Control
and Cybernetics
VOL. 15 (1986) No. 1

Approximation and adaptive policies in discounted dynamic programming*)

by

ONÉSIMO HERNÁNDEZ-LERMA

Departamento de Matemáticas
Centro de Investigación del I.P.N.
Apartado Postal 14-740
07000 México, D.F., Mexico

An iterative procedure meant to approximate the optimal reward function of infinite-horizon discounted dynamic programming problems with Polish state and action spaces is considered. The procedure is then used to determine an asymptotically optimal policy and it is also combined with a consistent parameter estimation scheme to determine an asymptotically optimal policy for decision models depending on unknown parameters. The latter policy is compared with the "principle of estimation and control" recently introduced by Schäl (1981), which is extended here to Polish state-space decision problems.

KEYWORDS: Markov decision process, Polish state space, Discounted reward criterion, Nonstationary value-iteration, Asymptotically optimal policies, Adaptive policies.

1. Introduction

In this paper we consider an iterative procedure meant to approximate the optimal reward function of infinite-horizon discounted dynamic programming problems with Polish (i.e., complete separable metric) state and action spaces. The procedure is then used to: (i) determine an asymptotically optimal policy, and (ii) it is combined with a strongly consistent parameter estimation scheme to determine an asymptitically optimal (adaptive) policy for decision models depending on unknown parameters. The policy obtained in (ii) is compared with the "principle of estimation and control" introduced by Schäl [22] for the adaptive control of de-

*) This research was supported in part by the Consejo Nacional de Ciencia y Tecnologia, under the grant PCCBBNA 020630.

numerable-state semi-Markov processes and extended here to Polish state-space Markov decision problems.

Our motivation in considering the problems indicated in the previous paragraph was born from our interest in Markov decision processes with incomplete state information (MDP-ISI) *and* depending on unknown parameters. We are thus confronted with a decision problem combined with state identification (sometimes called a filtering problem) and parameter estimation. However, it is well-known [11, 18, 20, 21, 23] that in many cases of interest a MDP-ISI can be reduced to a Markov decision process (MDP) in the usual sense, but in which the state space, say, S, of the original problem is replaced by the space S' of probability measures on S. Therefore, since S' turns out to be a Polish space in most of the usual cases (cf. cited references), it seemed natural to begin by extending to the case of a Polish state-space previously known results for MDP's with unknown parameters and denumerable (possibly finite) state space. And this is essentially what we do in the present paper: the nonstationary value-iteration (NVI) scheme introduced by Federgruen and Schweitzer [2] for MDP's with *finite* state and action spaces, as well as the adaptive policies considered by Schäl [22] and Hernàndez-Lerma and Marcus [5, 9] are extended here to the case of Polish state and action spaces. This is a first step towards the solution of the MDP-ISI and unknown parameters; the main difficulty involved in obtaining a complete solution is briefly discussed in Section 6.

Our results are also related to approximations of dynamic programs obtained under quite general conditions by Langen [15] and Whitt [24]. However, by restricting ourselves to *discounted* dynamic programming models we are able to show (uniform) convergence of our approximation schemes with very simple and short proofs.

We begin in Section 2 by introducing the decision models we are concerned with. In section 3, the NVI scheme of Federgruen and Schweitzer [2] is extended to decision models with Polish state and action spaces. The NVI scheme is used in Section 4 to determine an asymptiotically optimal (AO) policy for adaptive decision models, i.e., decision models depending on unknown parameters. Also in section 5, our results are briefly compared with the "principle of estimation and control" [22]. extended here to MDP's with Polish state space.

2. The decision model

To avoid unnecessary repetitions we shall agree that a topological (respectively, product) space is always endowed with the Borel (respec., product) σ-algebra. The Cartesian product of the sets A and B is denoted by AB.

As usual [3, 10, 11, 16, ...] to state the (discounted) dynamic programming problem we need to specify a decision model, the collection of admissible policies, and the objective function. This is done as follows.

The *decision model* (S, A, q, r, β) satisfies:

(A1) (a) The state space S is a Polish (=complete separable metric) space.

(b) The action set A is Polish. For each $x \in S$, the set of admissible actions in state x, denoted by $A(x)$, is a nonempty measurable subset of A. Let $K := \{(x, a) : x \in S,\ a \in A(x)\}$ be a measurable subset of (the product space) SA.

(c) $q(x, a, \cdot)$, for $(x, a) \in K$, is the transition law: when the system is in state x and action $a \in A(x)$ is chosen, the system moves to a new state according to the probability distribution $q(x, a, \cdot)$ on S.

(d) $r: K \to \boldsymbol{R}$ is the (measurable) reward function.

(e) $0 \leqslant \beta < 1$ is the discount factor.

In addition, we shall assume the following.

(A2) (a) There exists a constant R such that $|r(x, a)| \leqslant R$ for all $(x, a) \in K$. Moreover for each $x \in S$,

(b) $A(x)$ is compact,

(c) $a \to r(x, a)$ is continuous on $A(x)$, and

(d) $a \to \int q(x, a, dy)\, u(y)$ is continuous on $A(x)$ for each bounded measurable function $u: S \to \boldsymbol{R}$.

Let X_n and A_n be the state and action at the n-th stage, respectively, $n = 0, 1, \ldots$ A given realization of $(X_0, A_0, X_1, A_1, \ldots)$ is denoted by $(x_0, a_0, x_1, a_1, \ldots)$.

A *policy* d is a sequence $d = (d_0, d_1, \ldots)$, where $d_n(h_n, \cdot)$ is a conditional probability measure on the Borel sets of A, given the history of the process $h_n = (x_0, a_0, \ldots, x_{n-1}, a_{n-1}, x_n)$, and it satisfies

$$d_n(h_n, A(x_n)) = 1, \quad n = 0, 1, \ldots$$

A *Markov* policy is a sequence $(f_0, f_1, \ldots)$ of functions $f_n \in F$, where F is the collection of all measurable functions $f: S \to A$ such that $f(x) \in A(x)$ for all $x \in S$. As usual we identify F with the set of *stationary* policies, i.e., Markov policies of the form $(f, f, \ldots)$, $f \in F$.

Finally, the objective function is

$$v(d, x) := E_x^d \left[\sum_{n=0}^{\infty} \beta^n r(X_n, A_n) \right] \tag{1}$$

the expected total discounted reward when policy d is used and the initial state is x. A policy d is said to be *optimal* if it satisfies $v(d, x) = v^*(x)$, $x \in S$, where v^* is the *optimal reward function* defined by

$$v^*(x) = \sup_d v(d, x), \quad x \in S. \tag{2}$$

As mentioned in the Introduction, we are interested in obtaining a procedure approximating v^* and in determining an asymptotically optimal stationary policy, which is done in Sections 3 and 4, respectively. The results are then applied (Section 5) to decision processes depending on unknown parameters. An important role is played by the following well-known result [3, 4, 10, 11, 16]:

PROPOSITION 1. Assume (A1, A2). Then (a) v^* is a bounded measurable function and it satisfies the *optimality equation*

$$v^*(x) = \sup_{a \in A(x)} \left[r(x, a) + \beta \int_S q(x, a, dy)\, v^*(y) \right], \quad x \in S. \tag{3}$$

(b) A stationary policy $f \in F$ is optimal if, and only if, it satisfies

$$v^*(x) = r(x, f(x)) + \beta \int q(x, f(x), dy)\, v^*(y), \quad x \in S.$$

The existence of an optimal stationary policy is ensured under (A1, A2).

NOTATION. $B(S)$ denotes the space of real-valued bounded measurable functions u on S with the supremum norm $\|u\| = \sup_x |u(x)|$. $M(S)$ is the space of finite signed measures μ on S with the total variation norm $\|\mu\|$ (see, e.g., [19]).

We shall use further the following obvious facts. For any $u \in B(S)$ and $\mu \in M(S)$,

$$\left| \int u\, d\mu \right| \leqslant \|u\|\, \|\mu\|. \tag{4}$$

If $u, v \in B(S)$, then (see, e.g., [11, Lemma 3.3])

$$|\sup_x u(x) - \sup_x v(x)| \leqslant \sup_x |u(x) - v(x)|. \tag{5}$$

3. Nonstationary value-iteration

The nonstationary value-iteration (NVI) scheme introduced by Federgruen and Schweitzer [2, Theor. 3.1 (a)] for finite state and action spaces is extended in Theorem 1 below to the decision model (S, A, q, r, β) described in Section 2.

Consider a sequence of decision models (S, A, q_n, r_n, β), $n = 0, 1, \ldots$, each of which satisfies Assumptions (A1) and (A2), and such that they "converge" to (S, A, q, r, β) in the following sense.

(A3) As $n \to \infty$,

(a) $\eta(n) := \sup_{(x,a) \in K} |r_n(x, a) - r(x, a)| \to 0$,

and

(b) $\pi(n) := \sup_{(x,a) \in K} \|q_n(x, a, \cdot) - q(x, a, \cdot)\| \to 0$,

where $\| \ \|$ denotes the total variation norm.

Note that (A3) is equivalent to:

(A3′). As $\eta \to \infty$,

$$\bar{\eta}(n) := \sup_{m \geqslant n} \eta(m) \to 0 \quad \text{and} \quad \bar{\pi}(n) := \sup_{m \geqslant n} \pi(m) \to 0.$$

Also note that both sequences $\bar{\eta}(n)$ and $\bar{\pi}(n)$, $n=0, 1, \ldots$, are non-increasing.

Now let $v_n(\cdot)$, $n=0, 1, \ldots$, be the functions in $B(S)$ defined by

$$v_0(x) := \sup_{a \in A(x)} r_0(x, a), \quad x \in S,$$

and for $n=1, 2, \ldots$,

$$v_n(x) := \sup_{a \in A(x)} \left[r_n(x, a) + \beta \int q_n(x, a, dy) v_{n-1}(y) \right], \quad x \in S. \tag{6}$$

Note that, for all n, v_n and the optimal reward function v^* in (2) are bounded:

$$\|v^*\| \leqslant c_1 \quad \text{and} \quad \|v_n\| \leqslant R \sum_{k=0}^{n} \beta^k \leqslant c_1, \tag{7}$$

where $c_1 = R/(1-\beta)$.

THEOREM 1. *If* (A1, A2, A3) *hold, then* $\|v_n - v^*\| \to 0$. *More precisely,*

(a) $\|v_n - v^*\| \leqslant c \cdot \max\{\bar{\eta}([n/2]), \ \bar{\pi}([n/2]), \ \beta^{[n/2]}\}$, $n \geqslant 0$, *where* $c = (1+\beta c_1)/ /(1-\beta) + 2c_1 = (1+\beta c_1 + 2R)/(1-\beta)$, *and* $[r]$ *denotes largest integer* $\leqslant r$. *Moreover, if the sequences* $\eta(n)$ *and* $\pi(n)$ *in* (A3) *are themselves non-increasing, then* $\bar{\eta}$ *and* $\bar{\pi}$ *can be substituted by* η *and* π, *respectively, to obtain*:

(b) $\|v_n - v^*\| \leqslant c \cdot \max\{\eta([n/2]), \ \pi([n/2]), \ \beta^{[n/2]}\}$.

P r o o f. The proof is essentially the same as that of Theorem 3.1 (a) in [2], but is included here for completeness. First note that, by (7), we can apply (5) to functions v_{n+1} and v^* (with v^* as in (3)). That is, for any x in S,

$$\begin{aligned} &|v_{n+1}(x) - v^*(x)| \leqslant \\ &\leqslant \sup_{a \in A(x)} \left| r_{n+1}(x, a) - r(x, a) + \beta \int q_{n+1}(x, a, dy) v_n(y) - \beta \int q(x, a, dy) v^*(y) \right|. \end{aligned}$$

Now inside the absolute value on the right-hand side, add and substract the term $\beta \int q_{n+1}(x, a, dy) v^*(y)$, and then use the triangle inequality, the inequality (4), and take the supremum over all $x \in S$, to obtain

$$\|v_{n+1} - v^*\| \leqslant \eta(n+1) + \beta \|v^*\| \pi(n+1) + \beta \|v_n - v^*\|.$$

Therefore, for all $m = 1, 2, \ldots$,

$$\|v_{n+m} - v^*\| \leqslant \sum_{k=0}^{m-1} \beta^k [\eta(n+m-k) + \beta c_1 \pi(n+m-k)] + \beta^m \|v_n - v^*\|. \tag{8}$$

Now, since $\|v_n - v^*\| \leqslant 2c_1$ and $\bar{\eta}(n) \geqslant \eta(n+k)$ and $\bar{\pi}(n) \geqslant \pi(n+k)$ for all k, it follows from (8) that

$$\begin{aligned} \|v_{n+m} - v^*\| &\leqslant [\bar{\eta}(n) + \beta c_1 \bar{\pi}(n)]/(1-\beta) + 2c_1 \beta^m \\ &\leqslant c \cdot \max\{\bar{\eta}(n), \bar{\pi}(n), \beta^m\}. \end{aligned} \tag{8'}$$

Then making the substitution $k=n+m$ with $n=[k/2]$ and $m=k-[k/2]\geq[k/2]$, inequality (8′) reduces to

$$\|v_k-v^*\|\leq c\cdot\max\{\bar{\eta}([k/2]),\ \bar{\pi}([k/2]),\ \beta^{[k/2]}\},$$

which proves (a). Finally, to obtain (b) just note that if $\eta(n)$ and $\pi(n)$ are non-increasing, then (8′) holds when $\bar{\eta}$ and $\bar{\pi}$ are substituted by η and π, respectively. ∎

Several interesting applications of the NVI scheme are mentioned by Federgruen and Schweitzer in [2, Section 1]. Here we will use it to obtain asymptotically optimal policies (Section 4 below) and to obtain adaptive policies for decision processes depending on unknown parameters. A similar approach has been used in [6] to obtain finite-state approximations for denumerable MDP's.

4. Asymptotically optimal policies

Consider function $\varphi: K\to R$ defined by

$$\varphi(x, a)=r(x, a)+\beta\int q(x, a, dy)\,v^*(y)-v^*(x). \tag{9}$$

This function has been widely used [3, 4, 5, 17] as a measure of the "difference" between an optimal action in state x and any other action $a\in A(x)$. For instance, in terms of φ, Proposition 1 can be restated as follows:

PROPOSITION 1′. Assume (A1, A2). (a) *Optimality equation*: $\sup_{x\in A(x)}\varphi(x, a)=0$.

(b) *Optimality criterion.* A stationary policy f is optimal iff $\varphi(x, f(x))=0$ for all $x\in S$.

Here we use φ to define asymptotic optimality.

DEFINITION 1. A Markov policy $\{f_n\}$, i.e., a sequence of functions $f_n\in F$, is *asymptotically optimal* (AO) if, for each $x\in S$, $\varphi(x, f_n(x))\to 0$ as $n\to\infty$.

COMMENT. Asymptotic optimality is related to the following concept due to Schäl [22]. A policy d is asymptotically optimal in the sense of Schäl if, for every $x\in S$,

$$V_N(d, x)-E_x^d\,v^*(X_N)\to 0 \text{ as } N\to\infty, \tag{10}$$

where

$$V_N(d, x):=E_x^d\left[\sum_{n=N}^{\infty}\beta^{n-N}\,r(X_n, A_n)\right]$$

is the expected total reward from stage N onwards discounted at stage N. This concept of asymptotic optimality can be related to that in Definition 1 by the fact that [22, Theor. 4.12] (see also [5, 9]) the left-hand side of (10) can be written as

$$V_N(d, x)-E_x^d\,v^*(X_N)=E_x^d\left[\sum_{n=N}^{\infty}\beta^{n-N}\,\varphi(X_n, A_n)\right].$$

Thus a sufficient condition for (10) is the following: φ is a bounded function and $\varphi(X_n, A_n) \to 0$ P_x^d — almost surely as $n \to \infty$. ∎

We now use the NVI scheme (6) to define an A0 policy. First note that under Assumptions (A1) and (A2), for each $n=0, 1, \ldots$, there is a measurable function $f_n: S \to A$ such that $f_n(x) \in A(x)$, and

$$\begin{aligned} &v_0(x) = r_0(x, f_0(x)), \quad x \in S \\ &v_n(x) = r_n(x, f_n(x)) \leqslant \beta \int q_n(x, f_n(x), dy)\, v_{n-1}(y), \quad x \in S. \end{aligned} \tag{11}$$

This follows from standard measurable selection theorems; see, e.g., [3, 10, 16]. Thus $\{f_n\}$ is a Markov policy and we also have the following:

THEOREM 2. *Under the assumptions of Theorem* 1, $\{f_n\}$ *is* A0. *Furthermore, the asymptotic optimality is uniform in the sense that*

$$\|\varphi\|_n := \sup_{x \in S} \left|\varphi(x, f_n(x))\right| \to 0 \text{ as } n \to \infty.$$

Proof. From (9) and (11),

$$\begin{aligned} \varphi(x, f_n(x)) &= \varphi(x, f_n(x)) + v_n(x) - v_n(x) = \\ &= r(x, f_n(x)) - r_n(x, f_n(x)) + \beta \int q(x, f_n(x), dy)\, v^*(y) + \\ &\quad - \beta \int q_n(x, f_n(x), dy)\, v_{n-1}(y) + v_n(x) - v^*(x). \end{aligned}$$

On the right side, add and substract the term

$$\beta \int q_n(x, f_n(x), dy)\, v^*(y);$$

then a simple calculation (which uses (4)) shows that

$$\|\varphi\|_n \leqslant \eta(n) + \beta \|v^*\|\, \pi(n) + \beta \|v_{n-1} - v^*\| + \|v_n - v^*\|,$$

from which the desired result is immediately concluded. ∎

5. Adaptive policies

A Markov decision process, say $(S, A, q(\theta), r(\theta), \beta)$, depending on an unknown parameter θ is called an *adaptive* MDP (The name is sometimes used to include MDP's with incomplete state information, as in [11].) To solve these problems, a decision-maker has to identify or estimate the unknown parameter θ while seeking the optimal policy. Thus at each decision epoch, he has to estimate the parameter and "adapt" his actions to the estimated value; policies combining these two functions — parameter estimation and control actions — are called *adaptive* policies. An extensive survey on adaptive decision problems has been given recently by Kumar [13]; additional references can be found in [4, 5, 7–9, 17, 22].

In this section we consider an adaptive MDP $(S, A, q(\theta), r(\theta), \beta)$, where the transition law $q(x, a, \theta, \cdot)$ and the reward function $r(x, a, \theta)$ depend on an unknown parameter θ. In contrast to Bayesian problems [13, 11], we do not have *a priori* information about the true parameter value, except that it belongs to a given parameter set T, which is assumed to be a Polish space. For each $\theta \in T$, the decision model $(S, A, q(\theta), r(\theta), \beta)$ is assumed to satisfy conditions (A1) and the analogue of conditions (A2), namely:

($A2\theta$) (a) $|r(x, a, \theta)| \leqslant R$ for all $(x, a) \in K$, $\theta \in T$. Moreover for each $x \in S$ and $\theta \in T$,
(b) $A(x)$ is compact,
(c) $a \to r(x, a, \theta)$ is continuous on $A(x)$, and
(d) $a \to \int q(x, a, \theta, dy)\, u(y)$ is continuous on $A(x)$, for each $u \in B(S)$.

Under these assumptions, Proposition 1 (or 1′) holds for each fixed $\theta \in T$. In particular, if we define (cf. (1), (2) and (9))

$$v(d, x, \theta) := E_x^{d,\theta}\left[\sum_{n=0}^{\infty} \beta^n r(X_n, A_n, \theta)\right],$$

$$v^*(x, \theta) := \sup_d v(d, x, \theta),$$

and

$$\varphi(x, a, \theta) := r(x, a, \theta) + \beta \int q(x, a, \theta, dy)\, v^*(y, \theta) - v^*(x, \theta), \quad (x, a) \in K,$$

we can rewrite Proposition 1′ as follows.

PROPOSITION 1″. For fixed $\theta \in T$, (a) $\sup_{a \in A(x)} \varphi(x, a, \theta) = 0$; and (b) a stationary policy $f(\cdot, \theta)$ is optimal if, and only if, $\varphi(x, f(x, \theta), \theta) = 0$ for all $x \in S$.

Note that equation (a) in Proposition 1″ is equivalent to

$$v^*(x, \theta) = \sup \left[r(x, a, \theta) + \beta \int q(x, a, \theta, dy)\, v^*(y, \theta)\right], \quad x \in S;$$

cf. [4] section 1.1.

If $\theta \in T$ is the true (but known) parameter value we can approximate the optimal reward function $v^*(\cdot, \theta)$ using an appropriate version of Theorem 1, and an asymptotically optimal policy can be obtained from Theorem 2. To do this, the idea (roughly) is to consider the sequences

$$r_n(x, a) := r(x, a, \theta_n),\ q_n(x, a, \cdot) := q(x, a, \theta_n, \cdot),\ n \geqslant 0,$$

where $(x, a) \in K$ and $\{\theta_n\}$ is a sequence in T converging to θ. We require the θ-analogue of assumptions A3.

(A3θ) for any $\theta \in T$ and any sequence $\{\theta_n\}$ in T such that $\theta_n \to \theta$ as $n \to \infty$,

$$\eta(n, \theta) := \sup_{(x, a) \in K} |r(x, a, \theta_n) - r(x, a, \theta)| \to 0, \text{ and}$$

$$\pi(n, \theta) := \sup_{(x, a) \in K} \|q(x, a, \theta_n, \cdot) - q(x, a, \theta, \cdot)\| \to 0.$$

The θ-analogue of (A3′) holds for the non-increasing sequences

$$\bar{\eta}(n, \theta) = \sup_{m \geqslant n} \eta(m, \theta),\ \bar{\pi}(n, \theta) := \sup_{m \geqslant n} \pi(m, \theta).$$

Similarly, instead of the functions v_n in (6), we now consider

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

and for $n=1, 2, \ldots,$

$$\begin{aligned} v_n(x, \theta_n) &:= \sup_{a \in A(x)} \left[r(x, a, \theta_n) + \beta \int q(x, a, \theta_n, dy)\, v_{n-1}(y, \theta_{n-1}) \right] \\ &= r(x, f_n(x, \theta_n), \theta_n) + \beta \int q(x, f_n(x, \theta_n), \theta_n, dy)\, v_{n-1}(y, \theta_{n-1}), \end{aligned} \tag{12}$$

where, for each $x \in S$, $f_n(x, \theta_n)$ is a measurable maximizer of the right hand side of (12). Note that the right hand side of (12) depends on $\theta^{(n)} := (\theta_0, \theta_1, \ldots, \theta_n)$, so that, strictly speaking, we should write $v_n(x, \theta^{(n)})$ (respectively, $f_n(x, \theta^{(n)})$) instead of $v_n(x, \theta_n)$ (respect., $f_n(x, \theta_n)$). However, we shall keep the latter, shorter, notation. Then Theorems 1 and 2 can be summarized as follows.

Corollary 1. *Assume ($A1$, $A2\theta$, $A3\theta$) and let $\{\theta_n\}$ be any sequence in T converging to θ. Then*

(a) $\|v_n(\cdot, \theta_n) - v^*(\cdot, \theta)\| \to 0$ *as* $n \to \infty$, *and*

(b) $\{f_n(\cdot, \theta_n)\}$ *is asymptotically θ-optimal in the sense that* $\sup_{x \in S} |\varphi(x, f_n(x, \theta_n), \theta)| \to 0$ as $n \to \infty$.

Furthermore, (with the obvious changes in notation) the inequalities in Theorem 1 (a) and (b) also hold in the present case.

To define adaptive policies we first introduce the following definition, where $P_x^{d,\theta}$ denotes the probability measure when policy d is used, x is the initial state, and θ is the true parameter value; cf. [4, 12, 22].

Definition 2. A sequence $\hat{\theta}_n = \hat{\theta}(X_0, A_0, \ldots, X_{n-1}, A_{n-1}, X_n)$ of T-valued measurable functions is said to be a sequence of *strongly consistent* (SC) *estimators* of $\theta \in T$ if, as $n \to \infty$, $\hat{\theta}_n$ converges to θ $P_x^{d,\theta}$-almost surely for any $x \in S$ and any policy d.

Examples of SC estimators in adaptive Markov or semi-Markov decision processes can be seen in [4, 7, 8, 12, 14, 17, 22]. Given a sequence of SC estimators, an adaptive policy is obtained as follows.

Definition 3. Let $\{\hat{\theta}_n\}$ be a sequence of SC estimators of $\theta \in T$. The policy $\hat{d} = (\hat{d}_n, n=0, 1, \ldots)$ defined by

$$\hat{d}_n(x_0, A_0, \ldots, X_{n-1}, A_{n-1}, X_n) = f_n(X_n, \hat{\theta}_n)$$

is called the *NVI adaptive policy*.

Note that, since the convergence in Corollary 1 (b) is uniform in x, we obtain:

Corollary 2. *As $n \to \infty$,*

$$|\varphi(X_n, f_n(X_n, \hat{\theta}_n), \theta)| \to 0 \quad P_x^{\hat{d},\theta}\text{-almost surely.} \tag{13}$$

We can state (13) by saying that the NVI adaptive policy $\hat{d}$ is asymptotically optimal, although strictly speaking Definition 1 is not applicable here, since $\hat{d}$ is not a Markov policy.

To appreciate the goodness of the NVI policy, let us briefly compare it with the "principle of estimation and control (PEC)" introduced by Schäl [22], and which we now describe.

I. For each $\theta \in T$, let $g(\cdot, \theta) \in F$ be an optimal stationary policy (cf. Proposition 1'').

II. Let $\{\hat{\theta}_n\}$ be a sequence of SC estimators of θ, the true parameter value.

III. Define a policy $d' = (d'_n)$ by

$$d'_n(X_0, A_0, \ldots, X_{n-1}, A_{n-1}, X_n) = g(X_n, \hat{\theta}_n). \tag{14}$$

d' is called the *PEC policy*.

The PEC policy is known in adaptive control under the various names of "naive feedback controller", "self-turning regulator", and others, but is very well described as [17] "the method of substituting the estimates into optimal stationary control". The PEC policy has been widely used in decision processes with *average*-reward criterion [4, 7, 8, 14, 17], but to the best of our knowledge, Schäl's paper [22] was the first application to *discounted*-reward problems (with *denumerable* state space). To prove that d' is asymptotically optimal (see Theorem 3 (b) below) we need the following:

LEMMA 1. *Assume* $(A1, A2\theta, A3\theta)$. If $\theta_n \to \theta$, *then*

$$\|v^*(\cdot, \theta_n) - v^*(\cdot, \theta)\| \to 0. \tag{15}$$

Proof. For any x in S, we obtain from (5),

$$|v^*(x, \theta_n) - v^*(x, \theta)| \leq \sup_{a \in A(x)} \Big| r(x, a, \theta_n) - r(x, a, \theta) + \\ + \beta \int q(x, a, \theta_n, dy)\, v^*(y, \theta_n) - \beta \int q(x, a, \theta, dy)\, v^*(y, \theta) \Big|,$$

and therefore (using (7)),

$$\|v^*(\cdot, \theta_n) - v^*(\cdot, \theta)\| \leq \eta(n, \theta) + \beta c_1 \pi(n, \theta) + \beta \|v^*(\cdot, \theta_n) - v^*(\cdot, \theta)\|,$$

that is,

$$(1-\beta)\|v^*(\cdot, \theta_n) - v^*(\cdot, \theta)\| \leq \eta(n, \theta) + \beta c_1 \pi(n, \theta). \quad \blacksquare$$

THEOREM 3. *Under the assumptions of Lemma* 1 *we have*:

(a) *If* $\theta_n \to \theta$, *then*

$$\|\varphi(\cdot, g(\cdot, \theta_n), \theta)\| = \sup_x |\varphi(x, g(x, \theta_n), \theta)| \to 0 \quad as\ n \to \infty$$

(b) *The* PEC *policy* d' *is asymptotically optimal in the sense that, as* $n \to \infty$,

$$|\varphi(X_n, g(X_n, \hat{\theta}_n), \theta)| \to 0 \quad P_x^{d', \theta}\text{-almost surely for any } x \in S.$$

Proof. Part (a) can be proved as Theorem 1. First note that (cf. Proposition 1'' and I above) since $\varphi(x, g(x, \theta_n), \theta_n) = 0$, we can write

$$\varphi(x, g(x, \theta_n), \theta) = \varphi(x, g(x, \theta_n), \theta) - \varphi(x, g(x, \theta_n), \theta_n).$$

Next, using the definition of $\varphi(x, a, \theta)$ to expand the right hand side, a straightforward calculation shows that

$$\|\varphi(\cdot, g(\cdot, \theta_n), \theta)\| \leqslant \eta(n) + \beta \|v^*(\cdot, \theta)\| \pi(n) + \\ + (1+\beta) \|v^*(\cdot, \theta_n) - v^*(\cdot, \theta)\|,$$

so that (a) can be concluded from (A3θ) and (15). Finally since (a) holds, uniformly in x, for any sequence $\theta_n \to \theta$, (b) holds for any sequence $\{\hat{\theta}_n\}$ of SC estimators.

If follows from Corollary 2, Theorem 3 (b) and the comment following Definition 1, that the NVI and the PEC adaptive policies are both asymptotically optimal in the sense of Schäl [22]. Note also that our proof of the asymptotic optimality of d' (Theorem 3 (b)) is much more elementary than Schäl's proof [22, Theorem 5.21]. This is mainly due to the fact that, instead of the recurrency assumption 2.5 in [22], we have introduced the "uniform continuity" conditions (A3θ).

Finally, note that, from the point of view of applications, the main disadvantage of the PEC policy d' with respect to our NVI policy $\hat{d}$ is in step I above: d' requires to determine *in advance* the optimal stationary policies $g(\cdot, \theta)$ for all values of θ.

6. Concluding remarks

As noted in the Introduction the underlying motivation for the present work was our interest in Markov decision processes with incomplete state information (MDP-ISI) *and* depending on unknown parameters. Having transformed the original MDP-ISI to a MDP with complete state information in which the new state space is a space of probability measures [11, 18, 20, 21, 23] it might be thought that the adaptive policies (NVI and PEC) in Section 5 above are applicable. However, these adaptive policies are defined in terms of a sequence of SC estimators with are based on *complete* observations of the state (and action) sequence(s). Thus application of the results in Section 5 to an MDP-ISI there still remains the problem of showing that a sequence of SC estimators, based on *incomplete* state observations, can be constructed. We do not have an answer to this problem, at present, but perhaps results like those of Baum and Petrie [1] for finite-state non-controlled partially observed Markov chains might be extended to an MDP-ISI.

References

[1] Baum L. E., Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, **37** (1966) 1554–1563.

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

[3] Georgin J.-P. Contrôle des chaines de Markov sur des éspaces arbitraires. *Annales de l'Institut Henri Poincaré*, Section B, **16** (1978) 255–277.

[4] GEORGIN J.-P. Estimation et contrôle des chaines de Markov sur des éspaces arbitraires. Lecture Notes in Mathematics 636 (1978) 71–113.

[5] HERNÁNDEZ-LERMA O. Nonstationary value-iteration and adaptive control of discounted semi--Markov processes. *Journal of Mathematical Analysis and Applications*, 1985. To appear.

[6] HERNÁNDEZ-LERMA O. Finite-state approximations for denumerable multidimensional state discounted Markov decision processes. *Journal of Mathematical Analysis and Applications*, 1985. To appear.

[7] HERNÁNDEZ-LERMA O., MARCUS S. I. Adaptive control of service in queueing systems. *Systems and Control Letters*, **3** (1983) 283–289.

[8] HERNÁNDEZ-LERMA O., MARCUS S. I. Optimal adaptive control of priority assignment in queueing systems. *Systems and Control Letters*, **4** (1984) 65–72.

[9] HERNÁNDEZ-LERMA O., MARCUS S. I. Adaptive control of discounted Markov decision chains. *Journal of Optimization Theory and Applications*. To appear in Vol. 46, No. 2, June 1985.

[10] HIMMELBERG C. J., PARTHASARATHY T., VAN VLECK F. S. Optimal plans for dynamic programming problems. *Mathematics of Operations Research*, **1** (1976) 390–394.

[11] HINDERER K. Foundations of non-stationary dynamic programming with discrete time parameter, Lecture Notes in Operations Research 33 (Springer-Verlag, New York, 1970).

[12] KOLONKO M. Strongly consistent estimation in a controlled Markov renewal model. *Journal of Applied Probability*, **19** (1982) 532–545.

[13] KUMAR P. R. A survey of some results in stochastic adaptive control. Preprint, Dept. of Mathematics, University of Maryland, Baltimore County, 1984.

[14] KURANO M. Discrete-time markovian decision processes with an unknown parametr: average return criterion. *Journal of the Operations Research Society of Japan*, **15** (1972) 67–76.

[15] LANGEN H.-J. Convergence of dynamic programming models. *Mathematics of Operations Research*, **6** (1981), 493–512.

[16] MAITRA A. Discounted dynamic programming on compact metric spaces. Sankhya 30A (1968) 211–216.

[17] MANDL P. Estimation and control of Markov chains. *Advances in Applied Probability*, **6** (1974) 40–60.

[18] RHENIUS D. Incomplete information in Markovian decision models. *Annals of Statistics*, **2** (1974) 1327–1334.

[19] ROYDEN H. L. Real Analysis. New York, Macmillan, 1968.

[20] SAWAKI K., ICHIKAWA A. Optimal control for partially observable Markov decision processes over an infinite horizon. *Journal of the Operations Research Society of Japan*, **21** (1978) 1–15.

[21] SAWARAGI Y., YOSHIKAWA T. Discrete time Markovian decision processes with incomplete state observation. *Annals of Mathematical Statistics*, **41** (1970) 78–86.

[22] SCHÄL M. Estimation and control in discounted stochastic dynamic programming. Preprint No. 428, SFB 72, Universität Bonn, Bonn, 1981.

[23] WAKUTA K. Semi-Markov decision processes with incomplete state observation-discounted cost criterion. *Journal of the Operations Research Society of Japan*, **25** (1982) 351–362.

[24] WHITT W. Approximations of dynamic programs, I. *Mathematics of Operations Research*, **3** (1978) 231–243.

Received, May 1985.

Aproksymacja i strategie optymalne w programowaniu dynamicznym z dyskontem

W pracy przedstawiono iteracyjną metodę aproksymacji funkcji Bellmana w zadaniach programowania dynamicznego z dyskontem przy nieskończonym horyzoncie czasowym w niezupełnych Banacha (polskich) przestrzeniach stanów i sterowania. Opisana metoda jest następnie zastosowana

do określenia asymptotycznie optymalnego rozwiązania a po połączeniu ze zbieżną metodą estymacji parametrów także do określenia optymalnej strategii podejmowania decyzji. Otrzymaną strategię porównano ze strategią wynikającą z zaproponowanej przez Schäla w 1981 r. zasady estymacji i sterowania, którą w pracy uogólniono na zagadnienia decyzyjne w przestrzeniach niezupełnych Banacha.

Аппроксимация и адаптационные стратегии в динамическом программировании с переоценкой

В работе рассматривается итеративный метод аппроксимации функции Бельмана для задач динамического программирования с переоценкой при бесконечном временном горизонте в неполных банаховых польских пространствах состояний и управлений. Рассмотрена процедура используется также для определения оптимальной стратегии. Рассматриваются тоже модели с неизвестными параметрами. Нашу итерационную процедуру объединено с некоторым методом состоятельного оценивания и получено процедуру нахождения асимптотически оптимальной стратегии. Эту стратегию сравнено с „принципом оценивания и управления", введённым Шелём. Принцип Шеля обобщено на банаховые польские пространства.