta, t)} |r(x, a, \theta) - r(x, a, t)| \tag{4.3}$$

$$\varepsilon(\delta, t) := \sup_{(x,a) \in K, \theta \in B(\delta, t)} \Sigma p(x, y, a, \theta)\, R_t(y) \tag{4.4}$$

Assumption 4.3 below. For $t \in \Theta$ and $\delta > 0$, let $B(\delta, t)$ be defined by

(i) $\Delta(\delta, \tau) < \infty$,

(ii) $\varepsilon(\delta, t) < \infty$,

(iii) $R_\tau(x)$ is finite for every $x \in S$; see Remark 4.2 below.

REMARK 4.2. From now on, τ denotes an element of Θ such that Assumption 4.3 is satisfied.

The main consequence of Assumption 4.3 that we are going to use is the following result.

LEMMA 4.2. *Let τ be as in Assumption* 4.3. *Then, for every $\delta > 0$, $\mathscr{R}_{B(\delta,\tau)}(x)$ is finite for every $x \in S$. More precisely, for every $\delta > 0$,*

(i) $\mathscr{R}_{B(\delta,\tau)} \leqslant R_\tau + \Delta(\delta, \tau)/(1-\beta) + \beta\varepsilon(\delta, \tau)/(1-\beta)$.

(ii) *There exists a finite positive number $c(\delta)$ such that $\mathscr{R}_{B(\delta,\tau)} \leqslant c(\delta)(R_\tau + 1)$.*

Proof. It is clear that (ii) follows immediately from (i). To prove (i) observe that:

$$\|H_B R_\tau\| = \varepsilon(\delta, \tau) \qquad \text{where} \qquad B := \mathrm{B}(\delta, \tau).$$

Then, it follows easily that, for $n = 1, 2, ...,$

$$\|H_B^n R_\tau\| \leqslant \varepsilon(\delta, \tau),$$

and then, we obtain

$$\sum_{n=0}^{\infty} \beta^n H_B^n R_\tau \leqslant R_\tau + \beta\varepsilon(\delta, \tau)/(1-\beta). \tag{4.5}$$

Finally, observe that $R_B \leqslant R_\tau + \Delta(\delta, \tau)$. Using this inequality, the monotonicity property of H_B and (4.5), we obtain (i).

COROLLARY 4.1. *Let* $\{\theta_n | n = 0, 1, 2, ...\}$ *be a convergent sequence in* Θ,

$$\theta := \lim_{n \to \infty} \theta_n,$$

and take

$$M := \{\theta, \theta_0, \theta_1, ...\}$$

Then, under Assumption 4.3, $\mathscr{R}_M(x) < \infty$ *for every* $x \in S$. *Moreover,*

$$\mathscr{R}_M \leqslant c(\delta)(R_\tau + 1),$$

where $\delta > 0$ *is selected in such a way that* $M \subset B(\delta, \tau)$ *and* $c(\delta)$ *is the number appearing in Lemma* 4.2 (ii).

Proof. The result follows immediately from Lemma 4.3 observing that $\mathscr{R}_M \leqslant \mathscr{R}_B$ whenever $M \subset B$.

Now, Theorems 4.1 and 4.2 below, represent a solution to Problem (*A*).

THEOREM 4.1. *Suppose that*

(i) *Assumptions* 4.1–4.3 *hold*;

(ii) *Given* $\varepsilon > 0$ *and* $\delta > 0$, $\theta \in \Theta$ *and* $y \in S$, *there exists a nonnegative integer* r *such that*

$$\sup_{\substack{a \in A(y) \\ t \in B(\delta, \theta)}} \sum_{z \notin F_r} p(y, z, a, t)\,(R_\tau(z) + 1) \leqslant \varepsilon \tag{4.6}$$

Then, for every $\theta \in \Theta$, $y \in S$ *and every policy* D,

$$v_n(\hat{\theta}_0^n)(y) \to v^*(\theta)(y) \qquad P_x^{D,\theta} \text{ — almost surely,}$$

where $\hat{\theta}_0^n := (\hat{\theta}_0, ..., \hat{\theta}_n)$.

Proof. Observe that Lemma 4.1 (ii) and the measurability of the $\hat{\theta}_n'$s, imply that $v_n(\hat{\theta}_0^n)(y)$ is measurable for every $y \in S$ and $n = 0, 1, 2, ...$ Now, let $\{\theta_n | n = 0, 1, 2, ...\}$ be an arbitrary sequence in Θ converging to $\theta \in \Theta$ and take $M := \{\theta, \theta_0, \theta_1, ...\}$. We are going to show that the conditions in Theorem 3.1 are satisfied.

Let $y \in S$, $\varepsilon > 0$ and $\delta > 0$. Using assumption (ii), we see that there exists a nonnegative integer r such that

$$\sup \sum_{z \notin F_r} |p(y, z, a, t) - p(y, z, a, \theta)|\,(R_\tau(z) + 1) \leqslant 2\varepsilon, \tag{4.7}$$

where sup is taken over $a \in A(y)$ and $t \in B(\delta, \theta)$.

Now, Lemma 4.1 (i) implies, since F_r is a finite set, that we can find $\delta_1 \leqslant \delta$ such that

$$\sup \Big[|r(y, a, t) - r(y, a, \theta)| + \beta \sum_{z \in F_r} |p(y, z, a, t) - p(y, z, a, \theta)|\,(R_\tau(z) + 1)\Big] \leqslant \varepsilon, \tag{4.8}$$

where sup is taken over $a \in A(y)$ and $t \in B(\delta_1, \theta)$.

From (4.7), (4.8) and Corollary 4.1, we obtain that

$$E(y, M, \theta_n, \theta) \to 0 \qquad \text{as} \qquad n \to \infty. \tag{4.9}$$

Finally, it is clear that assumption (ii) implies that condition (ii) in Theorem 3.3 holds and (4.9) is precisely condition (i). Thus, we conclude that, for every $y \in S$,

$$v(\theta_0^n)(y) \to v^*(\theta)(y) \qquad \text{as} \qquad n \to \infty,$$

Now, the result follows since $\hat{\theta}_n \to \theta$ $P_x^{D,\theta}$ — almost surely. ■

Under additional assumptions a result stronger than that of Theorem 4.1 can be obtained as follows.

THEOREM 4.2. *Suppose that Assumptions* 4.1–4.3 *hold, and that for each* $\theta \in \Theta$ *the following is satisfied*:

(i) $\Delta(\delta, \theta) \to 0$ *as* $\delta \to 0$ (cf. (4.3)).

(ii) $\sup_{\substack{(x,a)\in K \\ t\in B(\delta,\theta)}} \sum_{y\in S} |p(x, y, a, t) - p(x, y, a, \theta)| (R_\tau(y)+1) \to 0$ *as* $\delta \to 0$.

(iii) $\sup_{(x,a)\in K} \sum_{y\notin F_r} p(x, y, a, \theta)(R_\tau(y)+1) \to 0$ *as* $r \to \infty$.

Then, for every $\theta \in \Theta$, $x \in S$ *and every policy* D,

$$\|v_n(\hat{\theta}_0^n) - v^*(\theta)\|_n + \left\| H_\theta |v_n(\hat{\theta}_0^n) - v^*(\theta)| \right\| \to 0 \quad P_x^{D,\theta} \text{ — almost surely.}$$

Proof. Let $\{\theta_n | n = 0, 1, 2, \ldots\}$ be a convergent sequence in Θ. We are going to show that the conditions in Theorem 3.4 (ii) are satisfied. Let $\theta = \lim \theta_n$ and $M := \{\theta, \theta_0, \theta_1, \ldots\}$. Then, assumptions (i) and (ii), together with Corollary 4.1 imply that

$$E(S, M, \theta_n, \theta) \to 0 \qquad \text{as} \qquad n \to \infty. \tag{4.10}$$

Now, Corollary 4.1 and assumption (iii) imply that

$$\varepsilon(r, M, \theta) \to 0 \qquad \text{as} \qquad r \to \infty, \tag{4.11}$$

and since F_r is finite for $r = 0, 1, 2, \ldots$, we have that

$$\mathscr{R}_M \text{ is bounded on the sets } F_r. \tag{4.12}$$

From (4.10)–(4.12) and Theorem 3.4 (ii) we conclude that, for each $\theta \in \Theta$,

$$\|v_n(\theta_0^n) - v^*(\theta)\|_n + \left\| H_\theta |v_n(\theta_0^n) - v^*(\theta)| \right\| \to 0 \qquad \text{as} \qquad n \to \infty. \tag{4.13}$$

The result follows observing that $\hat{\theta}_n \to \theta$ $P_x^{D,\theta}$ — almost surely. ■

In Theorems (4.1) and 4.2, the policy D is an arbitrary policy and can make no use of the registered history of the process. At the n-th stage, the observed information vector i_n is used to obtain the estimation $\hat{\theta}_n(i_n)$ and, as n increases, the controller is gaining knowledge about the true parameter value (because $\{\theta_n\}$ is a sequence of SC estimators of θ) and, it is desirable to use this knowledge to choose actions that, in some sense, are "nearly optimal", at least as $n \to \infty$.

DEFINITION 4.1. (TNVI adaptive policies; cf. [6]). Suppose that Assumptions 4.1 and 4.2 hold and let $\{\theta_0, \theta_1, ...\}$ be any sequence in Θ. Let d be an arbitrary stationary policy. For each $n = 0, 1, 2, ...$, define a function $f_n(\theta_0^n, \cdot): S \to A$ as follows:

$$f_n(\theta_0^n, x) := \arg\max_{a \in A(x)} [r(x, a, \theta_n) + \beta \Sigma p(x, y, a, \theta_n) v_{n-1}(\theta_0^{n-1})(y)] \text{ if } x \in F_n,$$

$$f_n(\theta_0^n, x) := d(x) \text{ if } x \notin F_n,$$

where $v_k(\theta_0^k)$ stands for the k-th function produced by the TNVI scheme in Definition 3.1, with $v_{-1} \equiv 0$. Now, define a (deterministic) policy $\hat{D} = \{\hat{D}_n | n = 0, 1, 2, ...\}$ as follows:

$$\hat{D}_n(I_n) := f_n(\hat{\theta}_0^n(I_n), X_n), \qquad n = 0, 1, 2, ...$$

for $I_n = (X_0, A_0, ..., X_{n-1}, A_{n-1}, X_n) \in H_n$. The policy $\hat{D}$ thus constructed is called a TNVI adaptive policy.

REMARK 4.3. We suppose that $f_n(\theta_0^n, x)$ can be selected in such a way that $f_n(\cdot, x)$ becomes a measurable mapping from Θ^{n+1} in $A(x)$ for each $x \in S$. This is possible if Assumption 4.2 holds (see, for instance, [12], Theorem 12.1). Observe that Theorems 4.1 and 4.2 hold when D is substituted by $\hat{D}$. Finally, note that $\hat{D}$ depends on what stationary policy d is employed to define $\hat{D}(I_n)$ if $X_n \notin F_n$. However, we do not indicate explicitly this dependence.

To study the asymptotic optimality properties of $\hat{D}$, we introduce the following definition.

DEFINITION 4.2. Under Assumption 4.3, define $\varphi: K \times \Theta \to R$ by

$$\varphi(x, a, \theta) := r(x, a, \theta) + \beta \Sigma p(x, y, a, \theta) v^*(\theta)(x).$$

REMARK 4.4. $\varphi(x, a, \theta)$ has been used as a measure of "goodness" of taking action a when the present state is x and θ is the true parameter value; see, for instance [1], [6] or [13]. As a consequence of the optimality equation, we have that $\varphi \leqslant 0$. The relation of φ to asymptotic optimality is given by the following relation, whose proof can be found, for instance in [6] or [13].

For every policy D, $\theta \in \Theta$ and $x \in S$,

$$\sum_{n=N}^{\infty} \beta^{n-N} E_x^{D,\theta} \varphi(X_n, A_n, \theta) = v_N(x, \theta) - E_x^{D,\theta} v^*(\theta)(X_N) \tag{4.14}$$

(cf. Definition 1.1 (ii)).

Using (4.14) and the fact that $\varphi \leqslant 0$, we obtain the following result.

LEMMA 4.4. *A policy D is* ADOS (*when θ is the true parameter value*), *if and only if, for every $x \in S$,*

$$E_x^{D,\theta}\,\varphi\,(X_n, A_n, \theta) \to 0 \qquad as \qquad n \to \infty.$$

The next Lemma will play an important role in the answer to problem (B).

LEMMA 4.5. *Let* $\{\theta_n | n = 0, 1, 2, ...\}$ *be a sequence in* Θ *converging to* θ *and let* $M := \{\theta, \theta_0, \theta_1, ...\}$. *Then*

(i) $\varphi\,(x, a, \theta) \leqslant 2\mathscr{R}_\theta\,(x)$ *for every* $(x, a) \in K$

(ii) $\varphi\,(x, f_n\,(\theta_0^n, x), \theta) \leqslant E\,(x, M, \theta_n, \theta) + H_\theta\,d_{n-1}\,(x) + d_n\,(x),\ x \in F_n$.

(iii) $\left\|\varphi\,(\cdot, f_n\,(\theta_0^n, \cdot), \theta)\right\|_n \leqslant E\,(S, M, \theta_n, \theta) + \|H_\theta\,d_{n-1}\| + \|d_n\|_n$

Proof. (i) $\varphi\,(x, a, \theta) \leqslant |r\,(x, a, \theta)| + \beta\Sigma p\,(x, y, a, \theta)|\,v^*(\theta)\,(y)| + |v^*(\theta)\,(x)|$

$$\leqslant R_\theta\,(x) + \beta H_\theta\,\mathscr{R}_\theta\,(x) + \mathscr{R}_\theta\,(x) \leqslant$$

$$\leqslant R_\theta\,(x) + \mathscr{R}_\theta\,(x) - R_\theta\,(x) + \mathscr{R}_\theta\,(x) = 2\mathscr{R}_\theta\,(x).$$

(ii) Writing a instead of $f_n\,(\theta_0^n, x)$ we have that

$$\varphi\,(x, a, \theta) = \left[r\,(x, a, \theta) - r\,(x, a, \theta_n) + \beta\Sigma\,(p\,(x, y, a, \theta) -\right.$$

$$\left. - p\,(x, y, a, \theta_n)\,v_{n-1}\,(y)\right] + [r\,(x, a, \theta_n) +$$

$$+ \beta\Sigma p\,(x, y, a, \theta_n)\,v_{n-1}\,(y) - v^*(x)] + \left[\Sigma p\,(x, y, a, \theta)\,(v^*(y) - v_{n-1}\,(y))\right]$$

where v_k stand for the k-th function produced by the TNVI scheme with $v_{-1} \equiv 0$ and v^* stands for $v^*(\theta)$. In the right hand side of the above equality, the first and third terms in brackets are bounded in absolute value by $E\,(x, M, \theta_n, \theta)$ and $H_\theta\,|v^* - v_{n-1}|\,(x)$ respectively, while, for $x \in F_n$, the second one is $d_n\,(x)$. This proves (ii) and (iii) follows from taking supremum over $x \in F_n$.

Now, Theorems 4.3 and 4.4 below refer to the asymptotic optimality properties of the TNVI adaptive policy $\hat{D}$ and represent our solution to problem (B).

THEOREM 4.3. *Suppose that the conditions in Theorem* 4.2 *hold. Then, for every* $\theta \in \Theta$ *and every* $x \in S$,

(i) $\varphi\,(X_n, \hat{D}_n\,(I_n), \theta)\,I_{F_n}\,(X_n) \to 0$ *as* $n \to \infty$ $P_x^{\hat{D},\theta}$ *— almost surely, where* I_B *stands for the indicator function of the set* B.

(ii) $\varphi\,(X_n, \hat{D}_n\,(I_n), \theta) \to 0$ *in* $P_x^{\hat{D},\theta}$ *— measure.*

Proof. (i) We have shown that, under the conditions in Theorem 4.2, (4.10) and (4.13) hold. Then, from Lemma 4.5 (iii) we obtain that, for every sequence $\{\theta_n | n = 0, 1, 2, ...\}$ converging to $\theta \in \Theta$,

$$\left\|\varphi\,(\cdot, f_n\,(\theta_0^n, \cdot), \theta)\right\|_n \to 0 \qquad \text{as} \qquad n \to \infty.$$

This gives the result, since $\hat{\theta}_n \to \theta$ $P_x^{D,\theta}$ — almost surely.

(ii) Let D be any policy. Then, for $n \geqslant 1$, and $r \geqslant 0$

$$E_x^{D,\theta}\left(\mathscr{R}_\theta(X_n)\, I_{S-F_r}(X_n) | X_{n-1}=y,\, D_{n-1}=a\right)= \\ =\sum_{z\notin F_r} p(y,z,a,\theta)\,\mathscr{R}_\theta(z)\leqslant \varepsilon_1(r,\theta)$$

where,

$$\varepsilon_1(n,\theta):=\sup_{(y,a)\in K}\sum_{z\notin F_n} p(y,z,a,\theta)\,\mathscr{R}_\theta(z). \tag{4.15}$$

Thus,

$$E_x^{D,\theta}\left(\mathscr{R}_\theta(X_n)\, I_{S-F_r}(X_n)\right)\leqslant \varepsilon_1(r,\theta). \tag{4.16}$$

and then, from Lemma 4.5 (i), we conclude that

$$E_x^{D,\theta}\left|\varphi\left(X_n, D_n(I_n),\theta\right) I_{S-F_n}(X_n)\right|\leqslant 2\varepsilon_1(n,\theta), \qquad n\geqslant 1. \tag{4.17}$$

In particular, (4.17) holds with $D=\hat{D}$ and from assumption (ii) in Theorem 4.2 we obtain that $\varepsilon_1(n,\theta)\to 0$ as $n\to\infty$. Finally, since L^1 convergence is stronger than convergence in measure, we conclude that

$$\varphi\left(X_n, \hat{D}_n(I_n),\theta\right) I_{S-F_n}(X_n)\to 0 \qquad \text{in} \qquad P_x^{\hat{D},\theta}\text{ — measure.}$$

This fact, together with (i) proves (ii).

REMARK 4.5. We observe that the proof of (4.16) has general character, that is, depends **only** on the definition of $\varepsilon_1(n,\theta)$.

THEOREM 4.4. *Suppose that*

(i) *Conditions on Theorem* 4.1 *hold.*

(ii) $\sup_{(x,a)\in K}\sum_{y\notin F_n} p(x,y,a,\theta)\left(R_\tau(y)+1\right)\to 0$ *as* $n\to\infty$ *for every* $\theta\in\Theta$.

Then, for every $\theta\in\Theta$, $x\in S$,

$$E_x^{\hat{D},\theta}\left(\varphi\left(X_n,\hat{D}_n(I_n),\theta\right)\right)\to 0 \qquad as \qquad n\to\infty,$$

that is, the TNVI *adaptive policy* $\hat{D}$ *is* ADOS.

Proof. Let $\theta\in\Theta$. Observe that, from Lemma 4.2, there exists a constant c such that $\mathscr{R}_\theta\leqslant c(R_\tau+1)$. Then, assumption (ii) is equivalent to the following:

$$\varepsilon_1(n,\theta)\to 0 \qquad \text{as} \qquad n\to\infty. \tag{4.18}$$

Let $\{\theta_n | n=0,1,2,\ldots\}$ be a sequence in Θ converging to θ and take $M:=\{\theta,\theta_0,\theta_1,\ldots\}$. Then, we note that under conditions of Theorem 4.1 we have, for every $x\in S$,

$$E(x,M,\theta_n,\theta)+d_n(x)\to 0 \qquad \text{as} \qquad n\to\infty. \tag{4.19}$$

Now, let c_1 be a constant such that,

$$\mathscr{R}_M\leqslant c_1(R_\tau+1)$$

(Corollary 4.1). Then, it follows that, for $k=0,1,2,\ldots$ and $x\in S$,

$$H_\theta d_n(x) \leqslant \|d_n\|_k + 2c_1 \sup_{a \in A(x)} \sum_{x \notin F_k} p(x, y, a, \theta)\,(R_\tau(y)+1).$$

Now, taking lim sup as $n \to \infty$, we obtain, using (4.19) and the finiteness of the sets F_k, that

$$\limsup H_\theta d_n(x) \leqslant 2c_1 \sup_{a \in A(x)} \sum_{x \notin F_k} p(x, y, a, \theta)\,(R_\tau(y)+1).$$

and, letting k go to ∞ and using assumption (ii) in Theorem 4.1 we obtain that, for every $x \in S$,

$$H_\theta d_n(x) \to 0 \qquad \text{as} \qquad n \to \infty. \tag{4.20}$$

From (4.19), (4.20) and Lemma 4.5 (ii) we obtain, using the finiteness of the sets F_r, that

$$\left\|\varphi\left(\cdot, f_n(\theta_0^n, \cdot), \theta\right)\right\|_r \to 0 \quad \text{as} \quad n \to \infty \quad \text{for} \quad r = 0, 1, 2, \ldots \tag{4.21}$$

Now, observe that (4.21) and Lemma 4.5 (i) allow us to conclude, using the bounded convergence theorem, that

$$E_x^{\hat{D},\theta}\left|\varphi\left(X_n, \hat{D}_n(I_n), \theta\right) I_{F_r}(X_n)\right| \to 0 \quad \text{as } n \to \infty,\ r = 0, 1, 2, \ldots \tag{4.22}$$

and then, for $r = 0, 1, 2, \ldots$

$$\limsup E_x^{\hat{D},\theta}\left|\varphi\left(X_n, \hat{D}_n(I_n), \theta\right)\right| = \limsup E_x^{\hat{D},\theta}\left|\varphi\left(X_n, \hat{D}_n(I_n), \theta\right) I_{S-F_r}(X_n)\right|.$$

Finally, using (4.16) and Lemma 4.5 (i), we obtain that the right-hand side of the above equality is less than or equal to $2\varepsilon_1(r, \theta)$ and using (4.18) we obtain, since r is arbitrary, that

$$E_x^{\hat{D},\theta}\left|\varphi\left(X_n, D_n(I_n), \theta\right)\right| \to 0 \qquad \text{as} \qquad n \to \infty. \qquad \blacksquare$$

5. Concluding remarks

We have seen in Example 3.1 that the continuity requirement (3.1) can be too weak to ensure that the sequence $\{v_n\}$ produced by the TNVI scheme converges to $v^*(\theta)$. On the other hand, the stronger condition (3.6) and an appropriate selection of the seed are enough to ensure that $\{v_n\}$ converges pointwise to $v^*(\theta)$. However, an important problem is to estimate $|v_n(x) - v^*(\theta)(x)|$ for $x \in S$, and under the conditions of Theorem 3.2, such an estimation is possible if we have a priori knowledge about ε-optimal policies when θ is the true parameter value (i.e. policies D satisfying $\|v(D, \theta) - v^*(\theta)\| \leqslant \varepsilon$); since it is unrealistic to assume such a priori knowledge, we had to look for a different approach and, in certain way, this justifies the conditions on the tails of the transition probabilities imposed in sections 3 and 4. Now, let us analyze briefly the basic Assumptions 4.1–4.3 which were used to solve problems (A) and (B). It is clear that the application

of the TNVI scheme to solve these problems, require some estimation method like that in Assumption 4.1 and an assumption of this type seems to be unvoidable. On the other hand, the continuity requirements in Assumption 4.2 seem natural and are satisfied in most practical cases. Finally, among the conditions in the (boundedness) Assumption 4.3, the most restrictive one is condition (ii). However, it can be weakened and our solution to problems (A) and (B) still holds. In fact, the only consequence of Assumption 4.3 that was used in section 4 was Lemma 4.2 (ii) and it holds under Assumption 4.3′ below.

ASSUMPTION 4.3′. There exists $\tau \in \Theta$ such that, for every $\delta > 0$ the following is satisfied:

(i) $\Delta(\delta, \tau) < \infty$

(ii) There exist constants $\alpha(\delta) \geqslant 0$ and $c_1(\delta) \geqslant 0$ such that, for every $x \in S$,

$$\sup_{a \in A(x), \theta \in B(\delta, \tau)} \Sigma p(x, y, a, \theta) R_\tau(y) \leqslant \alpha(\delta) R_\tau(x) + c_1(\delta),$$

and $\alpha(\delta)\beta < 1$.

(iii) $R_\tau(x)$ is finite for every $x \in S$.

So, Assumption 4.3′ can take the place of Assumption 4.3 and our solution to problems (A) and (B) still holds.

On the other hand, we note two important features of the TNVI adaptive policy $\hat{D}$:

(i) For each $n = 0, 1, 2, \ldots$ and each $i_n \in H_n$, the determination of $\hat{D}_n(i_n)$ depends only on a finite number of states.

(ii) The application of the TNVI adaptive policy $\hat{D}$ does not require a priori knowledge of a stationary optimal policy for every parameter value; such a priori knowledge is needed for the application of the "principle of estimation and control" of Schäl ([13]).

Finally, we mention that active research on the application of the ideas in [6] as well as those in the present paper to problems of priority assignment in queueing systems is presently in progress ([2]).

Acknowledgement. The author is grateful to Professor O. Hernández-Lerma for having suggested this problem and for his permanent encouragement, aside of many corrections and enlightening discussions.

References

[1] CAVAZOS-CADENA R. Finite-state approximations for denumerable state discounted Markov decision processes. Submitted to *Journal of Applied Mathematics and Optimization*, 1984.

[2] CAVAZOS-CADENA R., HERNÁNDEZ-LERMA O. Estimation and control of priority assignment in discrete time queues-discounted cost criterion, 1984. In preparation.

[3] DOSHI B. T., SHREVE S. Strong consistency of a modified maximum likelihood estimator for controlled Markov chains. *Journal of Applied Probability*, **17** (1980), 726–732.
[4] FEDERGRUEN A., SCHWEITZER P. J. Non-stationary Markov decision problems with converging parameters. *Journal of Optimization Theory and Applications*, **34** (1981), 207–241.
[5] HARRISON J. Discrete dynamic programming with unbounded rewards. *Ann. Math. Statist.* **43** (1972), 636–644.
[6] HERNÁNDEZ-LERMA O., MARCUS S. I. Adaptive control of discounted Markov decision chains. *J. Optim. Theory Appl.* 1984. To appear.
[7] KOLONKO M. Strongly consistent estimation in a controlled Markov renewal model. *Journal of Applied Probability*. **19** (1982), 532–545.
[8] LIPPMAN S. A. On dynamic programming with unbounded rewards. *Management Sci.* **21** (1975), 1225–1233.
[9] MANDL P. Estimation and control in Markov chains. *Advances in Applied Probability*, **6** (1974), 40–60.
[10] MUNKRES J. R. Topology, a First Course. Prentice-Hall, 1975.
[11] ROSS S. M. Applied probability models with optimization applications. San Francisco, Hoden-Day, 1976.
[12] SCHÄL M. Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. *Z. Wahrsch. Verw. Gebiete*, **32** (1975), 179–196.
[13] SCHÄL M. Estimation and control in discounted stochastic dynamic programming. Preprint No. 428, University of Bonn, Institute of Applied Mathematics, 1981.
[14] WESSELS J. Markov programming by succesive approximations with respect to weighted supremum norms. *J. Math. Anal. Appl.* **58** (1979), 326–335.

Received, September 1985.

Aproksymacja i sterowanie adaptacyjne procesami decyzyjnymi Markowa z dyskontem, nieograniczonymi nagrodami i przeliczalną liczbą stanów

W pracy rozważa się procesy decyzyjne Markowa z dyskontem, nieograniczonymi nagrodami i przeliczalną liczbą stanów, które zależą od nieznanych parametrów. Podobnie jak Hernández-Lerma i Marcus [6] stosujemy schemat iteracyjny wartości niestacjonarnych (Fredergruen, Schweitzer [4]) z procedurą dla procesów ze skończoną liczbą stanów (wprowadzoną w [1]). Pozwala nam to, wraz z metodą uzyskiwania estymatorów zgodnych, na znalezienie optymalnych globalnych zdyskontowanych nagród odpowiadających prawdziwym wartościom parametrów. Zaproponowano również asymptotycznie optymalną strategię adaptacyjną.

Аппроксимация и адаптивное управление марковскими процессами принятия решений с дисконтированием, неограниченным премированием с перечислимым числом состояний

В работе рассматриваются марковские процессы принятия решений с дисконтированием, неограниченными премиями и перечислимым числом состояний, которые зависят от неизвестных параметров. Также как Хернандес-Лерма и Маркус [6] используем

итерационную схему нестационарных значений (Федергруэн, Швайцер [4]) с процедурой для процессов с конечным числом состояний (введенной в [1]). Это, вместе с методом получения согласованных оценок, позволяет находить оптимальные глобальные дисконтированные премии, соответствующие действительным значениям параметров. Предложена также асимптотически оптимальная адаптивная стратегия.