Onésimo Hernández-Lerma

Approximation and adaptive control of Markov processes: Average reward criterion

Kybernetika, Vol. 23 (1987), No. 4, 265--288

Persistent URL: http://dml.cz/dmlcz/125648

KYBERNETIKA — VOLUME 23 (1987), NUMBER 4

APPROXIMATION AND ADAPTIVE CONTROL OF MARKOV PROCESSES: AVERAGE REWARD CRITERION

ONÉSIMO HERNÁNDEZ-LERMA

Several procedures to approximate the optimal value of average-reward controlled Markov processes with Borel state and control spaces are introduced. The procedures are then used to obtain (i) optimal policies, and (ii) optimal *adaptive* policies for control processes depending on unknown parameters. The latter include the well known "method of substituting the estimates into optimal stationary controls". The approximation procedures are based on a nonstationary version of the value-iteration scheme.

1. INTRODUCTION

In this paper we introduce several procedures to approximate the optimal value of infinite-horizon average-reward controlled Markov processes (CPM's) with Borel state and control spaces. The procedures are then used to determine (i) optimal policies, and (ii) optimal *adaptive* policies for CMP's depending on unknown parameters. The policies obtained in (ii) include the "method of substituting the estimates into optimal stationary controls", also known as the "principle of estimation and control (PEC)", introduced independently by Kurano [22] and Mandl [24, 25], and a policy based on the "nonstationary value-iteration" scheme proposed by Federgruen and Schweitzer [6] for finite state Markov decision processes, extended here to Borel CMP's. Related adaptive policies in [1, 2, 3, 10] are also briefly discussed.

A common feature of most of these adaptive policies is that they can be obtained by suitable modifications to the standard *value iteration* (VI) scheme, also known as the "method of successive approximations". Thus an important part of this paper is the extension to Borel CMP's of VI results by White [31], Hordijk et al. [20], Tijms [29] and many other authors. Related results for *discounted* reward CMP's are given in [12, 13, 16, 28].

Organization of the paper

In Section 2, we introduce the CMP we will be dealing with, together with the basic assumptions on it. Additional assumptions are introduced in later sections, as needed. Also in Section 2 we summarize the required background material, namely, optimality conditions (Theorem 2.4) and ergodicity conditions (2.6).

Section 3 is on the VI scheme. The main result is Theorem 3.9 in which several uniform approximations to the optimal average reward are provided. This theorem extends well-known results, and, for completeness, a proof of it is given in an appendix (Section 7). At the end of Section 3, we extend to Borel spaces several results by Baranov on "successive averagings" for Markov decision processes with finite state and control spaces.

In Section 4, we consider a sequence of CMP's and give conditions for it to converge to a limiting CMP; uniform approximations to the limiting optimal reward as well as optimal policies are also obtained. These results, which are important in themselves, are then used (in Section 6) to prove the optimality of the PEC adaptive policy.

Section 5 deals with a Nonstationary Value Iteration (NVI) scheme, which – as in Section 4 – also provides uniform approximations to the optimal value function, as well as optimal policies. The main difference to notice between the results in Section 4 and the NVI approximations is that the latter are *recursive*, whereas the former are not. The NVI Theorem 5.7 extends results in [6] on finite Markov decision processes.

Section 6 is on adaptive CMP's; the results in Sections 3, 4 and 5 are used to obtain approximations and optimal adaptive policies for CMP's with unknown parameters, provided that a consistent parameter-estimation scheme is given. We thus present new results on the PEC adaptive policy [9, 11, 22, 23, 24, 25], the VI adaptive policy [1, 2], and the NVI adaptive policy studied for discounted reward problems in [12, 13, 16, 17]. It is also shown how to obtain other adaptive policies, as those in [3] and [10].

Finally, in Section 7, a proof of the VI Theorem 3.9 is given.

Terminology and notation

A *Borel space,* say X, is a measurable subset of a complete separable metric space, endowed with the Borel sigma-algebra $\mathscr{B}(X)$. The Cartesian product of sets X and Y is denoted by XY. We denote by $M(X)$ the space of real-valued bounded measurable functions on X. Given two Borel spaces X and Y, a *stochastic kernel* (or conditional probability measure) on X given Y is a function $q(\mathrm{d}x \mid y)$ such that for each $y \in Y$, $q(\cdot \mid y)$ is a probability measure on X, and for each Borel set $B \in \mathscr{B}(X)$, $q(B \mid \cdot)$ is a Borel-measurable function on Y. For a signed measure μ on $\mathscr{B}(X)$, $\|\mu\|$ denotes the total variation norm [26]. For a real-valued function v, $\|v\|$ and $\|v\|_s$ denote

the supremum norm and the span seminorm, respectively, i.e.,

$$\|v\| := \sup_x |v(x)|, \quad \text{and} \quad \|v\|_s := \sup_x v(x) - \inf_x v(x).$$

"iff" means "if and only if", and "a.s." means "almost-surely".

2. PRELIMINARIES

In this section we first introduce the control model we will be dealing with, together with the underlying assumptions. We then give some basic optimality conditions in Theorem 2.4 below, assuming the existence of a solution to the so-called optimality equation (OE); finally, we present several (ergodicity) conditions under which one such solution is insured to exist.

Controlled Markov processes (CMP's)

A CMP is characterized by four objects (X, A, q, r) where:

(a) X is the *state space*, which is assumed to be a Borel space.

(b) A is the *action* (or *control*) *set*, a Borel space. For each $x \in X$, the set of admissible actions (or controls) in state x is denoted by $A(x)$ and is assumed to be a non-empty measurable subset of A. We also assume that the set of admissible state-action pairs

$$\mathbb{K} := \{(x, a) \mid x \in X \text{ and } a \in A(x)\}$$

is a measurable subset of the product space XA. The elements (x, a) in $\mathbb{K}$ sometimes will be denoted by k.

(c) $q(\mathrm{d}x \mid k)$, the so-called *law of motion* (or *transition law*), is a stochastic kernel on X given $\mathbb{K}$.

(d) $r: \mathbb{K} \to \mathbb{R}$ is a measurable function denoting the one-step *reward* (or return or income) function.

A CMP models a system that is observed at times $t = 0,1, \ldots$, with states and actions denoted by x_t and a_t at time t. If the system is in state $x_t = x \in X$ at time t and we take the action $a_t = a \in A(x)$, we are paid the reward $r(x, a)$ and the system moves to a new state $x_{t+1} = x'$ according to the probability distribution $q(\cdot \mid x, a)$ on X. Once the transition into x' has occurred, a new control $a' \in A(x')$ is chosen and the process is repeated. An example of a CMP is a difference equation control model of the form

$$\textbf{2.1} \qquad x_{t+1} = F(x_t, a_t, \xi_t), \quad \text{where} \quad t = 0, 1, \ldots; \quad x_0 \text{ is given},$$

and the disturbance sequence $\{\xi_t\}$ are i.i.d. random elements independent of x_0 [5, 10, 16, 19].

In adaptive control problems q and r will be allowed to depend on unknown parameters θ, in which case the CMP will be written as $(X, A, q(\theta), r(\theta))$.

The vector $h_t := (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t)$, where $(x_i, a_i) \in \mathbb{K}$ for all $i = 0, 1, \ldots$ $\ldots$, is called the *history* of the CMP up to time t. For each $t \geqq 0$, h_t is a vector in the space of histories H_t, where $H_0 := X$, and $H_t := \mathbb{K}H_{t-1}$ for $t = 1, 2, \ldots$.

A *policy* is a sequence $\delta = \{\delta_t\}$, where $\delta_t(\cdot \mid h_t)$ is a conditional probability measure on A given H_t, and satisfying the constraint

$$\delta_t(A(x_t) \mid h_t) = 1 \quad \text{for all} \quad h_t \in H_t \quad \text{and all} \quad t \geqq 0 .$$

A *Markov policy* is a sequence $(f_0, f_1, \ldots)$ of functions $f_t \in \mathbb{F}$, where $\mathbb{F}$ is the collection of all measurable functions $f: X \to A$ such that $f(x) \in A(x)$ for all $x \in X$. As usual, we identify $\mathbb{F}$ with the set of all *stationary* policies, i.e., Markov policies of the form $(f, f, \ldots)$, which will be simply denoted by $f \in \mathbb{F}$.

We are concerned in this paper with the problem of maximizing the long-run average expected reward per unit time, or simply, the *average reward* given by

$$\textbf{2.2} \qquad J(\delta, x) := \liminf_{n \to \infty} n^{-1} \mathsf{E}_x^{\delta} \sum_{t=0}^{n-1} r(x_t, a_t)$$

when the policy δ is used and the initial state is x. In 2.2, E_x^{δ} denotes the expectation with respect to the probability measures P_x^{δ} induced by δ and x, together with the transition law q; see, e.g. [5, 19]. In adaptive control models $(X, A, q(\delta), r(\theta))$, we shall write such a probability as $P_x^{\delta,\theta}$ when the true parameter value is θ.

A policy δ^* is said to be (average) optimal if it satisfies $J(\delta^*, x) = J^*(x)$ for all $x \in X$, where

$$J^*(x) := \sup_{\delta} J(\delta, x) , \quad x \in X .$$

Actually under the conditions imposed below (see, e.g., Theorems 2.4 and 2.9) it will follow that $J^*(x)$ is identically constant:

$$J^*(x) = j^* \quad \text{for all} \quad x \in X .$$

Throughout this paper, the CMP (X, A, q, r) is assumed to satisfy the following.

2.3 Assumptions.

(a) For every state $x \in X$, the control set $A(x)$ is a compact subset of A.

(b) $|r(k)| \leqq R < \infty$ for all $k = (x, a) \in \mathbb{K}$, and $r(x, a)$ is a continuous function of $a \in A(x)$ for each x in X.

(c) $\int_X v(y)\, q(\mathrm{d}y \mid x, a)$ is a continuous function of $a \in A(x)$ for each $x \in X$ and each $v \in M(X)$.

Remark. All the results in this paper hold true if in 2.3(b) and (c) we replace "continuous" by "upper semi-continuous". This follows from the measurable selection theorems in [18] or [27].

The following is a well-known result [8, 11, 29].

2.4 Theorem. Suppose there exist a constant j^* and a function v^* in $M(X)$ such that

(OE) $\quad j^* + v^*(x) = \max_{a \in A(x)} \{r(x, a) + \int_X v^*(y)\, q(\mathrm{d}y \mid x, a)\}$ for all $x \in X$.

Then:

(a) $\sup_{\delta} J(\delta, x) \leqq j^*$ for all $x \in X$.

(b) If $f \in \mathbb{F}$ is a stationary policy such that $f(x)$ maximizes the right side of (OE) for all $x \in X$, i.e.,

$$j^* + v^*(x) = r(x, f(x)) + \int v^*(y)\, q(\mathrm{d}y \mid x, f(x)) \quad \text{for all} \quad x \in X,$$

then f is optimal and $J(f, x) = j^*$ for all $x \in X$.

(c) For any policy δ and any x in X,

$$\lim_{n \to \infty} n^{-1} \sum_{t=0}^{n-1} r(x_t, a_t) = j^* \quad P_x^\delta\text{-a.s.}$$

iff

$$\lim_{n \to \infty} n^{-1} \sum_{t=0}^{n-1} \phi(x_t, a_t) = 0 \quad P_x^\delta\text{-a.s.},$$

where $\phi(x, a)$ is the function on $\mathbb{K}$ defined by

$$\phi(x, a) := r(x, a) + \int_X v^*(y)\, q(\mathrm{d}y \mid x, a) - j^* - v^*(x), \quad (x, a) \in \mathbb{K}.$$

2.5 Remarks. (a) If j^* and v^* are as in Theorem 2.4, it is then said that $\{j^*, v^*(\cdot)\}$ is a *solution* to the *optimality equation* (OE).

(b) Mandl [24] introduced $\phi(x, a)$ as a "measure of the difference" between $a \in A(x)$ and an optimal action in state $x \in X$. Notice that the (OE) can also be written as

$$\max_{a \in A(x)} \phi(x, a) = 0.$$

Moreover, Theorem 2.4(b) can be re-stated as follows: If $f \in \mathbb{F}$ is such that

$$\phi(x, f(x)) = 0 \quad \text{for all} \quad x \in X,$$

then f is optimal. Note that one such policy f exists: see the measurable selection theorems in [5, 18, 27, ...].

Ergodicity conditions

The question now is: Under what conditions does there exist a solution $\{j^*, v^*(\cdot)\}$ to the optimality equation (OE)? Before giving an answer, let us introduce the following.

2.6 Ergodicity conditions.

(1) There exists a state $x^* \in X$ and a positive number α_0 such that $q(\{x^*\} \mid k) \geqq \alpha_0$ for all $k \in \mathbb{K}$.

(2) There exists a measure μ on X such that $\mu(X) > 0$ and $q(\cdot \mid k) \geqq \mu(\cdot)$ for all $k \in \mathbb{K}$.

(3) There exists a measure ν on X such that $\nu(X) < 2$ and $q(\cdot \mid k) \leqq \nu(\cdot)$ for all $k \in \mathbb{K}$.

(4) There exists a number $\alpha < 1$ such that for all k and k' in $\mathbb{K}$, $\|q(\cdot \mid k) - q(\cdot \mid k')\| \leqq 2\alpha$, where $\| \ \|$ denotes the variation norm for signed measures.

(5) For any stationary policy $f \in \mathbb{F}$ there exists a probability measure p_f on X such that $\|q_f^t(\cdot \mid x) - p_f(\cdot)\| \leqq c_t$ for all $x \in X$ and $t = 0, 1, \ldots$, where the numbers c_t do not depend on x and f, and $\sum_t c_t < \infty$. Here $q_f^t(\cdot \mid x)$ denotes the t-step transition probability measure of the Markov process $\{x_t\}$ when the stationary policy f is used, given that the initial state is $x_0 = x$; see Remark 2.7 below.

2.7 Remark. (a) The t-step transition probability $q_f^t(\cdot \mid x) = q^t(\cdot \mid x, f(x))$ in 2.6(5) is given recursively by

$$q_f^t(B \mid x) = \int_X q_f^{t-1}(B \mid y)\, q_f(\mathrm{d}y \mid x) \quad \text{for all} \quad B \in \mathcal{B}(X) \quad \text{and} \quad t \geqq 1,$$

where $q_f^0(\cdot \mid x) := \delta_x(\cdot)$ is the unit measure concentrated on x. Note that

$$q_f^1(\cdot \mid x) = q_f(\cdot \mid x) = q(\cdot \mid x, f(x)) \quad \text{for all} \quad x \in X.$$

(b) In 2.6(5), it is easily verified that p_f is the unique invariant probability measure of the state process $\{x_t\}$. Furthermore, the average reward $J(f, x)$ in 2.2 becomes a constant $j(f)$:

$$J(f, x) = \liminf_{n \to \infty} n^{-1} \sum_{t=0}^{n-1} \int_X r(y, f(y))\, q_f^t(\mathrm{d}y \mid x) =$$
$$= \int r(y, f(y))\, p_f(\mathrm{d}y) := j(f) \quad \text{for all} \quad x \in X.$$

2.8 Lemma. The following implications hold for the ergodicity conditions 2.6:

$$(1) \to (2) \to (4) \to (3)$$
$$\downarrow$$
$$(5)$$

This again is a well-known result; for a proof see, e.g., Georgin [8]. In particular, Georgin uses results by Ueno [30] to show that the constants c_t in 2.6(5) can be chosen

as $c_t = 2\alpha^t$ for all $t \geqq 0$, where α is the number in 2.6(4). Other conditions sufficient for 2.6(5) are given in, e.g., [10, 29].

The following result is proved (e.g.) in [8, 11, 23].

2.9 Theorem. In addition to Assumptions 2.3, suppose that either 2.6(1), 2.6(2) or 2.6(3) hold. Then there exists a solution $\{j^*, v^*(\cdot)\}$ to (OE).

There are other ways to obtain a solution to (OE) using directly 2.6(4) or 2.6(5), and viewing the average-reward control problem as a "limit" as β tends to 1 of β-discounted reward problems: see the references cited above.

Motivated by our interest in adaptive control problems, we now turn to the question of how to obtain (if possible, uniform) approximations to the optimal reward j^*.

3. VALUE ITERATION

A common approach to obtain approximations of a solution $\{j^*, v^*(\cdot)\}$ to (OE) is the method of successive approximations or *value iteration* (VI) introduced by White [31] for average reward problems. In this section we summarize and extend to Borel spaces X and A ideas in [31, 7, 20, 29, ...]. We also extend to Borel spaces results by Baranov [4] on successive averagings for finite state and control spaces.

Throughout the remainder part of the paper we suppose that the CMP (X, A, q, r) satisfies the following.

3.1 Assumptions. In addition to Assumptions 2.3, we suppose:

(a) The ergodicity condition 2.6(4) holds, and

(b) There exists a solution $\{j^*, v^*(\cdot)\}$ to the optimality equation (OE).

To begin with, we rewrite (OE) in Theorem 2.4 as

$$j^* + v^*(x) = Tv^*(x), \quad x \in X, \tag{3.2}$$

where T is the operator on $M(X)$ defined by

$$Tv(x) := \max_{a \in A(x)} \{r(x, a) + \int_X v(y)\, q(\mathrm{d}y \mid x, a)\} \quad \text{for all} \quad x \in X. \tag{3.3}$$

Clearly $Tv \in M(X)$ if $v \in M(X)$: see the measurable selection theorems in [5, 18, 27]. A basic result is to show (Theorem 3.9) that T is a *contraction* operator with respect to the *span seminorm*

$$\|v\|_s := \sup_x v(x) - \inf_x v(x).$$

Uniform approximations

Let $\{v_t\}$ be a sequence in $M(X)$ defined recursively by

$$v_t := Tv_{t-1} = T^t v_0 , \quad t = 1, 2, \ldots ,$$

where $v_0 \in M(X)$ is arbitrary; that is, for $t = 1, 2, \ldots$,

3.4 $$v_t(x) = \max_{a \in A(x)} \{r(x, a) + \int v_{t-1}(y)\, q(\mathrm{d}y \mid x, a)\} \quad \text{for all} \quad x \in X .$$

$v_t(x)$ may be interpreted as the maximal expected reward for a planning horizon of t epochs when the initial state is x and the terminal reward $v_0(y)$ is obtained when the final state is $x_t = y$; that is, for any $t \geqq 1$ and $x \in X$,

$$v_t(x) = \sup_{\delta} \mathrm{E}_x^{\delta}\{\sum_{i=0}^{t-1} r(x_i, a_i) + v_0(x_t)\} .$$

Clearly, as $t \to \infty$, the functions v_t might not converge to a function in $M(X)$: take, e.g., $r(x, a)$ identical to a nonzero constant. We shall see, however, that appropriate transformations of v_t do converge.

Let $\{e_t\}$ be the sequence in $M(X)$ defined by

3.5 $$e_t(x) := v_t(x) - tj^* - v^*(x) , \quad \text{where} \quad t = 0, 1, \ldots \quad \text{and} \quad x \in X .$$

Notice that, for all $t \geqq 0$ and $x \in X$,

3.6 $$e_{t+1}(x) = \max_{a \in A(x)} \{\phi(x, a) + \int e_t(y)\, q(\mathrm{d}y \mid x, a)\} ,$$

where $\phi(x, a)$ is the function defined in Theorem 2.4(c). We also have the following.

3.7 Lemma.

(a) $\|e_{t+1}\| \leqq \|e_t\| \leqq \|v_0 - v^*\|$ for all $t \geqq 0$.

(b) The sequence $e_t^+ := \sup_x e_t(x)$ is nonincreasing, whereas $e_t^- := \inf_x e_t(x)$ is nondecreasing.

(c) $\|e_{t+1}\|_s \leqq \|e_t\|_s$ for all $t \geqq 0$, where $\|\cdot\|_s$ denotes the span seminorm.

Proof. (a) Since $\phi(x, a) \leqq 0$, it follows from 3.6 that

$$|e_{t+1}(x)| \leqq \|e_t\| \quad \text{for all} \quad t \geqq 0 \quad \text{and all} \quad x \text{ in } X ,$$

and therefore, $\|e_t\|$ is nonincreasing. For $t = 0$, we have $\|e_0\| = \|v_0 - v^*\|$.

(b) Similarly, 3.6 yields

$$e_{t+1}(x) \leqq \sup_y e_t(y) = e_t^+ .$$

Thus $e_{t+1}^+ \leqq e_t^+$, which proves the first part of (b).

To prove the second part, let $f \in F$ be a stationary policy such that $\phi(x, f(x)) = 0$ for all $x \in X$; see Remark 2.5(b). Then, from 3.6,

$$e_{t+1}(x) \geqq \int e_t(y)\, q_f(\mathrm{d}y \mid x) \geqq \inf_y e_t(y) = e_t^- .$$

(c) This follows from (b). □

By Lemma 3.7, the sequence $\{e_t\}$ is uniformly bounded and decreases in both the sup norm and the span seminorm. We will show below (Theorem 3.9(b)) that $\{e_t\}$ converges exponentially fast to a constant.

3.8 Definition. Let $\delta = \{f_t\}$ be a Markov policy such that $f_t(x)$ maximizes the right side of equation 3.4 for all $t \geqq 1$ and $x \in X$; we take $f_0 \in \mathbb{F}$ arbitrary.

3.9 Theorem. Under Assumption 3.1, we have:

(a) $\|Tu_1 - Tu_2\|_s \leqq \alpha\|u_1 - u_2\|_s$ for all u_1 and u_2 in $M(X)$, where $\alpha < 1$ is the constant in 2.6(4).

(b) There is a constant c such that

$$\sup_x |e_t(x) - c| \leqq \|e_t\|_s \leqq \alpha^t \|e_0\|_s \quad \text{for all} \quad t = 0, 1, \ldots.$$

(c) Let $V_t^+ := \sup_x w_t(x)$ and $V_t^- := \inf_x w_t(x)$, where

$$w_t(x) := v_t(x) - v_{t-1}(x) \quad \text{for all} \quad x \in X .$$

Then V_t^+ is a nonincreasing sequence, whereas V_t^- is nondecreasing, and both sequences converge exponentially fast to j^*, i.e., for all $t \geqq 1$,

$$0 \leqq V_t^+ - j^* \leqq 2\alpha^{t-1}\|e_0\|_s$$

and

$$0 \leqq j^* - V_t^- \leqq 2\alpha^{t-1}\|e_0\|_s .$$

(d) $V_t^- \leqq J(f_t, x) \leqq j^* \leqq V_t^+$ for all $x \in X$ and $t \geqq 1$.

(e) $\sup_x |w_t(x) - j^*| \leqq 2\alpha^{t-1}\|e_0\|_s$ for all $t \geqq 1$.

(f) $\sup_x |J(f_t, x) - j^*| \leqq \sup_x |w_t(x) - j^*|$ for all $t \geqq 1$.

(g) $\sup_x |(v_t(x) - v_t(z)) - (v^*(x) - v^*(z))| \to 0$ as $t \to \infty$ for all $z \in X$.

(h) $\sup_x |\phi(x, f_t(x))| \leqq 2\alpha^{t-1}\|e_0\|_s \to 0$ as $t \to \infty$, and therefore, by Theorem 2.4(c), the Markov policy δ in Definition 3.8, that is, the policy that uses the control $a_t := f_t(x_t)$ at time t, is optimal.

This theorem provides several uniform approximations to the optimal average reward j^*. Observe also that the stationary policy f_t that maximizes the right side of 3.4 may be regarded as approximately optimal for the infinite horizon problem when t is sufficiently large; this is made precise in parts (d), (f) and (h). Theorem 3.9 is essentially contained in the papers by Tijms [29] and Acosta Abreu [1]; however, since these papers are still unpublished, a proof will be given here: see Section 7.

Successive averagings

As a direct consequence of Theorem 3.9 we will now extend to Borel spaces a result of Baranov [4] on successive averagings for Markov decision process with finite state and action spaces.

Let v_t be the VI functions in 3.4 and define $u_t := t^{-1}v_t$. Using 3.4 we can write the u_t iteratively:

3.10 $$u_t = Q_t u_{t-1} \quad \text{for all} \quad t \geqq 1\,, \quad \text{with} \quad u_0(\cdot) := 0\,,$$

where Q_t is the operator on $B(X)$ given by

$$Q_t\, v(x) := \max_{a \in A(x)} \{t^{-1} r(x, a) + (t-1)\, t^{-1} \int v(y)\, q(\mathrm{d}y \mid x, a)\}\,, \quad x \in X\,.$$

For each $t \geqq 1$, the operator Q_t is a contraction with modulus $(t-1)/t$, and therefore, there exists a unique function u_t^* in $M(X)$ such that

3.11 $$u_t^* = Q_t u_t^* \quad \text{for all} \quad t \geqq 1\,.$$

From Theorem 3.9, we then obtain the following.

3.12 Corollary. Suppose that the assumptions of Theorem 3.9 hold. Then, as $t \to \infty$,
(a) $\sup_x |u_t(x) - j^*| \to 0$.
(b) $\|u_t^* - u_t\| \to 0$.
(c) $\sup_x |u_t^*(x) - j^*| \to 0$.

Proof. Part (a) follows from Theorem 3.9(b), whereas (b) follows from:

$$\|u_t^* - u_t\| \leqq (t-1)\, \|u_t - u_{t-1}\| =$$
$$= \|w_t - u_t\| \leqq \sup_x |w_t(x) - j^*| + \sup_x |u_t(x) - j^*|\,,$$

where w_t are the functions in Theorem 3.9(c) and (e).

Part (c) follows from (a) and (b). □

3.13 Remark. Observe that the policy $\delta = \{f_t\}$ in Definition 3.8 is such that $f_t(x)$ also maximizes the right side of the "successive averagings" equation 3.10. Thus from Theorem 3.9(h), we have obtained by a different approach another conclusion in [4]: The policy δ defined via the successive averagings 3.10 is optimal. It can also be proved, using Corollary 3.12(b), that the policy $\delta' = \{f_t'\}$ such that $f_t'(x)$ attains the maximum on the right side of 3.11, is optimal.

4. APPROXIMATING MODELS

Let (X, A, q_t, r_t), where $t = 0, 1, \ldots$, be a sequence of CMP's. In this section we give conditions under which the average-optimal reward of the t-models converges to the optimal reward of a "limit" control model (X, A, q, r); in the following section we shall study the convergence of a nonstationary version of the value iteration functions in Section 3.

Sometimes we shall write as q_∞ and r_∞ the transition law q and the reward function r in the limiting control model (X, A, r, q).

4.1 Assumptions. The control model (X, A, q_t, r_t) satisfies Assumptions 3.1 for all $0 \leqq t \leqq \infty$. Moreover, the sequence $\{r_t\}$ is uniformly bounded and 2.6(4) holds uniformly in t; that is,

(a) $|r_t(k)| \leqq R < \infty$ for all $k \in \mathbb{K}$ and $0 \leqq t \leqq \infty$, and

(b) $\sup_{t,k,k'} \|q_t(\cdot \mid k) - q_t(\cdot \mid k')\| \leqq 2\alpha$, with $\alpha < 1$, where the sup is over all k and k' in $\mathbb{K}$ and all $0 \leqq t \leqq \infty$. In addition:

(c) $\varrho(t) \to 0$ and $\pi(t) \to 0$ as $t \to \infty$, where

$$\varrho(t) := \sup_k |r_t(k) - r(k)| \quad \text{and} \quad \pi(t) := \sup_k \|q_t(\cdot \mid k) - q(\cdot \mid k)\|,$$

and the sup is over all $k \in \mathbb{K}$.

Thus for each t, all the results in Sections 2 and 3 hold. In particular, for each t, there is a bounded solution $\{j_t^*, v_t^*(\cdot)\}$ of the optimality equation in Theorem 2.4 or 3.2, i.e.,

$$\textbf{4.2} \qquad j_t^* + v_t^*(x) = \max_{a \in A(x)} \{r_t(x, a) + \int v_t^*(y)\, q_t(\mathrm{d}y \mid x, a)\} := T_t v_t^*(x), \quad \text{for all } x \in X \text{ and } 0 \leqq t \leqq \infty,$$

where T_t is the operator on $M(X)$ defined by

$$\textbf{4.3} \qquad T_t v(x) := \max_{a \in A(x)} \{r_t(x, a) + \int v(y)\, q_t(\mathrm{d}y \mid x, a)\}, \quad v \in M(X), \quad x \in X.$$

For the limit control model $(X, A, q, r) = (X, A, q_\infty, r_\infty)$, we write (sometimes) $j^* = j_\infty^*$ and $v^* = v_\infty^*$, so that 4.2 and 4.3 hold for $t = \infty$.

For each t, let $f_t^* \in \mathbb{F}$ be a stationary policy such that $f_t^*(x)$ maximizes the right side of 4.2 for all $x \in X$. The main results of this section are that j_t^* and v_t^* converge to j^* and v^*, respectively, and that the Markov policy $\delta^* = \{f_f^*\}$, which takes the action

$$a_t := f_t^*(x_t) \quad \text{at time} \quad t = 0, 1, \ldots,$$

is optimal for the limiting control model. We summarize these results as follows.

4.4 Theorem. Under Assumptions 4.1, we have:

(a) For every $0 \leqq t \leqq \infty$, $f \in \mathbb{F}$ and $x \in X$,

$$\|q_{t,f}^n(\cdot \mid x) - q_f^n(\cdot \mid x)\| \leqq n\, \pi(t) \quad \text{for all} \quad n = 0, 1, \ldots,$$

where $q_{t,f}^n(\cdot \mid x)$ denotes the n-step transition probability for the t-model when the stationary policy $f \in \mathbb{F}$ is used and the initial state is x. Here we use the notation

$$q_{\infty,f}^n(\cdot \mid x) = q_f^n(\cdot \mid x).$$

(b) $|j_t(f) - j(f)| \leqq R\, \pi(t)$ for all $0 \leqq t \leqq \infty$ and $f \in \mathbb{F}$, where $j_t(f) = J_t(f, x)$ denotes the average reward for the t-model when the policy $f \in \mathbb{F}$ is used; see Remark 2.7(b). Here, $j_\infty(f) = j(f)$.

(c) $|j_t^* - j^*| \leqq R\, \pi(t)$ for all $0 \leqq t \leqq \infty$.

(d) $\|v_t^* - v^*\|_s \leqq b_0 \cdot \max\{\varrho(t), \pi(t)\}$, where $\|\ \|_s$ denotes the span seminorm and $b_0 := (2 + 2\|v^*\| + R)/(1 - \alpha)$.

(e) $\delta^* = \{f_t^*\}$ is average optimal for the limiting control model.

Proof. (a) For $n = 0$, (a) holds trivially, since $q^0_{t,f}(\cdot \mid x) = \delta_x(\cdot)$ for all t, f and x; see Remark 2.7(a). For $n = 1$, the inequality follows from the definition of $\pi(t)$. For $n > 1$, the inequality in (a) is easily verified by induction.

(b) Since

$$j_t(f) = \lim_{n\to\infty} n^{-1} \int r_t(y, f(y))\, q^n_{t,f}(\mathrm{d}y \mid x) \quad \text{for all} \quad 0 \leqq t \leqq \infty\,,$$

where $j_\infty(f) = j(f)$, we obtain

$$|j_t(f) - j(f)| \leqq \lim_n n^{-1}\{\varrho(t) + nR\,\pi(t)\} = R\,\pi(t)\,.$$

(c) With the notation of part (b), we can write

$$j_t^* = \sup_f j_t(f) \quad \text{for all} \quad 0 \leqq t \leqq \infty\,,$$

where the sup is over all $f \in \mathbb{F}$, and $j_\infty^* = j^*$. Thus (by (b))

$$|j_t^* - j^*| \leqq \sup_f |j_t(f) - j(f)| \leqq R\,\pi(t)\,.$$

(d) By Theorem 3.9(a) applied to the t-model,

$$\|T_t v_t^* - T_t v^*\|_s \leqq \alpha \|v_t^* - v^*\|_s\,.$$

On the other hand, from 4.2, $v_t^* = T_t v_t^* - j_t^*$ for all $0 \leqq t \leqq \infty$ (where $T = T_\infty$ as in 3.3), so that

$$\begin{aligned}\|v_t^* - v^*\|_s &\leqq \|T_t v_t^* - T_t v^*\|_s + \|T_t v^* - Tv^*\|_s + |j_t^* - j^*| \leqq \\ &\leqq \alpha\|v_t^* - v^*\|_s + 2[\varrho(t) + \|v^*\|\,\pi(t)] + R\,\pi(t)\,,\end{aligned}$$

which implies (d).

(e) By Remark 2.5(b), the optimality equation 4.2 for the t-model can also be written as

$$\max_{a\in A(x)} \phi_t(x, a) = 0 \quad \text{for all} \quad x \in X \quad \text{and} \quad 0 \leqq t \leqq \infty\,,$$

where

4.5 $$\phi_t(x, a) := r_t(x, a) + \int v_t^*(y)\, q_t(\mathrm{d}y \mid x, a) - j_t^* - v_t^*(x)\,.$$

Note also that $\phi_t(x, f_t^*(x)) = 0$ for all x and t. Thus by 4.5 and the definition of $\phi(x, a)$ in Theorem 2.4(c), we can expand

$$\phi(x, f_t^*(x)) = \phi(x, f_t^*(x)) - \phi_t(x, f_t^*(x))\,,$$

and then a straightforward calculation using parts (c) and (d) yields

$$|\phi(x, f_t^*(x))| \leqq \varrho(t) + (R + \|v^*\|)\,\pi(t) + \|v_t^* - v^*\|_s \to 0 \quad \text{as} \quad t \to \infty\,.$$

Since this convergence is uniform in x, we conclude from Theorem 2.4(c) that δ^* is optimal for the limiting control model. □

The approximations given by Theorem 4.4 have an inconvenience: To use parts (c), (d) or (e), *first* we have to solve the optimality equation 4.2 for each $t = 0, 1, \ldots$. It would be better, of course, to have a *recursive* approximation scheme. We will present one such scheme in the following section which unfortunately, to converge, it requires assumptions some restrictive than 4.1 above.

5. NONSTATIONARY VALUE ITERATION

We shall consider again the sequence of control models (X, A, q_t, t_t) which "converge" to the control model $(X, A, q, r) = (X, A, q_\infty, r_\infty)$. Now, however, we impose stronger assumptions.

5.1 Assumptions. In addition to Assumptions 4.1, we now suppose that

$$\bar{\varrho} := \sum_{t=0}^{\infty} \varrho(t) \leqq \infty \quad \text{and} \quad \pi := \sum_{t=0}^{\infty} \pi(t) < \infty .$$

The necessity of the new assumptions when doing nonstationary value iteration (NVI) is discussed by Federgruen and Schweitzer [6, page 232], who introduced the NVI scheme for Markov decision processes with *finite* state and control spaces. In this section we extend the NVI scheme to Borel control models; we will follow a development somewhat parallel to that in Section 3.

Thus instead of the VI functions f_t in 3.4, we now define the NVI functions $\bar{v}_t$ as follows: For all $t \geqq 0$ and $x \in X$,

$$\textbf{5.2} \qquad \bar{v}_{t+1}(x) := T_t \bar{v}_t(x) = \max_{a \in A(x)} \{ r_t(x, a) + \int \bar{v}_t(y)\, q_t(\mathrm{d}y \mid x, a) \} ,$$

where $\bar{v}_0 \in M(X)$ is arbitrary. And then, instead of the function e_t in 3.5, we now introduce, for all $t \geqq 0$ and $x \in X$,

$$\textbf{5.3} \qquad d_t(x) := \bar{v}_t(x) - tj^* - v^*(x) ,$$

$$d_t^+ := \sup_x d_t(x) , \quad d_t^- := \inf_x d_t(x) ,$$

and

$$\textbf{5.4} \qquad c_t(x) := d_t(x) + v^*(x) = \bar{v}_t(x) - tj^* .$$

5.5 Lemma. The sequence $\{d_t(\cdot)\}$ (and therefore $\{c_t(\cdot)\}$) is uniformly bounded.

Proof. By definitions 5.2 and 5.3,

$$d_{t+1}(x) = T_t \bar{v}_t(x) - (t+1) j^* - v^*(x) ,$$

or

$$d_{t+1}(x) = \max_{a \in A(x)} \{ r_t(x, a) - r(x, a) + \int d_t(y)\, q_t(\mathrm{d}y \mid x, a) + \\ + \int v^*(y) [q_t(\mathrm{d}y \mid x, a) - q(\mathrm{d}y \mid x, a)] + \phi(x, a) \} .$$

By Remark 2.5(b), $\phi(x, a) \leqq 0$, and therefore,

$$d_{t+1}(x) \leqq \sup_y d_t(y) + \varrho(t) + b_1 \pi(t), \quad \text{where} \quad b_1 := \|v^*\|,$$

which implies

$$d_{t+1}^+ \leqq d_t^+ + \varrho(t) + b_1 \pi(t) \quad \text{for all} \quad t \geqq 0.$$

Similarly,

$$d_{t+1}^- \geqq d_t^- - \varrho(t) - b_1 \pi(t) \quad \text{for all} \quad t \geqq 0.$$

Thus,

$$d_{t+1}^+ \leqq d_0^+ + \sum_{i=0}^{t} [\varrho(i) + b_1 \pi(i)] \leqq d_0^+ + \bar{\varrho} + b_1 \bar{\pi} \quad \text{for all} \quad t \geqq 0,$$

and similarly,

$$d_{t+1}^- \geqq d_0^- - (\bar{\varrho} + b_1 \bar{\pi}) \quad \text{for all} \quad t \geqq 0. \qquad \square$$

In Theorem 5.7 below we use the following notation

$$\gamma(t) := \varrho(t) + M \pi(t), \quad \text{and} \quad \gamma^c(t) := \varrho^c(t) + M \pi^c(t),$$

where

5.6
$$\varrho^c(t) := \sum_{i=t}^{\infty} \varrho(i) \quad \text{and} \quad \pi^c(t) := \sum_{i=t}^{\infty} \pi(i),$$

and M is upper bound for all $\|d_t\|_s$ and $\|c_t\|_s$; such an M exists by Lemma 5.5. Theorem 5.7 is the NVI analogue of the value-iteration Theorem 3.9(b), (e), (g), and (h).

5.7 Theorem. Suppose that Assumptions 5.1 hold. Then:

(a) There exists a constant c such that

$$\sup_x |d_t(x) - c| < (1 + 2M) \cdot D_t \quad \text{for all} \quad t \geqq 0,$$

where

$$D_t := \max \{\varrho^c([t/2]), \ \pi^c([t/2]), \ \alpha^{[t/2]}\}.$$

(b) $\sup_x |\bar{v}_t(x) - \bar{v}_{t-1}(x) - j^*| = \sup_x |d_t(x) - d_{t-1}(x)| \leqq (1 + 2M)(D_t + D_{t-1})$ for all $t \geqq 1$.

(c) $\sup_x |(\bar{v}_t(x) - \bar{v}_t(z)) - (v^*(x) - v^*(z))| \to 0$ as $t \to \infty$ for all $z \in X$.

(d) Let $\bar{\delta} = \{\bar{f}_t\}$, with $\bar{f}_t \in \mathbb{F}$, be the Markov policy such that, for each $t \geqq 0$, $\bar{f}_t(x)$ maximizes the right side of 5.2 for all $x \in X$. Then $\bar{\delta}$ is average optimal for the limiting control model (X, A, q, r).

Parts (a) and (b) of this theorem extend to Borel control models some parts of Theorem 4.1 in Federgruen and Schweitzer [6]. We shall refer to the policy $\bar{\delta}$ in part (d) as an *NVI policy*. In Section 6 we will introduce an *adaptive* NVI policy.

Proof. (a) Let $c_t(\cdot)$ be the function in 5.4; then for all $t \geqq 0$ and $x \in X$,

$$c_{t+1}(x) = T_t c_t(x) - j^* = T c_t(x) - j^* + T_t c_t(x) - T c_t(x),$$

where T (or T_∞) is the operator in 3.3. Thus, since

$$T_t\, c_t(x) - Tc_t(x) \leqq \varrho(t) + M\,\pi(t) =: \gamma(t) \quad \text{for all} \quad t \geqq 0 \quad \text{and} \quad x \in X\,,$$

we have

$$c_{t+1}(x) \leqq T\,c_t(x) - j^* + \gamma(t) \quad \text{for all} \quad t \geqq 0 \quad \text{and} \quad x \in X\,,$$

and therefore,

$$c_{t+n}(x) \leqq T^n\, c_t(x) - nj^* + \sum_{i=t}^{t+n-1} \gamma(i) \leqq T^n\, c_t(x) - nj^* + \gamma^c(t)\,.$$

Substraction of $v^*(x)$ on each side of the latter inequality implies (see 5.4)

$$d_{t+n}(x) \leqq e_{t,n}(x) + \gamma^c(t) \quad \text{for all} \quad t \geqq 0\,, \quad n \geqq 1 \quad \text{and} \quad x \in X\,,$$

where

$$e_{t,n}(x) := T^n\, c_t(x) - nj^* - v^*(x) \quad \text{(cf. 3.5)}\,.$$

A similar argument shows that

$$T_t\, c_t(x) - T c_t(x) \geqq -\gamma(t)$$

and then

$$d_{t+n}(x) \geqq e_{t,n}(x) - \gamma^c(t)\,,$$

so that

$$|d_{t+n}(x) - e_{t,n}(x)| \leqq \gamma^c(t) = \varrho^c(t) + M\,\pi^c(t)\,.$$

On the other hand, by Theorem 3.9(b), there exists some constant c such that, for all $t \geqq 0$ and $n \geqq 1$,

$$\sup_x |e_{t,n}(x) - c| \leqq \alpha^n \|e_{t,0}\|_s = \alpha^n \|d_t\|_s \leqq M\alpha^n\,.$$

Therefore,

$$|d_{t+n}(x) - c| \leqq |d_{t+n}(x) - e_{t,n}(x)| + |e_{t,n}(x) - c| \leqq$$

$$\leqq \varrho^c(t) + M\,\pi^c(t)\,M\alpha^n \leqq (1 + 2M)\,.\,\max\{\varrho^c(t), \pi^c(t), \alpha^n\}\,.$$

Let $m = t + n$, with $t = [m/2]$ and $n = m - t \geqq m - [m/2] \geqq [m/2]$; then the above inequality becomes

$$|d_m(x) - c| \leqq (1 + 2M) \max\{\varrho^c([m/2]), \pi^c([m/2]), \alpha^{[m/2]}\}\,,$$

for all $m \geqq 1$ and $x \in X$, which proves part (a).

(b) and (c): Both follow from (a) and 5.4.

(d) It suffices to verify that

$$\sup_x |\phi(x, \bar{f}_t(x))| \to 0 \quad \text{as} \quad t \to \infty\,, \tag{1}$$

so that (d) follows from Theorem 2.4(c). To simplify the notation let us write $\bar{f}_t(x)$ as a_t; then $\phi(x, \bar{f}_t(x))$ in Theorem 2.4(c) becomes

$$\phi(x, a_t) = r(x, a_t) + \int v^*(y)\, q(\mathrm{d}y \mid x, a_t) - j^* - v^*(x)\,, \tag{2}$$

and 5.2 becomes

$$\bar{v}_{t+1}(x) = r_t(x, a_t) + \int \bar{v}_t(y)\, q_t(\mathrm{d}y \mid x, a_t)\,.$$

Then on the right side of (2) add and substract $\bar{v}_{t+1}(x) - \bar{v}_t(x)$ and then an obvious computation using parts (b) and (c) yields (1). □

Nonstationary successive averagings

A nonstationary version of the "successive averaging" functions in 3.10 is obtained as follows.

Let $\bar{u}_t(x) := t^{-1}\, \bar{v}_t(x)$, where $t = 1, 2, \ldots$ and $x \in X$, and $\bar{v}_t$ are the NVI functions. From 5.2, the $\bar{u}_t$ satisfy

5.8 $$\bar{u}_{t+1}(x) = \bar{Q}_t\, \bar{u}_t(x)\,, \quad \text{for all} \quad t \geqq 0 \quad \text{and} \quad x \in X; \quad u_0(\cdot) := 0\,,$$

where

$$\bar{Q}_t\, v(x) := \max_{a \in A(x)} \{(t+1)^{-1}\, r_t(x, a) + t(t+1)^{-1} \int \bar{u}_t(y)\, q_t(\mathrm{d}y \mid x, a)\,.$$

Clearly, for each $t = 0, 1, \ldots$, the operator $\bar{Q}_t$ is a contraction with modulus $t/(t+1)$, and therefore, there exists a unique function $\bar{u}_t^* \in M(X)$ such that

5.9 $$\bar{u}_t^* = \bar{Q}_t \bar{u}_t^* \quad \text{for all} \quad t = 0, 1, \ldots,$$

and as a consequence of Theorem 5.7, we obtain:

5.10 Corollary. Suppose that the assumptions of Theorem 5.7 hold. Then, as $t \to \infty$,
(a) $\sup_{x \in X} |\bar{u}_t(x) - j^*| \to 0$,
(b) $\|\bar{u}_t - \bar{u}_t^*\| \to 0$,
(c) $\sup_{x \in X} |\bar{u}_t(x) - j^*| \to 0$.

Notice that, with the obvious changes in notation, Remark 3.13 also holds in the present, nonstationary case.

Discounted-like NVI

A review of the results above will show that the measures $q_t(\cdot \mid k)$ may be subprobabilities, i.e. $q_t(X \mid k) \leqq 1$, provided that they satisfy Assumption 5.1. In particular, we may take

$$q_t'(\cdot \mid k) := \beta_t q_t(\cdot \mid k)\,, \quad t = 0, 1, \ldots,$$

where $\{\beta_t\}$ is sequence of positive numbers increasing to 1 and such that

$$\sum_{t=0}^{\infty} (1 - \beta_t) < \infty\,.$$

In such a case, the NVI functions, which we now denote by h_t' (instead of $\bar{v}_t$ as in 5.2),

are defined by

5.11
$$h'_{t+1}(x) := U_t\, h'_t(x) :=$$
$$:= \max_{a \in A(x)} \{r_t(x, a) + \beta_t \int h'_t(y)\, q_t(\mathrm{d}y \mid x, a)\}\,, \quad t = 0, 1, \ldots,$$

where $h'_0 \in M(X)$ is arbitrary. Note that U_t is a contraction operator with modulus β_t.

In terms of U_t, Gordienko [10] studies average-optimal policies for discrete-time systems of the form 2.1, where the noise sequence $\{\xi_t\}$ has *unknown* distribution. In Gordienko's paper, $r_t(k) = r(k)$ for all t, and $q_t(\cdot \mid k)$ is the empirical process of $\{\xi_t\}$. For similar problems in the discounted-reward case, see [16, 17].

6. ADAPTIVE CONTROL MODELS

In this section we consider control models $(X, A, q(\theta), r(\theta))$ depending on a parameter θ that takes values in a Borel space Θ. For each θ in Θ, the θ-model $(X, A, q(\theta), r(\theta))$ is such that the transition law and the one-step reward function depend on θ, i.e., we have $q(\cdot \mid k, \theta)$ and $r(k, \theta)$, where $k \in \mathbb{K}$, but everything remains essentially the same as in Sections 2 and 3 except for notational changes. For instance, instead of the long-run average expected reward per unit time $J(\delta, x)$ in 2.2, we now have

$$J(\delta, x, \theta) := \liminf_{n \to \infty} n^{-1}\, \mathsf{E}_x^{\delta, \theta} \sum_{t=0}^{n-1} r(x_t, a_t, \theta)$$

for each θ in Θ, where $\mathsf{E}_x^{\delta, \theta}$ denotes the expectation with respect to the probability measure $P_x^{\delta, \theta}$ when θ is the true parameter value.

The program for this section is as follows. Firstly, we summarize some of the results in Sections 2 and 3; the idea is to put the parametric models in the appropriate setting. Secondly, we re-state some of the approximation results in Sections 4 and 5, and then those results are used to obtain several *adaptive* policies.

Preliminaries

In the first part of this section we suppose that the θ-analogue of Assumptions 4.1 are valid; namely, we suppose:

6.1 Assumptions. $(X, A, q(\theta), r(\theta))$ satisfies Assumptions 2.3 for all θ in Θ; that is,

- $A(x)$ is a compact subset of A for all x in X.
- $r(x, a, \theta) \in M(\mathbb{K}\Theta)$, and for each $x \in \Theta$, the function $r(x, a, \theta)$ is continuous in $a \in A(x)$.
- $\int_X v(y, \theta)\, q(\mathrm{d}y \mid x, a, \theta)$ is continuous function of $a \in A(x)$ for each $x \in X$, $\theta \in \Theta$ and $v \in M(X\Theta)$.

Moreover,

(a) $|r(k, \theta)| \leqq R < \infty$ for all $k \in \mathbb{K}$ and $\theta \in \Theta$.

(b) $\sup_{\theta,k,k'} \|q(\cdot \mid k, \theta) - q(\cdot \mid k', \theta)\| \leqq 2\alpha$, where $\alpha < 1$, and the sup is over all θ in Θ, and k and k' in $\mathbb{K}$.

(c) For any $\theta \in \Theta$ and any sequence $\{\theta_t\}$ in Θ such that $\theta_t \to \theta$, it holds that

$$\varrho(t, \theta) \to 0 \text{ and } \pi(t, \theta) \to 0 \quad \text{as } t \to \infty,$$

where

$$\varrho(t, \theta) := \sup_k |r(k, \theta_t) - r(k, \theta)|$$

and

$$\pi(t, \theta) := \sup_k \|q(\cdot \mid k, \theta_t) - q(\cdot \mid k, \theta)\| .$$

(d) There exist bounded functions $j^*(\theta) \in M(\Theta)$ and $v^*(x, \theta) \in M(X, \Theta)$ such that

6.2 $$j^*(\theta) + v^*(x, \theta) = T_\theta v^*(x, \theta) \quad \text{for all} \quad x \in X ,$$

where, for each $\theta \in \Theta$, T_θ is the operator on $M(X\Theta)$ defined by

6.3 $$T_\theta v(x, \theta) := \max_{a \in A(x)} \{(r(x, a, \theta) + \int_X v(y, \theta)\, q(\mathrm{d}y \mid x, a, \theta)\} .$$

Conditions sufficient for (b) and (d) may be obtained as in Section 2. Of course, implicit in the description of the θ-control model is the fact that $q(\cdot \mid k, \theta)$ is a Borel-measurable stochastic kernel on X given $\mathbb{K}\Theta$.

On the other hand, existence of solution $j^*(\theta)$, $v^*(\cdot, \theta)$ of the θ-optimality equation in Assumption 6.1(d) allows us to use the θ-version of Theorem 2.4 to obtain the following.

The principle of estimation and control (PEC)

Let f^* be a measurable function from $x\Theta$ to A such that, for each $\theta \in \Theta$ and $x \in X$, the action $f^*(x, \theta) \in A(x)$ maximizes the right side of the (OE) 6.3. (Such a function f^* exists by the measurable selection theorems in [18, 27].) By Theorem 2.4(b), $f^*(\cdot, \theta) \in \mathbb{F}$ is an optimal stationary policy for the θ-model, and from Theorem 4.4 we may conclude the following.

6.4 Theorem. Suppose that Assumptions 6.1 hold and let $\{\theta_t\}$ be any sequence in Θ converging to $\theta \in \Theta$. Then:

(a) $|j^*(\theta_t) - j^*(\theta)| \leqq R\, \pi(t, \theta) \quad \text{for all} \quad t \geqq 0.$

(b) $\|v^*(\cdot, \theta_t) - v^*(\cdot, \theta)\|_s \leqq \text{constant} \cdot \max\{\varrho(t, \theta), \pi(t, \theta)\}.$

(c) Let $\{\hat{\theta}_t\}$ be a sequence of strongly consistent (SC) estimators of $\theta \in \Theta$ (that is, a sequence of measurable functions $\hat{\theta}_t$ from H_t to Θ such that, as $t \to \infty$, $\hat{\theta}_t$ converges to θ $P_x^{\delta,\theta}$-a.s. for all policy δ and all $x \in X$), and let $\delta^* = \{\delta_t^*\}$ be the policy defined by

$$\delta_t^*(h_t) := f^*(x_t, \hat{\theta}_t(h_t)) \quad \text{for all} \quad h_t \in H_t \quad \text{and} \quad t \geqq 0 .$$

Then δ^* is optimal for the θ-model. □

To obtain this theorem it suffices to make the substitutions

6.5 $$r_t(k) := r(k, \theta_t) \quad \text{and} \quad q_t(\cdot \mid k) := q(\cdot \mid k, \theta_t)$$

in Theorem 4.4.

We call δ^* as PEC *adaptive policy* and it was originally introduced by Kurano [22] and Mandl [24]. Examples of SC estimators are given by those authors; see also [9, 14, 15, 28].

Nonstationary value iteration

In addition to 6.1, let us *assume*

6.6 For any θ and θ_t as in 6.1(c),

$$\sum_{t=0}^{\infty} \varrho(t, \theta) < \infty \quad \text{and} \quad \sum_{t=0}^{\infty} \pi(t, \theta) < \infty .$$

Now, instead of the NVI functions $\bar{v}_t(x)$ in 5.2, define

$$\bar{v}_{t+1}(x, \theta_t) := T_{\theta_t} \bar{v}_t(x, \theta_{t-1}) \quad \text{where} \quad t \geqq 0 \quad \text{and} \quad \bar{v}_0(\cdot) := 0 .$$

That is,

$$\bar{v}_1(x, \theta_0) := \max_{a \in A(x)} r(x, a, \theta_0)$$

and for $t \geqq 1$,

6.7 $$\bar{v}_{t+1}(x, \theta_t) := \max_{a \in A(x)} \{ r(x, a, \theta_t) + \int \bar{v}_t(y, \theta_{t-1})\, q(\mathrm{d}y \mid x, a, \theta_t) \} .$$

For each $t \geqq 0$, let $\bar{f}_t(\cdot, \theta_t) \in \mathbb{F}$ be such that $\bar{f}_t(x, \theta_t)$ maximizes the right side of 6.7 for all $x \in X$, and

$$\bar{f}_0(x, \theta_0) := \arg\max_{a \in A(x)} r(x, a, \theta_0) \quad \text{for all} \quad x \in X .$$

Strictly speaking, we should write $\bar{f}_t$ as $\bar{f}(\cdot, \theta_t, \theta_{t-1}, \ldots, \theta_0)$, but we shall keep the shorter notation $\bar{f}_t(\cdot, \theta_t)$.

Using again the substitution 6.5 we can rewrite Theorem 5.7 as follows.

6.8 Theorem. Suppose that Assumptions 6.1 and 6.6 hold and let $\theta_t \to \theta \in \Theta$. Then:

(a) There exists a constant $c(\theta)$ such that

$$\sup_{x \in X} |d_t(x, \theta) - c(\theta)| \to 0 \quad \text{as} \quad t \to \infty ,$$

where $d_t(x, \theta) := \bar{v}_t(x, \theta_{t-1}) - tj^*(\theta) - v^*(x, \theta)$.

(b) $\sup_{x \in X} |\bar{v}_{t+1}(x, \theta_t) - \bar{v}_t(x, \theta_{t-1}) - j^*(\theta)| \to 0$ as $t \to \infty$.

(c) $\sup_{x \in X} |[\bar{v}_{t+1}(x, \theta_t) - \bar{v}_{t+1}(z, \theta_t)] - [v^*(x, \theta) - v^*(z, \theta)]| \to 0$ as $t \to \infty$ for all $z \in X$.

(d) Let $\{\hat{\theta}_t\}$ be a sequence of SC estimators of θ satisfying 6.6 $P_x^{\delta, \theta}$-a.s. for every

policy δ and $x \in X$, and let $\bar{\delta} = \{\bar{\delta}_t\}$ be the policy defined by

$$\bar{\delta}_t(h_t) := \bar{f}_t(x_t, \hat{\theta}_t(h_t)) \quad \text{for all} \quad h_t \in H_t \quad \text{and} \quad t \geqq 0 .$$

Then $\bar{\delta}$ is optimal for the θ-model.

We call δ and NVI *adaptive policy*. Other NVI adaptive policies can be defined via suitable variants of the NVI scheme of Section 5. For instance, for the discounted-like NVI functions in 5.11 see Gordienko [10], or for the nonstationary successive averagings 5.8 see Baranov [3]. For adaptive policies in terms of the value iteration functions v_t in Section 3 see [1, 2].

For discounted reward problems these policies (except Baranov's) are further discussed in [12, 13, 16, 17, 28].

7. APPENDIX: Proof of Theorem 3.9

(a) Let u_1 and u_2 be arbitrary functions in $M(X)$, and let g_1 and g_2 in $\mathbb{F}$ be stationary policies such that

$$Tu_i(x) = r(x, g_i(x)) + \int u_i(y)\, q(\mathrm{d}y \mid x, g_i(x)) \quad \text{for all} \quad x \in X \quad \text{and} \quad i = 1, 2 .$$

Of course,

$$Tu_i(x) \geqq r(x, g_j(x)) + \int u_i(y)\, q(\mathrm{d}y \mid x, g_j(x)) \quad \text{if} \quad j \neq i .$$

Then for any two arbitrary points x and x' in X,

$$\textbf{7.1} \qquad (Tu_1 - Tu_2)(x) - (Tu_1 - Tu_2)(x') \leqq \int_X (u_1(y) - u_2(y))\, \lambda(\mathrm{d}y) ,$$

where λ is the finite signed measure on X given by

$$\lambda(\cdot) := q(\cdot \mid x, g_1(x)) - q(\cdot \mid x', g_2(x')) .$$

By the Jordan-Hahn Decomposition Theorem [26], there exist disjoint measurable sets X^+ and X^- whose union is X and such that

$$\|\lambda\| = \lambda(X^+) - \lambda(X^-) \leqq 2\alpha .$$

where $\|\cdot\|$ denotes the total variation norm, and the latter inequality comes from 2.6(4). Moreover, since $\lambda(X^+) + \lambda(X^-) = 0$, we have $\lambda(X^+) \leqq \alpha$. Thus in inequality 7.1 we obtain

$$\begin{aligned} \int_X (u_1(y) - u_2(y))\, \lambda(\mathrm{d}y) &= \left(\int_{X^+} + \int_{X^-}\right)(u_1 - u_2)\, \lambda(\mathrm{d}y) \leqq \\ &\leqq \int_{X^+} \sup_y (u_1 - u_2)\, \lambda(\mathrm{d}y) + \int_{X^-} \inf_y (u_1 - u_2)\, \lambda(\mathrm{d}y) \leqq \\ &\leqq \|u_1 - u_2\|_s\, \lambda(X^+) \leqq \alpha \|u_1 - u_2\|_s . \end{aligned}$$

Since x and x' are arbitrary points in X, the desired conclusion follows.

(b) The operator T in 3.3, applied to the function $v^*(\cdot) + (t - 1) j^*$, satisfies

$$T(v^* + (t - 1) j^*) = Tv^* + (t - 1) j^* = v^* + tj^* \quad \text{by 3.2.}$$

Thus

$$\|e_t\|_s = \|v_t - tj^* - v^*\|_s = \|Tv_{t-1} - T(v^* + (t-1)j^*)\|_s \leqq$$
$$\leqq \alpha\|v_{t-1} - (t-1)j^* - v^*\|_s = \alpha\|e_{t-1}\|_s ,$$

so that

$$\|e_t\|_s \leqq \alpha^t\|e_0\|_s \quad \text{for all} \quad t \geqq 0 .$$

From this inequality and Lemma 3.7(b) we conclude that there exists a constant c satisfying part (b).

(c) By Definition 3.8 of $f_t \in \mathbb{F}$, we have

$$v_t(x) = r(x, f_t(x)) + \int v_{t-1}(y)\, q(\mathrm{d}y \mid x, f_t(x)) \quad \text{for all} \quad x \in X .$$

On the other hand,

$$v_{t-1}(x) \geqq r(x, f_t(x)) + \int v_{t-2}(y)\, q(\mathrm{d}y \mid x, f_t(x)) \quad \text{for all} \quad x \in X ,$$

and therefore,

$$v_t(x) - v_{t-1}(x) \leqq \sup_y \{v_{t-1}(y) - v_{t-2}(y)\} = V_{t-1}^+ .$$

Thus $V_t^+ \leqq V_{t-1}^+$, and a similar proof yields $V_t^- \geqq V_{t-1}^-$, which proves the first part of (c).

To prove the second part, note that, from 3.5,

7.2 $$w_t(x) = e_t(x) - e_{t-1}(x) + j^*$$

and therefore,

$$V_t^+ = \sup_x \{e_t(x) - e_{t-1}(x)\} + j^* ,$$
$$V_t^- = \inf_x \{e_t(x) - e_{t-1}(x)\} + j^* .$$

Finally, since $\|w_t\|_s = V_t^+ - V_t^- \leqq 2 \sup_x |e_t(x) - e_{t-1}(x)|$, the desired result follows from part (b).

(d) In part (c), we have already shown that $V_t^- \leqq j^* \leqq V_t^+$ for all t. On the other hand,

$$J(f_t, x) \leqq j^* \quad \text{for all} \quad t \geqq 0 \quad \text{and} \quad x \in X .$$

Thus it only remains to prove the first inequality in (d).

To prove this, let us first simplify the notation writing $f_t = g \in \mathbb{F}$ for any fixed $t \geqq 1$; then 3.4 becomes

7.3 $$v_t(x) = r(x, g(x)) + \int v_{t-1}(y)\, q_g(\mathrm{d}y \mid x) .$$

Now, by Assumption 3.1(a), Lemma 2.8 and Remark 2.7(b).

$$J(g,x) = \int r(y,g(y))\, p_g(\mathrm{d}y) = j(g) \quad \text{for all} \quad x \in X .$$

Again by Remark 2.7(b), p_g is the invariant probability measure of $q_g(\cdot \mid x)$ and

therefore, integrating both sides of 7.3 with respect to p_g, we get

$$\int v_t(x)\, p_g(\mathrm{d}x) = \int r(x, g(x))\, p_g(\mathrm{d}x) + \iint v_{t-1}(y)\, q_g(\mathrm{d}y \mid x)\, p_g(\mathrm{d}x) = \\ = j(g) + \int v_{t-1}(y)\, p_g(\mathrm{d}y)\,.$$

Consequently,

7.4 $$j(g) = \int w_t(y)\, p_g(\mathrm{d}y) \quad \text{for all} \quad t \geqq 1\,, \quad \text{where} \quad g = f_t\,.$$

The latter implies

$$V_t^- \leqq j(f_t) \leqq V_t^+ \quad \text{for all} \quad t \geqq 1\,.$$

(e) Follows from (b) and equation 7.2.

(f) Follows from (e) and equation 7.4.

(g) Follows from (b).

(h) To simplify the notation let us write $a_t = f_t(x)$, so that $v_t(x)$ in 3.4 can be written as

$$v_t(x) = r(x, a_t) + \int v_{t-1}(y)\, q(\mathrm{d}y \mid x, a_t)\,,$$

and $\phi(x, f_t(x))$ becomes

$$\phi(x, a_t) = r(x, a_t) + \int v^*(y)\, q(\mathrm{d}y \mid x, a_t) - j^* - v^*(x)\,.$$

On the right side of the latter equation, add and substract $v_t(x)$ and $(t-1)\,j^*$, to obtain

$$\phi(x, a_t) = e_t(x) - \int e_{t-1}(y)\, q(\mathrm{d}y \mid x, a_t)\,,$$

and then (h) is concluded using (b). This completes the proof of Theorem 3.9. □

ACKNOWLEDGEMENTS

After completion of this paper, Prof. Manfred Schäl brought to our attention two related works by Mandl and Hübner [32] and Hordijk and Tijms [33], both on finite-state controlled Markov chains. In the former work [32], several conditions are given for the asymptotic normality of the rewards under various sequences θ_t converging to the "true" parameter value; in [33], a discounted-like iterative method (similar to 5.11 above) is presented. We would like to thank Prof. Schäl for bringing to our attention these papers, and to Prof. Mandl and an unknown reviewer for several helpful remarks.

(Received October 29, 1986.)

REFERENCES

[1] R. S. Acosta Abreu: Control of Markov chains with unknown parameters and metric state space. Submitted for publication. In Spanish.

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

[3] V. V. Baranov: Recursive algorithms of adaptive control in stochastic systems. Cybernetics *17* (1981), 815—824.

[4] V. V. Baranov: A recursive algorithm in markovian decision processes. Cybernetics *18* (1982), 499—506.

[5] D. P. Bertsekas and S. E. Shreve: Stochastic Optimal Control — The Discrete Time Case. Academic Press, New York 1978.

[6] A. Federgruen and P. J. Schweitzer: Nonstationary Markov decision problems with converging parameters. J. Optim. Theory Appl. *34* (1981), 207—241.
[7] A. Federgruen and H. C. Tijms: The optimality equation in average cost denumerable state semi-Markov decision problems, recurrency conditions and algorithms. J. Appl. Probab. *15* (1978), 356—373.
[8] P. J. Georgin: Contrôle de chaînes de Markov sur des espaces arbitraires. Ann. Inst. H. Poincaré *B 14* (1978), 255—277.
[9] J. P. Georgin: Estimation et contrôle de chaînes de Markov sur des espaces arbitraires. In: Lecture Notes Mathematics 636. Springer-Verlag, Berlin—Heidelberg—New York—Tokyo 1978, pp. 71—113.
[10] E. I. Gordienko: Adaptive strategies for certain classes of controlled Markov processes. Theory Probab. Appl. *29* (1985), 504—518.
[11] L. G. Gubenko and E. S. Statland: On controlled, discrete-time Markov decision processes. Theory Probab. Math. Statist. *7* (1975), 47—61.
[12] O. Hernández-Lerma: Approximation and adaptive policies in discounted dynamic programming. Bol. Soc. Mat. Mexicana *30* (1985). In press.
[13] O. Hernández-Lerma: Nonstationary value-iteration and adaptive control of discounted semi-Markov processes. J. Math. Anal. Appl. *112* (1985), 435—445.
[14] O. Hernández-Lerma and S. I. Marcus: Adaptive control of service in queueing systems. Syst. Control Lett. *3* (1983), 283—289.
[15] O. Hernández-Lerma and S. I. Marcus: Optimal adaptive control of priority assignment in queueing systems. Syst. Control Lett. *4* (1984), 65—75.
[16] O. Hernández-Lerma and S. I. Marcus: Adaptive policies for discrete-time stochastic control systems with unknown disturbance distribution. Submitted for publication, 1986.
[17] O. Hernández-Lerma and S. I. Marcus: Nonparametric adaptive control of discrete-time partially observable stochastic systems. Submitted for publication, 1986.
[18] C. J. Himmelberg, T. Parthasarathy and F. S. Van Vleck: Optimal plans for dynamic programming problems. Math. Oper. Res. *1* (1976), 390—394.
[19] K. Hinderer: Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter. (Lecture Notes in Operations Research and Mathematical Systems 33.) Springer-Verlag, Berlin—Heidelberg—New York 1970.
[20] A. Hordijk, P. J. Schweitzer and H. Tijms: The asymptotic behaviour of the minimal total expected cost for the denumerable state Markov decision model. J. Appl. Probab. *12* (1975), 298—305.
[21] P. R. Kumar: A survey of some results in stochastic adaptive control. SIAM J. Control Optim. *23* (1985), 329—380.
[22] M. Kurano: Discrete-time markovian decision processes with an unknown parameter — average return criterion. J. Oper. Res. Soc. Japan *15* (1972), 67—76.
[23] M. Kurano: Average-optimal adaptive policies in semi-Markov decision processes including an unknown parameter. J. Oper. Res. Soc. Japan *28* (1985), 252—366.
[24] P. Mandl: Estimation and control in Markov chains. Adv. Appl. Probab. *6* (1974), 40—60.
[25] P. Mandl: On the adaptive control of countable Markov chains: In: Probability Theory, Banach Centre Publications 5, PWB-Polish Scientific Publishers, Warsaw 1979, pp. 159—173.
[26] H. L. Royden: Real Analysis. Macmillan, New York 1968.
[27] M. Schäl: 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.
[28] M. Schäl: Estimation and control in discounted stochastic dynamic programming. Preprint No. 428, Institute for Applied Math., University of Bonn, Bonn 1981.
[29] H. C. Tijms: On dynamic programming with arbitrary state space, compact action space

and the average reward as criterion. Report BW 55/75, Mathematisch Centrum, Amsterdam 1975.
[30] T. Ueno: Some limit theorems for temporally discrete Markov processes. J. Fac. Science, University of Tokyo *7* (1957), 449—462.
[31] D. J. White: Dynamic programming, Markov chains, and the method of successive approximations. J. Math. Anal. Appl. *6* (1963), 373—376.
[32] P. Mandl and G. Hübner: Transient phenomena and self-optimizing control of Markov chains. Acta Universitatis Carolinae — Math. et Phys. *26* (1985), 1, 35—51.
[33] A. Hordijk and H. Tijms: A modified form of the iterative method of dynamic programming. Ann. Statist. *3* (1975), 1, 203—208.

Dr. Onésimo Hernández-Lerma, Department of Mathematics, Centro de Investigación del I.P.N., Apartado Postal 14-740, México, D. F. 07000. Mexico.