KYBERNETIKA — VOLUME 17 (1981), NUMBER 6

ON THE EXISTENCE OF STATIONARY OPTIMAL POLICIES IN DISCRETE DYNAMIC PROGRAMMING

KAREL SLADKÝ

The paper deals with cumulative optimality criteria in multiplicative Markov decision chains. Polynomial bounds on functional equations of discrete dynamic programming obtained in [Sladký (1981)] are employed to establish a family of optimality criteria for cumulative rewards having a nice property that a stationary optimal policy exists.

1. INTRODUCTION

The paper is devoted to multiplicative Markov decision chains that can be roughly defined as classical Markov decision chains where transition probability matrices are replaced by general nonnegative matrices. Our analysis will be based on the results for functional equations of discrete dynamic programming obtained in [8] and, especially, in [9].

In the remainder of this section we introduce notions and notations used throughout the further text.

We are concerned with a dynamic system which at each time point $t = 0, 1, \ldots$ is observed and classified into a finite number of states from $I = \{1, 2, \ldots, N\}$. If at time t the system is found to be in state i, then a decision, say $f(i)$, must be chosen from a finite set $F(i)$ and, consequently, two things happen:

(i) A reward $r(i, f(i))$ is received, and

(ii) the systems moves into state j at time $t + 1$ with given transition rate $q(i, j; f(i))$ depending only on $i, j \in I$ and selected decision $f(i) \in F(i)$. We only assume that $q(i, j; f(i)) \geqq 0$ and $\sum_{j \in I} q(i, j; \cdot) > 0$ for any $i, j \in I$, $f(i) \in F(i)$; however, $q(i, j; \cdot)$'s need not be probabilities and so we do not necessarily assume $\sum_{j \in I} q(i, j; \cdot) \leqq 1$.

Let $F \equiv \mathop{\mathsf{X}}_{i=1}^{N} F(i)$ and the elements of F, denoted by generic f, be called decision vectors. Obviously, f is an N-vector whose i-th component $f(i) \in F(i)$ specifies the

decision in state i. Observe that F possesses so called "product property"; i.e. if $f_1, f_2 \in F$ then for any $i_1, i_2 \in I$ there exists $f \in F$ such that $f_1(i_1) = f(i_1)$, $f_2(i_2) = f(i_2)$. For what follows it will be convenient to introduce transition rate matrices $Q(f)$ ($N \times N$ matrices depending on $f \in F$) whose ij-th element $[Q(f)]_{ij} = q(i, j; f(i))$ and define similarly reward vectors $r(f)$ (N-column vectors) by $[r(f)]_i = r(i, f(i))$.

Remember that matrices, resp. (column) vectors, will be usually denoted by capital, resp. small, letters. Writing a matrix relation, symbol I, resp. e, is reserved for a unit matrix, resp. unit vector, being of an appropriate dimension; similarly, 0 is also used for a zero matrix of an appropriate dimension. For matrix C, C' denotes the transpose of C; $[C]_i$, resp. $[C]_{ij}$, is reserved for the i-th row, resp. ij-th element, of C. We write $C > B$, resp. $C \gg B$, iff each element of $C - B$ is nonnegative, resp. positive, and $C \neq B$. Similarly, $C \geqq B$ if either $C > B$ or $C = B$. We say that C is lexicographically greater than B (and write $C \succ B$) iff the first non-zero element of each row of $C - B$ is positive and $C \neq B$. Similarly, $C \succeqq B$ if either $C \succ B$ or $C = B$.

A (Markovian) policy controlling the system, say π, is identified by a sequence of decision vectors $\{f_t, t = 0, 1, \ldots\}$. So $\pi \equiv (f_t)$, where for each $i \in I$, $t = 0, 1, \ldots$ $f_t(i)$ specifies the decision at time t if the system is found to be in state i. A policy π is called stationary if $f_t \equiv f$ (i.e. the selected decision depends only on the current state).

To show usefulness of the presented sequential decision model, let us mention at least two of its possible applications:

(I) *Markov decision chains with multiplicative utility functions* (cf. [4], [6]).

Let us consider a Markov decision chain, i.e. our model where transition rates $q(i, j; \cdot)$ are replaced by transition probabilities $p(i, j; \cdot)$ and $\bar{r}(i, \cdot)$ denotes one-stage reward received in state i. Let (supposing policy $\pi \equiv (f_t)$ is followed and the system starts in state $i \in I$) $\bar{r}_t$ be the (random) reward earned at time point t. The utility $U(\ldots)$ of the sequence of rewards $\bar{r}_0, \bar{r}_1, \ldots, \bar{r}_T$ is assumed to be stationary and multiplicative, i.e. $U(r_0, r_1, \ldots, r_T) = \prod_{t=0}^{T} u(\bar{r}_t)$, where $u(\cdot)$ is a given positive function. Setting $q(i, j; \cdot) = p(i, j; \cdot)\, u(\bar{r}(i, \cdot))$ and $r(i, \cdot) = u(\bar{r}(i, \cdot))$, by an easy calculation we get that the vector of expected utilities up to time T, when policy $\pi \equiv (f_t)$ is followed, is given by

$$v^{(0)}(\pi; T) = Q(f_0)\, Q(f_1) \ldots Q(f_{T-1})\, r(f_T) \tag{1.0}$$

(observe that the i-th component denotes the expected utility if the system starts in state $i \in I$).

(II) *Controlled branching processes* (cf. [5], [6]).

Let us consider a controlled branching process with N types of individuals and $F(i)$ possible treatments to type i. Let $q(i, j; f(i))$ be the expected number of individuals

of type j arising from one individual of type i which is subject to treatment $f(i) \in F(i)$. Then the expected population of type j after T generations starting from one individual of type i when policy $\pi \equiv (f_t)$ is followed is obviously given by $[Q(f_0)\, Q(f_1) \ldots \ldots Q(f_{T-1})]_{ij}$. Denoting by $r(j, f(j))$ the terminal value of each individual of type j subject to treatment $f(j) \in F(j)$, then $v^{(0)}(\pi; T)$ (given by (1.0)) is the vector of values of the population after T generations when a treatment policy $\pi \equiv (f_t)$ is followed (notice that the i-th component of $v^{(0)}(\pi; t)$ denotes the value obtained from one individual of type i).

Supposing policy $\pi \equiv (f_t)$ is followed, we introduce for each time interval $\langle t_1, t_2 \rangle$ (where $t_1, t_2 = 0, 1, \ldots$) a vector of l-order cumulative rewards (for $l = 0, 1, \ldots$) denoted by $v^{(l)}(\pi; t_1, t_2)$. These values are defined recursively by

$$v^{(l+1)}(\pi; t_1, t_2) = \sum_{t=t_1}^{t_2} v^{(l)}(\pi; t_1, t) \tag{1.1}$$

where

$$v^{(0)}(\pi; t_1, t_2) = Q(f_{t_1})\, Q(f_{t_1+1}) \ldots Q(f_{t_2-1})\, r(f_{t_2}) \tag{1.1'}$$

and $v^{(0)}(\pi; t_1, t_1) = r(f_{t_1})$. From (1.1), (1.1') we get for any $l \geqq 1$ (cf. Remark 1.1)

$$v^{(l)}(\pi; t_1, t_2) = \sum_{t=t_1}^{t_2} \binom{t_2 - t + l - 1}{l - 1} v^{(0)}(\pi; t_1, t) \tag{1.2}$$

and, in particular, for $l = 1$ we have

$$v^{(1)}(\pi; t_1, t_2) = \sum_{t=t_1}^{t_2} Q(f_{t_1}) \ldots Q(f_{t-1})\, r(f_t)\,. \tag{1.2'}$$

Moreover, on abbreviating $v^{(l)}(\pi; 0, t_2)$ by $v^{(l)}(\pi; t_2)$, (1.2) can be written as

$$v^{(l)}(\pi; T) = \sum_{t=0}^{T} \binom{T - t + l - 1}{l - 1} Q(f_0) \ldots Q(f_{t-1})\, r(f_t)\,. \tag{1.3}$$

Remark 1.1. (1.2) can be easily verified by induction on l. For $l = 1$ (1.2) holds trivially; the induction step is immediate as $\left(\text{recall that } \sum_{m=0}^{t} \binom{m + l}{l} = \binom{t + l + 1}{l + 1}\right)$

$$v^{(l+1)}(\pi; t_1, t_2) = \sum_{t=t_1}^{t_2} \sum_{\tau=t_1}^{t} \binom{t - \tau + l - 1}{l - 1} v^{(0)}(\pi; t_1, \tau) =$$

$$= \sum_{\tau=t_1}^{t_2} v^{(0)}(\pi; t_1, \tau) \sum_{t=\tau}^{t_2} \binom{t - \tau + l - 1}{l - 1} = \sum_{\tau=t_1}^{t_2} \binom{t_2 - \tau + l}{l} v^{(0)}(\pi; t_1, \tau)\,.$$

Remark 1.2. If all $Q(f)$'s are (sub)-stochastic matrices $v^{(l)}(\pi; t_1, t_2)$ can be interpreted as vectors of expected values of "weighted" one-stage rewards. In particular, supposing that the system was found at time t_1 in state $j \in I$, then $[v^{(0)}(\pi; t_1, t_2)]_j$ specifies the expected reward to be earned at time t_2, $[v^{(1)}(\pi; t_1, t_2)]_j$ denotes the value

of total expected rewards to be earned in the time interval $\langle t_1, t_2\rangle$ and for $l > 1$ $[v^{(l)}(\pi; t_1, t_2)]_j$ corresponds (cf. (1.2)) to "weighted" total expected rewards to be earned in the time interval $\langle t_1, t_2\rangle$.

Recall that in dynamic programming models $v^{(1)}(\pi; 0, t) \equiv v^{(1)}(\pi; t)$ (for $t \to \infty$) is usually considered for evaluating the "quality" of policy $\pi \equiv (f_t)$; in virtue of (1.1), (1.2) optimality criteria based on $v^{(l)}(\pi; t) \equiv v^{(l)}(\pi; 0, t)$ for $l > 1$ (originally introduced by Veinott in [10] and further studied e.g. in [2, 3, 6, 7]), though it might have interesting economic interpretations, seem only to be a natural generalization of the standard "quality measure" of a policy $\pi \equiv (f_t)$.

For what follows, we shall need also the opposite time orientation for the considered model. To this order observe that by (1.2), (1.1′) for $t_2 \geqq t_1$ we get (we set $v^{(l)}(\pi; t, \tau) = 0$ if $\tau < t$)

$$v^{(l)}(\pi; t_1, t_2) = Q(f_{t_1})\, v^{(l)}(\pi; t_1 + 1, t_2) + \binom{t_2 - t_1 + l - 1}{l - 1} r(f_{t_1}). \tag{1.4}$$

Now fixing t_2, introducing "backward time" orientation with $n = t_2 - t_1$, and setting $v^{(l)}(\pi; t_1, t_2) = v^{(l)}(n + 1; \pi)$, $f_{t_1} = f^{(n)}$, (1.4) reads

$$v^{(l)}(n + 1; \pi) = Q(f^{(n)})\, v^{(l)}(n; \pi) + \binom{n + l - 1}{l - 1} r(f^{(n)}) \tag{1.5}$$

where $v^{(l)}(0; \pi) = 0$. Iterating (1.5) we immediately get

$$v^{(l)}(n + 1; \pi) = \sum_{m=0}^{n} \binom{n - m + l - 1}{l - 1} Q(f^{(n)}) \dots Q(f^{(n-m+1)})\, r(f^{(n-m)}) \tag{1.5′}$$

and, in particular, for π stationary, i.e. $\pi \equiv (f)$, instead of $v^{(l)}(n; \pi)$ we write $v^{(l)}(n; f)$; so by (1.5), (1.5′) and (1.3) we get

$$v^{(l)}(n + 1; f) = Q(f)\, v(n; f) + \binom{n + l - 1}{l - 1} r(f) = \tag{1.5*}$$

$$= \sum_{m=0}^{n} \binom{n - m + l - 1}{l - 1} (Q(f))^{n-m}\, r(f) = v^{(l)}(\pi; n).$$

Similarly, denoting (for t_2 fixed) $v^{(l)}(n) = \max_{\pi} v^{(l)}(n; \pi)$ then, obviously, $\{v^{(l)}(n), n = 0, 1, \dots\}$ satisfy the following dynamic programming recursion

$$v^{(l)}(n + 1) = \max_{f \in F} \left[Q(f)\, v^{(l)}(n) + \binom{n + l - 1}{l - 1} r(f) \right] = \tag{1.6}$$

$$= Q(\hat{f}^{(n)})\, v^{(l)}(n) + \binom{n + l - 1}{l - 1} r(\hat{f}^{(n)})$$

with $v^{(l)}(0) = 0$.

It is easy to verify (cf. Example I of Section 1 in [8]) that (1.6) can be written in a more compact form. Introducing an $(N + l) \times (N + l)$ matrix $M(f)$ by

$$M(f) = \begin{bmatrix} Q(f) & r(f) & 0 \\ 0 & 1 & e' \\ 0 & 0 & J \end{bmatrix} \tag{1.7}$$

with J being a $(l - 1) \times (l - 1)$ upper triangular matrix whose each entry on or above the diagonal equals 1 (consequently, also the dimension of the unit row vector e' equals $l - 1$), then by an elementary matrix calculation we get for $(N + l)$-column vector $z(n)$ defined recursively by

$$z(n + 1) = \max_{f \in F} M(f)\, z(n) = M(\hat{f}^{(n)})\, z(n) \tag{1.8}$$

with

$$z(0) = [\underbrace{0, 0, \ldots, 0}_{N}; \underbrace{1, 1, \ldots, 1}_{l}]' \tag{1.8'}$$

that

$$z(n) = \left[(v^{(l)}(n))', \binom{n + l - 1}{l - 1}, \ldots, \binom{n}{0}\right]'. \tag{1.9}$$

2. PRELIMINARIES

In this section we summarize some useful facts from the theory of nonnegative matrices together with some properties of discrete dynamic programming recursions obtained in [8] and [9].

Let us consider the set of nonnegative matrices $\{Q(f), f \in F\}$. According to the well-known Perron-Frobenius theorem $\sigma(f)$ (spectral radius of $Q(f)$) equals to the largest positive eigenvalue of $Q(f)$ and we can choose the corresponding eigenvector $u(f) > 0$. Recall that if $Q(f)$ is irreducible then even $u(f) \gg 0$, and $\sigma(f)$ is simple. Moreover, if $Q(f)$ is reducible, i.e., if by suitably permuting rows and corresponding columns of $Q(f)$ is is possible to write

$$Q(f) = \begin{bmatrix} Q_{(11)}(f) & Q_{(12)}(f) & \cdots & Q_{(1r)}(f) \\ 0 & Q_{(22)}(f) & \cdots & Q_{(2r)}(f) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_{(rr)}(f) \end{bmatrix} \tag{2.1}$$

where each $Q_{(ii)}(f)$ itself is an irreducible matrix having spectral radius $\sigma_{(i)}(f)$, necessary and sufficient conditions for $u(f) \gg 0$ can be easily formulated by means of accessibility between irreducible classes of $Q(f)$ (accessibility is defined in the same way as in Markov chain theory, see [8] for details). Recalling that $Q_{(ii)}(f)$ is called basic, resp. non-basic, class of $Q(f)$ iff $\sigma_{(i)}(f) = \sigma(f)$, resp. $\sigma_{(i)}(f) < \sigma(f)$, it holds

(cf. [8] or Theorem 7 of Chapter 13 in [1]): $u(f) \gg 0$ if and only if each non-basic, resp. basic, class of $Q(f)$ is accessible to some basic class, resp. is not accessible to any other irreducible class, of $Q(f)$.

In virtue of these facts we can establish that each $Q(f)$ can be also decomposed into (another) block-triangular form whose (generally reducible) diagonal submatrices are the "largest" possible submatrices of $Q(f)$ having strictly positive eigenvectors. More precisely, it can be shown (cf. Lemma 2.1 of [8]) that for any $Q(f)$, by possibly permuting rows and corresponding columns, we can write

$$Q(f) = \begin{bmatrix} Q_{11}(f) & Q_{12}(f) & \dots & Q_{1s}(f) \\ 0 & Q_{22}(f) & \dots & Q_{2s}(f) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Q_{ss}(f) \end{bmatrix} \tag{2.2}$$

where for each $i = 1, 2, \dots, s \equiv s(f)$

$$Q_{ii}(f)\, u_i(f) = \sigma_i(f)\, u_i(f) \tag{2.2'}$$

with $\sigma_i(f)$, resp. $u_i(f) \gg 0$, being the spectral radius, resp. corresponding right eigenvector, of $Q_{ii}(f)$ (in general reducible), and

$$(\sigma_1(f), \nu_1(f)) \succ (\sigma_2(f), \nu_2(f)) \succ \dots \succ (\sigma_s(f), \nu_s(f)). \tag{2.2''}$$

Here $\nu_i(f)$ is the index of $Q_{ii}(f)$, defined as a number of irreducible classes with spectral radius $\sigma_i(f)$ that are successively accessible from $Q_{ii}(f)$. Observe that

$$\sigma_i(f) = \sigma_{i+1}(f) \quad \Rightarrow \quad \nu_i(f) = \nu_{i+1}(f) + 1$$

$$\sigma_i(f) > \sigma_{i+1}(f) \quad \Rightarrow \quad \nu_i(f) = 1\,, \quad \text{and} \quad \nu_s(f) = 1\,.$$

Of course, the decomposition according to (2.2) determines also a similar decomposition of the state space I. For what follows it will be useful to label the elements of $Q_{ii}(f)$ by integers from $I_i(f)$ where

$$I = \bigcup_{i=1}^{s(f)} I_i(f) \quad \text{and} \quad I_i(f) \cap I_k(f) = \emptyset \quad \text{for any} \quad i \neq k\,.$$

Moreover, in [8] we have established, under the assumption that $\sigma_s(f) > 0$ for any $f \in F$, that a similar block-triangular decomposition is even possible with respect to the whole set $\{Q(f), f \in F\}$. These results (cf. [8], Theorem 3.2) can be summarized as:

Proposition 1. There exists $\hat{f} \in F$ (that can be found by a finite policy iteration method) such that the block-triangular decomposition according to (2.2), (2.2'), (2.2'') of matrix $Q(f)$ remains block-triangular for any $Q(f)$ with $f \in F$; i.e., denoting for any $f \in F$ by $Q_{ii}(f)$ submatrix of $Q(f)$ whose elements are labelled by integers from $I_i(\hat{f})$, then for any $i = 1, 2, \dots, s = s(\hat{f})$

$$Q_{ik}(f) = 0 \quad \text{for any} \quad k < i \tag{2.3}$$

together with

$$Q_{ii}(f)\,u_i \leqq Q_{ii}(\hat{f})\,u_i = \sigma_i u_i \tag{2.4}$$

where $\sigma_i \equiv \sigma_i(\hat{f}) > 0$ and $u_i \equiv u_i(f) \gg 0$ (similarly we abbreviate $v_i(\hat{f})$ by v_i).

Having introduced the uniform decomposition of the whole set $\{Q(f), f \in F\}$, determining also the partition of the state space I into classes $I_i(\hat{f})$ (where $i = 1, 2, \ldots, s \equiv s(\hat{f})$), and recalling that an irreducible class of $Q_{ii}(f)$ is called basic class of $Q_{ii}(f)$ iff its spectral radius equals to $\sigma_i(f)$, we further denote (cf. [8])

$$\bar{I}_i(\hat{f}; f) = \{j \in I_i(\hat{f}) : j \text{ belongs to some basic class of } Q_{ii}(f)\}$$

$$I_i^{\mathrm{R}}(\hat{f}) = \{j \in I_i(\hat{f}) : \exists f \in F \text{ with } \sigma_i(f) = \sigma_i \text{ such that } j \in \bar{I}_i(\hat{f}; f)\}$$

$$I_i^{\mathrm{T}}(\hat{f}) = I_i(\hat{f}) \setminus I_i^{\mathrm{R}}(\hat{f})\,.$$

Observe (cf. Remark 3.5 of [8]) that there need not exist any $f \in F$ such that $I_i^{\mathrm{R}}(\hat{f}) = \bar{I}_i(\hat{f}; f)$.

Now let us recall some properties of vector sequences $\{x(n; \pi),\ n = 0, 1, \ldots\}$ defined recursively by

$$x(n+1; \pi) = Q(f^{(n)})\,x(n; \pi) \quad \text{with} \quad x(0; \pi) \equiv x(0) > 0\,, \tag{2.5}$$

where policy π is identified in the "backward time" orientation by $\pi \equiv (\ldots, f^{(n)}, \ldots, f^{(1)}, f^{(0)})$. Obviously, $x(n; \pi) \geqq 0$ and also $x(n; \pi) \leqq x(n)$, where $x(n)$ obeys the dynamic programming recursion

$$x(n+1) = \max_{f \in F} Q(f)\,x(n) = Q(\hat{f}^{(n)})\,x(n) \quad \text{with} \quad x(0) > 0\,. \tag{2.5*}$$

In particular, for π stationary, i.e. $\pi \equiv (f)$, instead of $x(n; \pi)$ we simply write $x(n; f)$.

Employing the uniform block-triangular decomposition of $\{Q(f), f \in F\}$ introduced in Proposition 1, some upper bounds on $x(n; \pi)$ can be established (cf. Theorem 4.1 and Corollary 4.2 of [8]). Moreover, as it was indicated in [9], these bounds remain valid even in the case of not necessarily nonnegative matrices if the "uniform" block-triangular decomposition of $\{Q(f), f \in F\}$ is possible. So (for the rest of this section) we make:

Assumption A. There exists $\hat{f} \in F$, defining the partition of the state space I into classes $I_i(\hat{f})$, and the respective decomposition of any $Q(f)$, such that (2.2), (2.2′), (2.2″) hold for $f = \hat{f}$ and also (2.3), (2.4) are valid with $Q_{ii}(f) > 0$, but $Q_{ij}(f)$ not necessarily nonnegative for $j > i$.

Denoting by $x_i(n; \pi)$ subvector of $x(n; \pi)$ (and similarly by w_i subvector of w) whose elements are labelled by integers from $I_i(\hat{f})$ and introducing $\bar{x}_i(n; \pi) = \sigma_i^{-n}\,x_i(n; \pi)$, $\bar{x}_i(n) = \sigma_i^{-n}\,x_i(n)$ it holds:

Proposition 2. Let Assumption A be fulfilled. Then $u_i \gg 0$ in (2.4) can be selected such that for all $n = 0, 1, \ldots$

$$\bar{x}_i(n) \leqq \binom{n + v_i - 1}{v_i - 1} u_i \,. \tag{2.6}$$

In particular, if $s(\hat{f}) = 1$ and $u_1 \equiv u \geqq x(0)$, $\sigma \equiv \sigma(\hat{f})$ then

$$x(n) \leqq \sigma^n u \,. \tag{2.6'}$$

However, if in (2.6) some $v_i > 1$ and for $j = i + 1, \ldots, i + v_i - 1$ asymptotic behaviour of $\bar{x}_j(n)$ is known, some (finer) polynomial bounds on $\bar{x}_i(n)$ can be established. This topic was investigated in [9] (cf. Theorem 3.1 and Corollary 3.4 of [9]) and some useful results of this study are summarized in Proposition 3 (observe that $\sigma_i = \sigma_p \Rightarrow v_i - v_p = p - i$ and $v_i > 1 \Rightarrow \sigma_j = \sigma_i$ for $j = i, \ldots, i + v_i - 1$ with $\sigma_{i+v_i} < \sigma_i$).

Proposition 3. Let Assumption A hold, $v_{i_0} > 1$ for some $i_0 = 1, 2, \ldots, s \equiv s(\hat{f})$, and let for any $j = p, \ldots, i_0 + v_{i_0} - 1$ (with $p > i_0$) there exist vectors $w_j^{(k)}$ (with $k = 0, \ldots, v_j - 1$) such that

$$\lim_{n\to\infty} \left[\bar{x}_j(n) - \sum_{k=0}^{v_j - 1} \binom{n}{k} w_j^{(k)} \right] = 0 \,. \tag{2.7}$$

Then for any $i = i_0, \ldots, p - 1$ it is possible to compute (by a policy iteration algorithm) vectors $w_i^{(k)}$ (with $k = v_i - 1, \ldots, 1, 0$) such that

$$\lim_{n\to\infty} \binom{n}{p - i}^{-1} \left[\bar{x}_i(n) - \sum_{k=0}^{v_i - 1} \binom{n}{k} w_i^{(k)} \right] = 0 \,. \tag{2.8}$$

Moreover, if there exists some $\tilde{f} \in F$ fulfilling

$$\lim_{n\to\infty} [\bar{x}_j(n) - \bar{x}_j(n; \tilde{f})] = 0 \quad \text{for any} \quad j = p, \ldots, i_0 + v_{i_0} - 1 \tag{2.9}$$

then we can select $f^* \in F$ (with $f^*(k) = \hat{f}(k)$ for any $k \in I_m(\hat{f})$ with $m > p$) such that for any $i = i_0, \ldots, p - 1$

$$\lim_{n\to\infty} \binom{n}{p - i}^{-1} \left[\bar{x}_i(n; f^*) - \sum_{k=0}^{v_i - 1} \binom{n}{k} w_i^{(k)} \right] = 0 \tag{2.10}$$

and, consequently, also

$$\lim_{n\to\infty} \binom{n}{p - i}^{-1} [\bar{x}_i(n) - \bar{x}_i(n; f^*)] = 0 \,. \tag{2.11}$$

The following Proposition 4 shows that condition (2.7) of Proposition 3 is always fulfilled if we assume that (2.7) holds only for the components labelled from $I_j^R(\hat{f})$

(instead of all $I_j(\hat{f})$). To this order let us fix some $p = 1, \ldots, s \equiv s(\hat{f})$ and set (for any $f \in F$ and integer $m = p, \ldots, s$)

$$Q_{pp}(f) = \begin{bmatrix} Q_{pp}^{\mathrm{TT}}(f) & Q_{pp}^{\mathrm{TR}}(f) \\ Q_{pp}^{\mathrm{RT}}(f) & Q_{pp}^{\mathrm{RR}}(f) \end{bmatrix}, \quad Q_{pm}(f) = \begin{bmatrix} Q_{pm}^{\mathrm{T}}(f) \\ Q_{pm}^{\mathrm{R}}(f) \end{bmatrix}; \tag{2.12}$$

$$x_p(n) = \begin{bmatrix} x_{p\mathrm{T}}(n) \\ x_{p\mathrm{R}}(n) \end{bmatrix}; \quad w_p^{(k)} = \begin{bmatrix} w_{p\mathrm{T}}^{(k)} \\ w_{p\mathrm{R}}^{(k)} \end{bmatrix}, \quad u_p = \begin{bmatrix} u_{p\mathrm{T}} \\ u_{p\mathrm{R}} \end{bmatrix}$$

where the elements of $Q_{pp}^{\mathrm{RR}}(f)$, $x_{p\mathrm{R}}(n)$, $w_{p\mathrm{R}}^{(k)}$ and the rows of $Q_{pm}^{\mathrm{R}}(f)$ (resp. $Q_{pp}^{\mathrm{TT}}(f)$, $x_{p\mathrm{T}}(n)$, $w_{p\mathrm{T}}^{(k)}$ and $Q_{pm}^{\mathrm{T}}(f)$) are labelled by integers from $I_p^{\mathrm{R}}(\hat{f})$ (resp. $I_p^{\mathrm{T}}(\hat{f}) = I_p(\hat{f}) \setminus I_p^{\mathrm{R}}(\hat{f})$). Remember that $\bar{x}_{p\mathrm{T}}(n) = \sigma_p^{-n} x_{p\mathrm{T}}(n)$ and $\bar{Q}_{pm}(f) = \sigma_p^{-1} Q_{pm}(f)$. Observe that by (2.4) and Perron-Frobenius theorem we get for any $f \in F$

$$Q_{pp}^{\mathrm{TT}}(f)\, u_{p\mathrm{T}} < \sigma_p u_{p\mathrm{T}} \tag{2.13}$$

and recall (cf. definition of $I_i^{\mathrm{T}}(\hat{f})$ or (2.13)) that the spectral radius of each $Q_{pp}^{\mathrm{TT}}(f)$ (for $p = 1, 2, \ldots, s = s(\hat{f})$) is less than σ_p.

It can be easily recognized that a very special case of Proposition 4 (with $p = s = 1$ and $\sigma_p = 1$, $Q_{pp}^{\mathrm{RR}}(f) = 1$, $x_{p\mathrm{R}}(n) \equiv w_{p\mathrm{R}}^{(0)} = 1$) was treated in the dynamic programming literature in connection with so called transient dynamic programming that extends well-known discounted dynamic programming models (cf. [11]).

Proposition 4. Let for some $p = 1, 2, \ldots, s \equiv s(\hat{f})$ there exist $w_{p\mathrm{R}}^{(k)}$ (for $k = 0, 1, \ldots, v_p - 1$) such that

$$\lim_{n \to \infty} \left[\bar{x}_{p\mathrm{R}}(n) - \sum_{k=0}^{v_p - 1} \binom{n}{k} w_{p\mathrm{R}}^{(k)} \right] = 0 \tag{2.14}$$

and, moreover, if $v_p > 1$ let $w_j^{(k)}$ (for $j = p + 1, \ldots, p + v_p - 1$; $k = 0, \ldots, v_j - 1$) be selected such that for any $j = p + 1, \ldots, p + v_p - 1$

$$\lim_{n \to \infty} \left[\bar{x}_j(n) - \sum_{k=0}^{v_j - 1} \binom{n}{k} w_j^{(k)} \right] = 0\,. \tag{2.14'}$$

Then (by a policy iteration algorithm) it is possible to construct $w_{p\mathrm{T}}^{(k)}$ (for $k = 0, 1, \ldots, v_p - 1$) such that

$$\lim_{n \to \infty} \left[\bar{x}_{p\mathrm{T}}(n) - \sum_{k=0}^{v_p - 1} \binom{n}{k} w_{p\mathrm{T}}^{(k)} \right] = 0\,. \tag{2.15}$$

Furthermore, if there exists some $\tilde{f} \in F$ fulfilling

$$\lim_{n \to \infty} [\bar{x}_{p\mathrm{R}}(n) - \bar{x}_{p\mathrm{R}}(n; \tilde{f})] = 0 \tag{2.16}$$

and if $v_p > 1$ also

$$\lim_{n \to \infty} [\bar{x}_j(n) - \bar{x}_j(n; \tilde{f})] = 0 \quad \text{for all} \quad j = p + 1, \ldots, p + v_p - 1\,, \tag{2.16'}$$

then it is possible to select $f^* \in F$ (with $f^*(k) = \hat{f}(k)$ for any $k \in I_p^R(\hat{f})$ and any $k \in \in I_m(\hat{f})$ with $m > p$) such that

$$\lim_{n\to\infty} [\bar{x}_{pT}(n) - \bar{x}_{pT}(n; f^*)] = 0 . \tag{2.17}$$

To establish Proposition 4 we shall need Lemma 2.1 (very similar to Theorem 2.1 of [9]) together with Lemma 2.2. To simplify the notations, in Lemmas 2.1, 2.2 we shall delete indices p, T (so we write $Q(f)$, $u(f)$, $c^{(k)}(f)$ instead of $Q_{pp}^{TT}(f)$, $u_{pT}^{(k)}(f)$, $c_{pT}^{(k)}(f)$, respectively). However, the presented formulation of Lemma 2.1 well corresponds to the usual denotations in controlled Markov chains. In particular, if $v = 0$ and $Q(f) = \alpha P(f)$ with $P(f)$ stochastic and $\alpha \in (0, 1)$, the proof of Lemma 2.1 reduces to the well-known policy iteration algorithm for discounted Markov decision chains.

Lemma 2.1. Let for any $f \in F$, $Q(f) > 0$ with $\sigma(f) < 1$ and let $c^{(k)}(f)$ (for $k = = v, \ldots, 1, 0$) be given vectors. Then there exists vectors $u^{(k)}$ (with $k = v, \ldots, 1, 0$) and a nonincreasing sequence of (non-empty) sets of decision vectors $F \equiv \tilde{F}^{(v+1)} \supset \supset \tilde{F}^{(v)} \supset \ldots \supset \tilde{F}^{(1)} \supset \tilde{F}^{(0)} \neq \emptyset$ possessing the following property:
Denoting for $k = v, \ldots, 1, 0$ (we set $u^{(v+1)} = 0$)

$$\tilde{\psi}^{(k)}(f) = (Q(f) - I)u^{(k)} - u^{(k+1)} + c^{(k)}(f), \tag{2.1.1}$$

then for any $k = v, \ldots, 1, 0$

$$\tilde{\psi}^{(k)}(f) \leqq 0 \quad \text{for any} \quad f \in \tilde{F}^{(k+1)} \tag{2.1.2}$$

where $\tilde{F}^{(k)}$ are defined recursively by

$$\tilde{F}^{(k)} = \{f \in \tilde{F}^{(k+1)} : \tilde{\psi}^{(k)}(f) = 0\} \quad \text{with} \quad \tilde{F}^{(v+1)} \equiv F . \tag{2.1.3}$$

Moreover, there exists $\tilde{f} \in F$ such that

$$\tilde{\psi}^{(k)}(\tilde{f}) = 0 \quad \text{for any } k = v, \ldots, 1, 0 . \tag{2.1.4}$$

Proof. By policy iterations. First observe that (as $\sigma(f) < 1$) $Z(f) = (I - Q(f))^{-1}$ always exists, so $(Q(f) - I) Z(f) = -I$. Then by a direct calculation we can verify that

$$u^{(k)}(f) = -\sum_{j=k}^{v} (-Z(f))^{1+j-k} c^{(j)}(f) = Z(f)[c^{(k)}(f) - u^{(k+1)}(f)] \tag{2.1.5}$$

is a solution to the set of equations (for $k = v, \ldots, 1, 0$; $u^{(v+1)}(f) = 0$)

$$(Q(f) - I) u^{(k)}(f) - u^{(k+1)}(f) + c^{(k)}(f) = 0 . \tag{2.1.6}$$

Moreover, as $(Q(f) - I)$ is nonsingular, we easily conclude that $u^{(k)}(f)$'s given by (2.1.5) are the unique solution of (2.1.6).

Denoting for $f, g \in F$

$$\tilde{\psi}^{(k)}(g; f) = (Q(g) - I) u^{(k)}(f) - u^{(k+1)}(f) + c^{(k)}(g) \tag{2.1.7}$$

then by (2.1.6), (2.1.7) we get

$$(2.1.8)\quad u^{(k+1)}(g) - u^{(k+1)}(f) = (Q(g) - I)\,u^{(k)}(g) + c^{(k)}(g) - u^{(k+1)}(f) = \\ = (Q(g) - I)\,(u^{(k)}(g) - u^{(k)}(f)) + \tilde{\psi}^{(k)}(g; f)\,.$$

As $u^{(\nu+1)}(\cdot) \leqq 0$ and $(I - Q(\cdot))^{-1} > 0$ with at least positive diagonal entries, by (2.1.8) we immediately verify that

$$(2.1.9)\qquad \tilde{\psi}^{(\nu)}(g; f) > 0 \;\Rightarrow\; u^{(\nu)}(g) > u^{(\nu)}(f)\,.$$

Similarly, if $\tilde{\psi}^{(\nu)}(g; f) = 0$, repeating the above reasoning and recalling uniqueness of the solutions to (2.1.6), we conclude that

$$(2.1.9')\quad \tilde{\psi}^{(k)}(g; f) = 0 \;\; (\text{for } k = \nu, \ldots, m+1)\,, \;\; \tilde{\psi}^{(m)}(g; f) > 0 \;\Rightarrow \\ \Rightarrow\; u^{(k)}(g) = u^{(k)}(f) \;\; (\text{for } k = \nu, \ldots, m+1) \;\; \text{and} \;\; u^{(m)}(g) > u^{(m)}(f)\,.$$

To show that (2.1.2), (2.1.4) hold for suitably chosen $u^{(k)}$'s, let us construct a (finite) sequence of decision vectors $f_0, f_1, \ldots, f_r \equiv \tilde{f}$ with f_0 arbitrary and f_{n+1} obtained by the following improvement of f_n:

For given $f_n \in F$ calculate $u^{(k)}(f_n)$ (for $k = \nu, \ldots, 1, 0$) being the solution to (2.1.6) for $f = f_n$ and on the base of $u^{(k)}(f_n)$'s (cf. (2.1.5)) perform the improvement of f_n; i.e. select $f_{n+1} \neq f_n$ (if possible) such that (cf. (2.1.7))

$$(2.1.10)\qquad (\tilde{\psi}^{(\nu)}(f_{n+1}; f_n), \ldots, \tilde{\psi}^{(0)}(f_{n+1}; f_n)) \succ 0\,.$$

So by (2.1.9), (2.1.9')

$$(2.1.11)\quad \tilde{\psi}^{(k)}(f_{n+1}; f_n) = 0 \;\; (\text{for } k = \nu, \ldots, m+1)\,, \;\; \tilde{\psi}^{(m)}(f_{n+1}; f_n) > 0 \;\Rightarrow \\ \Rightarrow\; (u^{(\nu)}(f_{n+1}), \ldots, u^{(m)}(f_{n+1})) \succ (u^{(\nu)}(f_n), \ldots, u^{(m)}(f_n))$$

and, consequently, the elements of $\{f_n\}$ cannot recur. As F is finite, in a finite number of policy improvement steps we obtain $f_r \equiv \tilde{f}$ that cannot be further improved and $u^{(k)} \equiv u^{(k)}(\tilde{f})$ satisfy (2.1.2), (2.1.4). □

Lemma 2.2. Let for any $f \in F$ $Q(f) > 0$ with $\sigma(f) < 1$. Then for an arbitrary policy $\pi \equiv (f^{(k)})$

$$(2.2.1)\qquad \lim_{n\to\infty} \prod_{k=0}^{n-1} Q(f^{(k)}) = 0$$

and this convergence is exponential; i.e., there exists $\varrho \in (0, 1)$ and $u^0 \gg 0$ such that

$$(2.2.1')\qquad \prod_{k=0}^{n-1} Q(f^{(k)})\,u^0 \leqq \varrho^n u^0\,.$$

Moreover, if (for vectors $y(n)$, $h(f; n)$ having only appropriate dimensions) there exist $\hat{f}^{(n)}, f^* \in F$ such that for any $n \geqq n_0$

$$(2.2.2)\qquad y(n+1) \leqq Q(\hat{f}^{(n)})\,y(n) + h(\hat{f}^{(n)}; n)$$

$$(2.2.3)\qquad y(n+1) \geqq Q(f^*)\,y(n) + h(f^*; n)$$

where

$$\lim_{n\to\infty} h(f; n) = 0 \quad \text{for any} \quad f \in F \tag{2.2.4}$$

then also

$$\lim_{n\to\infty} y(n) = 0\,. \tag{2.2.5}$$

Proof. Adding sufficiently small $\varepsilon > 0$ to each element of $Q(f)$ (recall that the spectral radius of $Q(f)$ depends continuously on the elements of $Q(f)$) and applying (2.4) of Proposition 1 to the resulting (irreducible) matrix, we conclude that for an appropriate vector u^0 and number $\varrho \in (0, 1)$

$$Q(f)\,u^0 \leqq \varrho u^0 < u^0 \quad \text{where} \quad u^0 \gg 0\,. \tag{2.2.6}$$

Iterating (2.2.6) we get (2.2.1′) and, consequently, also (2.2.1) (observe that the ij-th element on the RHS of (2.2.1) must be nongreater than $\varrho^n[u^0]_i/[u^0]_j$) Now, by iterating (2.2.2) we get for any $n \geqq n_0$ and $m > 0$

$$y(n+m) \leqq \prod_{k=1}^{m} Q(\hat{f}^{(n+m-k)})\, y(n) + \sum_{l=0}^{m-1} \sum_{k=1}^{m-l-1} Q(\hat{f}^{(n+m-k)})\, h(\hat{f}^{(n+l)}; n+l) \tag{2.2.7}$$

(by our convention $\prod_{k=1}^{0} Q(\hat{f}^{(k)}) = I$). Similarly, by (2.4.3) we conclude that

$$y(n+m) \geqq (Q(f^*))^m\, y(n) + \sum_{l=0}^{m-1} (Q(f^*))^{m-l-1}\, h(f^*; n+l)\,. \tag{2.2.8}$$

Choosing $c(n) < 0$, $d(n) > 0$ such that

$$c(n) \leqq h(f; k) \leqq d(n) \quad \text{for any} \quad f \in F \quad \text{and} \quad k = n, n+1, \ldots, \tag{2.2.9}$$

by (2.2.1), (2.2.1′) we get for $\beta = \max_{i,j} [u^0]_i/[u^0]_j$ and $f^{(n)} = \hat{f}^{(n)}$ or $f^{(n)} = f^*$

$$(1-\varrho)^{-1}\,\beta\, c(n) \leqq \sum_{l=0}^{m-1} \sum_{k=1}^{m-l-1} Q(f^{(n+m-k)})\, h(f^{(n+l)}; n+l) \leqq \tag{2.2.10}$$

$$\leqq (1-\varrho)^{-1}\,\beta\, d(n)\,.$$

However, by (2.2.4) $c(n)$, $d(n)$ in (2.2.9) can be selected such that $c(n) \to 0$, $d(n) \to 0$ for $n \to \infty$; so (2.2.5) follows then immediately by inserting (2.2.10) into (2.2.7) and (2.2.8). □

Now we are in a position to present:

Proof of Proposition 4. Let us denote

$$y_j^{(p)}(n) = \bar{x}_j(n) - \sum_{k=0}^{v_j-1} \binom{n}{l} w_j^{(k)} \quad \text{for } j = p+1, \ldots, p+v_p-1,\ j = p\mathrm{T}, p\mathrm{R}$$

$$y_j^{(p)}(n) = \sigma_p^{-n}\, x_j(n) \qquad \text{for all} \quad j \geqq p + v_p\,.$$

Recalling the matrix decomposition according to (2.2), (2.11), by (2.5*) we get for arbitrary $w_{p\mathrm{T}}^{(k)}$'s:

$$y_{p\mathrm{T}}^{(p)}(n+1) = \bar{x}_{p\mathrm{T}}(n+1) - \sum_{k=0}^{\nu_p-1}\binom{n+1}{k} w_{p\mathrm{T}}^{(k)} = \tag{2.4.1}$$

$$= \bar{Q}_{pp}^{\mathrm{TT}}(\hat{f}^{(n)})\,\bar{x}_{p\mathrm{T}}(n) + \bar{Q}_{pp}^{\mathrm{TR}}(\hat{f}^{(n)})\,\bar{x}_{p\mathrm{R}}(n) + \sum_{j=p+1}^{s}\bar{Q}_{pj}^{\mathrm{T}}(\hat{f}^{(n)})\,(\sigma_p^{-n}\,x_j(n)) -$$

$$-\sum_{k=0}^{\nu_p-1}\binom{n}{k} w_{p\mathrm{T}}^{(k)} - \sum_{k=1}^{\nu_p-1}\binom{n}{k-1} w_{p\mathrm{T}}^{(k)} = \bar{Q}_{pp}^{\mathrm{TT}}(\hat{f}^{(n)})\,y_{p\mathrm{T}}^{(p)}(n) + \bar{Q}_{pp}^{\mathrm{TR}}(\hat{f}^{(n)})\,y_{p\mathrm{R}}^{(p)}(n) +$$

$$+\sum_{j=p+1}^{s}\bar{Q}_{pj}^{\mathrm{T}}(\hat{f}^{(n)})\,y_j^{(p)}(n) + (\bar{Q}_{pp}^{\mathrm{TT}}(\hat{f}^{(n)}) - I)\sum_{k=0}^{\nu_p-1}\binom{n}{k} w_{p\mathrm{T}}^{(k)} +$$

$$+\bar{Q}_{pp}^{\mathrm{TR}}(\hat{f}^{(n)})\sum_{k=0}^{\nu_p-1}\binom{n}{k} w_{p\mathrm{R}}^{(k)} + \sum_{j=p+1}^{p+\nu_p-1}\bar{Q}_{pj}^{\mathrm{T}}(\hat{f}^{(n)})\sum_{k=0}^{\nu_j-1}\binom{n}{k} w_j^{(k)} - \sum_{k=1}^{\nu_p-1}\binom{n}{k-1} w_{p\mathrm{T}}^{(k)}\,.$$

As (recalling that $\nu_j = \nu_p + p - j$ and changing summation)

$$\sum_{j=p+1}^{p+\nu_p-1}\bar{Q}_{pj}^{\mathrm{T}}(\hat{f}^{(n)})\sum_{k=0}^{\nu_j-1}\binom{n}{k} w_j^{(k)} = \sum_{k=0}^{\nu_p-1}\binom{n}{k}\sum_{j=p+1}^{\nu_p+p-k-1}\bar{Q}_{pj}^{\mathrm{T}}(\hat{f}^{(n)})\,w_j^{(k)}$$

by (2.4.1) we immediately get

$$y_{p\mathrm{T}}^{(p)}(n+1) = \bar{Q}_{pp}^{\mathrm{TT}}(\hat{f}^{(n)})\,y_{p\mathrm{T}}^{(p)}(n) + \bar{Q}_{pp}^{\mathrm{TR}}(\hat{f}^{(n)})\,y_{p\mathrm{R}}^{(p)}(n) + \tag{2.4.2}$$

$$+\sum_{j=p+1}^{s}\bar{Q}_{pj}^{\mathrm{T}}(\hat{f}^{(n)})\,y_j^{(p)}(n) + \sum_{k=0}^{\nu_p-1}\binom{n}{k}\psi_{p\mathrm{T}}^{(k)}(\hat{f}^{(n)})\,,$$

where (for $k = 0, 1, \ldots, \nu_p - 1$ with $w_{p\mathrm{T}}^{(\nu_p)} = 0$)

$$\psi_{p\mathrm{T}}^{(k)}(f) = (\bar{Q}_{pp}^{\mathrm{TT}}(f) - I)\,w_{p\mathrm{T}}^{(k)} - w_{p\mathrm{T}}^{(k+1)} + \bar{Q}_{pp}^{\mathrm{TR}}(f)\,w_{p\mathrm{R}}^{(k)} + \sum_{j=p+1}^{p+\nu_p-k-1}\bar{Q}_{pj}^{\mathrm{T}}(f)\,w_j^{(k)}\,. \tag{2.4.3}$$

As for any $f \in F$ spectral radius of $\bar{Q}_{pp}^{\mathrm{TT}}(f)$ is less than 1, by Lemma 2.1 we can select $w_{p\mathrm{T}}^{(k)}$'s such that for any $f \in F$

$$(\psi_{p\mathrm{T}}^{(\nu_p-1)}(f), \ldots, \psi_{p\mathrm{T}}^{(0)}(f)) \leqq 0 \tag{2.4.4}$$

and for some $f^* \in F$

$$(\psi_{p\mathrm{T}}^{(\nu_p-1)}(f^*), \ldots, \psi_{p\mathrm{T}}^{(0)}(f^*)) = 0\,. \tag{2.4.4'}$$

As each f in (2.4.4) is selected from a finite set F, there also exists some $n_0 < \infty$ such that for any $n \geqq n_0$

$$\sum_{k=0}^{\nu_p-1}\binom{n}{k}\psi_{p\mathrm{T}}^{(k)}(\hat{f}^{(n)}) \leqq 0\,. \tag{2.4.5}$$

By (2.4.2), (2.4.5) we conclude that for any $n \geqq n_0$

$$y_{p\mathrm{T}}^{(p)}(n+1) \leqq \bar{Q}_{pp}^{\mathrm{TT}}(\hat{f}^{(n)})\,y_{p\mathrm{T}}^{(p)}(n) + h_{p\mathrm{T}}(\hat{f}^{(n)}; n) \tag{2.4.6}$$

where

$$h_{p\mathrm{T}}(f;n) = \bar{Q}^{\mathrm{TR}}_{pp}(f)\, y^{(p)}_{p\mathrm{R}}(n) + \sum_{j=p+1}^{s} \bar{Q}^{\mathrm{T}}_{pj}(f)\, y^{(p)}_{j}(n)\,. \tag{2.4.6*}$$

Similarly, by (2.4.2), (2.4.4′) and (2.4.6*) we get

$$y^{(p)}_{p\mathrm{T}}(n+1) \geqq \bar{Q}^{\mathrm{TT}}_{pp}(f^*)\, y^{(p)}_{p\mathrm{T}}(n) + h_{p\mathrm{T}}(f^*;n)\,. \tag{2.4.6′}$$

As by assumptions (2.14), (2.14′) $\lim_{n\to\infty} h(f;n) = 0$, in virtue of Lemma 2.2 applied to (2.4.6), (2.4.6′) we immediately get (2.15).

To finish the proof, let us denote (for any stationary policy $\pi \equiv (f)$)

$$\bar{x}_{p\mathrm{T}}(n;f) = \sigma_p^{-n}\, x_{p\mathrm{T}}(n;f),\ y^{(p)}_{p\mathrm{T}}(n;f) = \bar{x}_{p\mathrm{T}}(n;f) - \sum_{k=0}^{\nu_p-1} \binom{n}{k} w^{(k)}_{p\mathrm{T}}$$

(where $w^{(k)}_{p\mathrm{T}}$'s satisfy (2.4.4), (2.4.4′)) and define similarly $y^{(p)}_{p\mathrm{R}}(n;f)$, $y^{(p)}_{pj}(n;f)$. Mimicking the reasoning used in (2.4.1), (2.4.2), we conclude that for $f^* \in F$ satisfying (2.4.4′) together with conditions (2.16), (2.16′) (i.e. $f^*(k) = \tilde{f}(k)$ for any $k \in I^{\mathrm{R}}_p(\hat{f})$ or $k \in I_i(\hat{f})$ with $i > p$)

$$y^{(p)}_{p\mathrm{T}}(n+1;f^*) = \bar{Q}^{\mathrm{TT}}_{pp}(f^*)\, y^{(p)}_{p\mathrm{T}}(n;f^*) + h_{p\mathrm{T}}(f^*;n)\,. \tag{2.4.7}$$

Applying again Lemma 2.2 to (2.4.7) and (2.4.6*) we immediately conclude that

$$\lim_{n\to\infty} \left[\bar{x}_{p\mathrm{T}}(n;f^*) - \sum_{k=0}^{\nu_p-1} \binom{n}{k} w^{(k)}_{p\mathrm{T}} \right] = 0 \tag{2.4.8}$$

and so (2.17) follows immediately by (2.4.8) and (2.15). □

Remark 2.3. In case that an exponential convergence is assumed in (2.14), (2.14′) (then, evidently, $h_{p\mathrm{T}}(f;n)$ tends exponentially fast to zero for $n \to \infty$ and any $f \in F$), using the same reasoning as in Lemma 2.2 we can also establish that the convergence in (2.15) is also exponential (observe that if the convergence in (2.2.4) is exponential, also the bounds in (2.2.10) converge exponentially to zero).

3. STATIONARY OPTIMAL POLICIES IN TRANSIENT DYNAMIC PROGRAMMING

This section deals with the existence of stationary optimal policies if cumulative rewards are considered in transient dynamic programming models. Transient dynamic programming can be identified as a special case of the presented dynamic programming model for which the additional condition

$$\sigma \equiv \max_{f\in F} \sigma(f) < 1 \tag{3.1}$$

is fulfilled. Notice that discounted dynamic programming, widely discussed in the

literature, turns out to be only a very special case of this dynamic programming model.

First, by employing the results of Proposition 4, we establish the asymptotic behaviour of the vector of l-order cumulative rewards $v^{(l)}(n)$, resp. $v^{(l)}(n; f)$, defined recursively by (1.6), resp. (1.5*); in particular, we show that for $n \to \infty$ $v^{(l)}(n)$, resp. $v^{(l)}(n; f)$, converges to some (vector) polynomial.

Theorem 3.1. Let (3.1) hold. Then, for any $l = 1, 2, \ldots$, there exist vectors $w^{(k,l)}$ $(k = 0, 1, \ldots, l-1)$ and a decision vector $f^*_{(l)} \in F$ such that

$$\lim_{n\to\infty} \left[v^{(l)}(n) - \sum_{k=0}^{l-1} \binom{n}{k} w^{(k,l)} \right] = 0 , \tag{3.1.1}$$

$$\lim_{n\to\infty} [v^{(l)}(n) - v^{(l)}(n; f^*_{(l)})] = 0 . \tag{3.1.2}$$

Proof. Let us introduce $M(f)$, $z(n)$ by (1.7), (1.8), (1.8′) and write $M(f)$ in the following block-triangular form

$$M(f) = \begin{bmatrix} M_{11}(f) & M_{12} & \ldots & M_{1l} \\ 0 & M_{22} & \ldots & M_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & M_{ll} \end{bmatrix} \tag{3.1.3}$$

where $M_{11}(f) = \begin{bmatrix} Q(f) & r(f) \\ 0 & 1 \end{bmatrix}$, $M_{1k} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ (for $k = 2, \ldots, l$; 0 denotes N-column zero vector), and for any $k \geqq i \geqq 2$ $M_{ik} = 1$. Recalling (3.1) the spectral radius of $M_{11}(f)$ equals 1 and for the spectral radius, resp. index, of each M_{ii} (with $i = 2, \ldots, l$) we have $\tilde{\sigma}_i = 1$, resp. $\tilde{\nu}_i = l - i + 1$.

Now, let for $i = 1, \ldots, l$ $z_i(n)$ be a subvector of $z(n)$ (cf. (1.8)) whose elements correspond to those of $M_{11}(f)$ and M_{ii}'s (so $z_i(n)$ is a scalar for each $i > 1$). As (cf. (1.8′)) we have chosen $z_1(0) = [0 \ 1]'$ (here 0 denotes N-dimensional zero row vector) and $z_i(0) = 1$ for any $i > 1$, by (1.9), (3.1.3) we get

$$z_1(n) = \left[(v^{(l)}(n))', \binom{n+l-1}{l-1} \right]' \quad \text{and} \quad z_i(n) = \binom{n+l-i}{l-i} \quad \text{for} \quad i = 2, \ldots, l .$$

Recalling the standard formula

$$\sum_{k=0}^{m} \binom{n}{k} \binom{p}{m-k} = \binom{n+p}{m} \tag{3.1.4}$$

we immediately get

$$\binom{n+l-i}{l-i} = \sum_{k=0}^{l-i} \binom{n}{k} \binom{l-i}{k} . \tag{3.1.4′}$$

Employing (3.1.4′) we can write

$$[z_1(n+1)]_{N+1} = \sum_{k=0}^{l-1} \binom{n}{k} [w_1^{(k,l)}]_{N+1} \quad \text{where} \quad [w_1^{(k,l)}]_{N+1} = \binom{l-1}{k} \tag{3.1.5}$$

and similarly for $i = 2, \ldots, l$ we have

$$z_i(n) = \sum_{k=0}^{l-i} \binom{n}{k} w_i^{(k,l)} \quad \text{with} \quad w_i^{(k,l)} = \binom{l-i}{k}. \tag{3.1.5'}$$

So the assumptions of Proposition 4 are satisfied (observe that (3.1.5), resp. (3.1.5'), corresponds to (2.14), resp. (2.14'), and $Q(f)$ constitutes a non-basic class of $M_{11}(f)$). Then by (2.15) of Proposition 4 it is possible to compute N-dimensional vectors $w^{(k,l)}$ (for $k = 0, \ldots, l-1$) such that (3.1.1) holds. As the basic class of $M(f)$ as well as all M_{ij} with $i, j > 1$ do not depend on f, conditions (2.16), (2.16') of Proposition 4 are trivially fulfilled and (3.1.2) follows then immediately by (2.17). □

Remark 3.2. Employing the facts mentioned in Remark 2.3, we can easily establish that the convergence in (3.1.1) and (3.1.2) is exponential; i.e. that there exist vectors $c \ll 0$, $d \gg 0$ and a number $\varrho \in (\sigma, 1)$ such that for each n

$$c\varrho^n \leqq v^{(l)}(n) - \sum_{k=0}^{l-1} \binom{n}{k} w^{(k,l)} \leqq d\varrho^n \tag{3.2.1}$$

$$c\varrho^n \leqq v^{(l)}(n) - v^{(l)}(n; f_{(l)}^*) \leqq d\varrho^n . \tag{3.2.2}$$

Moreover, by a careful examination of the proof of Theorem 3.1 we can establish:

Corollary 3.3. For any (fixed) $l = 1, 2, \ldots$ and all $k = 0, 1, \ldots, l-1$ $w^{(k,l)}$'s (coefficients of the vector polynomial in (3.1.1)) can be found (e.g. by the policy iteration algorithm used in the proof of Lemma 2.1) as a (unique) solution to the following set of equations considered for $k = l-1, \ldots, 0$

$$\max_{f \in F^{(k+1,l)}} \psi^{(k,l)}(f) = 0 \tag{3.3.1}$$

where (we set $w^{(l,l)} = 0$)

$$\psi^{(k,l)}(f) = \binom{l-1}{k} r(f) + (Q(f) - I) w^{(k,l)} - w^{(k+1,l)} \tag{3.3.1'}$$

and $F^{(k,l)}$ are defined recursively by

$$F^{(k,l)} = \{f \in F^{(k+1,l)} : \psi^{(k,l)}(f) = 0\} \quad \text{with} \quad F^{(l,l)} \equiv F . \tag{3.3.2}$$

Furthermore, for any $m = 1, 2, \ldots$

$$w^{(k+m,l+m)} = \sum_{j=0}^{m} \binom{m}{j} w^{(k+j,l)} \tag{3.3.3}$$

and

$$F^{(k,l)} = F^{(k+m,l+m)} \supset \ldots \supset F^{(0,l)} = F^{(m,l+m)} \supset \ldots \supset F^{(0,l+m)} \neq \emptyset . \tag{3.3.3'}$$

Proof. (3.3.1), (3.3.2) can be verified by applying the results of Proposition 4 to the specific structure given by (3.1.3), (3.1.5) and by employing Lemma 2.1. To this order observe that, for the considered specific structure, in (2.4.3) $\bar{Q}_{pp}^{TR}(f) = r(f)$,

$\bar{Q}^{\mathrm{T}}_{pj}(f) = 0$ for all $j > p$, and $w^{(k)}_{p\mathrm{R}} = \binom{l-1}{k}$. So (3.3.1′) follows immediately by (2.4.3) on replacing $\bar{Q}^{\mathrm{TT}}_{pp}(f)$, $w^{(k)}_{p\mathrm{T}}$ by $Q(f)$, $w^{(k,l)}$, respectively. Employing Lemma 2.1 we can easily verify the existence and compute solutions of (3.3.1) and (3.3.2).

To establish (3.3.3), (3.3.3′), first observe that by (3.1.4) we have (notice that $\binom{n}{m} = 0$ if $m > n$)

$$\binom{l+m-1}{k+m} = \sum_{j=0}^{m} \binom{m}{j}\binom{l-1}{k+j} \tag{3.3.4}$$

and by (3.3.1′), (3.3.4) we get

$$\sum_{j=0}^{m} \binom{m}{j} \psi^{(k+j,l)}(f) = \binom{l+m-1}{k+m} r(f) + (Q(f) - I) \sum_{j=0}^{m} \binom{m}{j} w^{(k+j,l)} - \tag{3.3.5}$$

$$- \sum_{j=0}^{m} \binom{m}{j} w^{(k+1+j,l)} .$$

As there exists unique solution of (3.3.1), (3.3.1′), (3.3.2) (notice that $w^{(k,l)}$'s are calculated successively for $k = l-1, \ldots, 0$) by recalling that

$$(\psi^{(l-1,l)}(f), \ldots, \psi^{(k,l)}(f)) \lesseqgtr 0 \quad \text{with equality iff} \quad f \in F^{(k,l)}$$

and using (3.3.5) for $k = l-1, \ldots, k$ (and comparing it with $\psi^{(m+k,l+m)}(f)$) we conclude that (3.3.3) and (3.3.3′) must hold. □

The results of Theorem 3.1 and Corollary 3.3 enable to present Theorem 3.4 establishing that optimal policies can be found in the class of stationary policies if various cumulative rewards according to (1.3) (observe that $v^{(l)}(\pi; t)$ are not necessarily bounded if $l > 1$) are considered in transient dynamic programming.

Theorem 3.4. Let (3.1) hold. Then to each $m = 1, 2, \ldots$ there exists stationary policy $\pi^*_{(m)} \equiv (f^*_{(m)})$ such that for all $l = 1, 2, \ldots, m$, and any policy $\pi \equiv (f_t)$

$$\liminf_{t \to \infty} [v^{(l)}(\pi^*_{(m)}; t) - v^{(l)}(\pi; t)] \geqq 0 . \tag{3.4.1}$$

Proof. Choosing $f^*_{(m)} \in F^{(0,m)}$ by (3.1.2) and (3.3.3′) we get

$$\lim_{n \to \infty} [v^{(l)}(n) - v^{(l)}(n; f^*_{(m)})] = 0 \quad \text{for any} \quad l = 1, \ldots, m . \tag{3.4.2}$$

Recalling that $v^{(l)}(t)$ are defined recursively by (1.6)

$$v^{(l)}(t+1) \geqq v^{(l)}(\pi; t) \quad \text{for any} \quad \pi \equiv (f_t) \quad \text{and each} \quad t = 0, 1, \ldots . \tag{3.4.3}$$

As $\pi^*_{(m)} \equiv (f^*_{(m)})$ is stationary, (cf. (1.5*)) $v^{(l)}(t+1; f^*_{(m)}) = v^{(l)}(\pi^*_{(m)}; t)$ and (3.4.1) follows then immediately by (3.4.2) and (3.4.3). □

Remark 3.5. The case with $l = 1$ is well-known from [11]. Recall that for any $\pi \equiv (f_t)$ $\lim_{t\to\infty} v^{(1)}(\pi; t) = v^{(1)}(\pi)$ always exists and that by (3.4.1) we get $v^{(1)}(\pi^*_{(1)}) \geqq \geqq v^{(1)}(\pi)$.

4. STATIONARY OPTIMAL POLICIES IN NORMALIZED DYNAMIC PROGRAMMING

This section is devoted to "normalized" dynamic programming, i.e. the presented dynamic programming model fulfilling the additional condition $\sigma \equiv \max_{f\in F} \sigma(f) = 1$. We establish a family of sensitive averaging optimality criteria having a nice property that a stationary optimal policy exists.

Our analysis heavily depends on the properties of the set $\{Q(f), f \in F\}$ summarized in Proposition 1. To this order the decomposition of $Q(f)$, mentioned in Proposition 1, will be employed. In particular, for the considered "normalized" dynamic programming model it will be useful to write

$$Q(f) = \begin{bmatrix} Q_{11}(f) & Q_{12}(f) & \dots & Q_{1p}(f) & Q_{1,p+1}(f) \\ 0 & Q_{22}(f) & \dots & Q_{2p}(f) & Q_{2,p+1}(f) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & Q_{pp}(f) & Q_{p,p+1}(f) \\ 0 & 0 & \dots & 0 & Q_{p+1,p+1}(f) \end{bmatrix} \tag{4.1}$$

where the spectral radius of each $Q_{ii}(f)$ (denoted by $\sigma_i(f)$) is nongreater than 1 for $i = 1, \dots, p$, resp. less than 1 for $i = p + 1$, and at least for one $\hat{f} \in F$

$$\sigma_i(\hat{f}) \equiv \sigma_i = 1 \quad \text{for any} \quad i = 1, \dots, p\,. \tag{4.2}$$

Observe that for the index of $Q_{ii}(\hat{f})$ it holds $v_i(\hat{f}) \equiv v_i = 1 - i + p$ and that $Q_{p+1,p+1}(f)$ can be vacuous. However, by adding to $Q(f)$ an auxiliary $(N + 1)$-th row (of dimension N), resp. $(N + 1)$-th column (of dimension $N + 1$), whose elements equal 0, resp. $\varepsilon > 0$, the structure of $Q(f)$ remains unchanged, only $\sigma_{p+1} > 0$. Recall that (for any $i = 1, \dots, p$, $p + 1$ and any $f \in F$) elements of each $Q_{ii}(f)$ are labelled by integers of $I_i(\hat{f})$ ($\hat{f} \in F$ fixed, $\bigcup_{i=1}^{p+1} I_i(\hat{f}) = I$), and that for any $i = 1, \dots, p$ and any $f \in F$

$$Q_{ii}(f)\, u_i \leqq Q_{ii}(\hat{f})\, u_i = u_i \quad \text{with} \quad u_i \gg 0\,. \tag{4.3}$$

Remember that N-column vectors $v^{(l)}(n; f)$, resp. $v^{(l)}(n)$, are defined recursively by (1.5*), resp. (1.6), and in virtue of (4.1) $v_i^{(l)}(n; f)$, resp. $v_i^{(l)}(n)$, $(i = 1, \dots, p, p + 1)$ denotes a subvector of $v^{(l)}(n; f)$, resp. $v^{(l)}(n)$, whose components are labelled by integers from $I_i(\hat{f})$.

First we establish some bounds on $v^{(l)}(n)$.

Theorem 4.1. Let (4.1), (4.2) hold and $l = 1, 2, \ldots$ be fixed. Then for $i = 1, \ldots, p, p+1$ there exists vectors $w_i^{(k,l)}$ (where $k = 0, \ldots, l + p - i$) and a decision vector $f_{(l)}^* \in F$ such that

$$\lim_{n\to\infty} \binom{n}{p-i+1}^{-1} \left[v_i^{(l)}(n) - \sum_{k=0}^{l+p-i} \binom{n}{k} w_i^{(k,l)} \right] = 0 , \tag{4.1.1}$$

$$\lim_{n\to\infty} \binom{n}{p-i+1}^{-1} [v_i^{(l)}(n) - v_i^{(l)}(n; f_{(l)}^*)] = 0 . \tag{4.1.2}$$

Proof. Similarly as in the proof of Theorem 3.1, let us introduce by (1.7) matrix $M(f)$ (being of dimension $N + l$) and write $M(f)$ in the following block-triangular form

$$M(f) = \begin{bmatrix} M_{11}(f) & \ldots & M_{1p}(f) & \ldots & M_{1,p+l} \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & \ldots & M_{pp}(f) & \ldots & M_{p,p+l} \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & \ldots & M_{p+l,p+l} \end{bmatrix} \tag{4.1.3}$$

Taking into account (1.7) and (4.1), $M_{ij}(f)$'s in (4.1.3) will be given by the following rules:

(i) For any $i = 1, \ldots, p$ we set

$$M_{ij}(f) = Q_{ij}(f) \quad \text{for} \quad j = i, \ldots, p \tag{4.1.4}$$

$$M_{i,p+1}(f) = [Q_{i,p+1}(f) \quad r_i(f)] \tag{4.1.4'}$$

$$M_{ij}(f) = 0 \quad \text{if} \quad j < i \quad \text{or} \quad j > p + 1 . \tag{4.1.4''}$$

(ii) We denote

$$M_{p+1,p+1}(f) = \begin{bmatrix} Q_{p+1,p+1}(f) & r_{p+1}(f) \\ 0 & 1 \end{bmatrix} \tag{4.1.5}$$

$$M_{p+1,j}(f) \equiv M_{p+1,j} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{for each} \quad j > p + 1 \tag{4.1.5'}$$

(here 0 stands for zero column vector being of the same dimension as $r_{p+1}(f)$), and set

$$M_{p+1,j}(f) \equiv M_{p+1,j} = 0 \quad \text{for each} \quad j < p + 1 . \tag{4.1.5''}$$

(iii) For any $i = p + 2, \ldots, p + l$ $M_{ij}(f) \equiv M_{ij}$ are scalars such that

$$M_{ij} = 1 \quad \text{for each} \quad j \geqq i , \quad \text{and} \tag{4.1.6}$$

$$M_{ij} = 0 \quad \text{if} \quad j < i . \tag{4.1.6'}$$

Now let us introduce by (1.8), (1.8′) an auxiliary sequence of $(N + l)$-dimensional vectors $\{z(n),\ n = 0, 1, \ldots\}$ and let $z_i(n)$ be a subvector of $z(n)$ whose elements correspond to those of $M_{ii}(f)$ (so for $i > p + 1$ $z_i(n)$ are scalars). Taking into account (4.1.3) by (1.8), (1.8′) we can verify that

$$z_i(n) = \begin{cases} \binom{n + p + l - i}{p + l - i} & \text{for} \quad i = p + 2, \ldots, p + l \\ \left[(v_{p+1}^{(l)}(n))', \binom{n + l - 1}{l - 1} \right]' & \\ v_i^{(l)}(n) \quad \text{for} \quad i = 1, \ldots, p\,. & \end{cases} \tag{4.1.7}$$

Denoting

$$\tilde{z}(n) = \begin{bmatrix} z_{p+1}(n) \\ \vdots \\ z_{p+l}(n) \end{bmatrix}, \quad \tilde{M}(f) = \begin{bmatrix} M_{p+1,p+1}(f) & \ldots & M_{p+1,p+l}(f) \\ \vdots & \ddots & \vdots \\ 0 & \ldots & M_{p+l,p+l}(f) \end{bmatrix}$$

and recalling that $\sigma_{p+1}(f) < 1$, by (4.1.5)–(4.1.6) we can easily see that the procedure used in the proof of Theorem 3.1 can be applied for $\tilde{z}(n + 1) = \max_{f \in F} \tilde{M}(f)\, \tilde{z}(n)$; so (cf. (3.1.5′)) for $i > p + 1$

$$z_i(n) = \sum_{k=0}^{l+p-i} \binom{n}{k} w_i^{(k,l)} \quad \text{with} \quad w_i^{(k,l)} = \binom{l + p - i}{k} \tag{4.1.8}$$

and for $i = p + 1$ (4.1.1) together with (4.1.2) follow then immediately by (3.1.1) and (3.1.2); i.e. for $n \to \infty$

$$z_{p+1}(n) = \begin{bmatrix} v_{p+1}^{(l)}(n) \\ \sum_{k=0}^{l-1} \binom{n}{k} \binom{l-1}{k} \end{bmatrix} \to \sum_{k=0}^{l-1} \binom{n}{k} \begin{bmatrix} w_{p+1}^{(k,l)} \\ \binom{l-1}{k} \end{bmatrix}. \tag{4.1.8′}$$

The proof can be finished by employing Proposition 3. To this order observe that by (4.1.8) and (4.1.1), (4.1.2) (already established for $i = p + 1$) conditions (2.7), (2.8) of Proposition 3 (written for $z_j(n)$ instead of $x_j(n)$ and considered for $j = p + 1, \ldots, p + l$ instead of $j = p, \ldots, i_0 + v_{i_0} - 1$) are trivially fulfilled. So (4.1.1), resp. (4.1.2), for $i = 1, \ldots, p$ follows immediately by applying (2.8), resp. (2.10′), of Proposition 3. □

In general, vectors $w_i^{(k,l)}$ in (4.1.1) can be constructed by the methods for finding coefficients of bounding polynomials in (2.8) of Proposition 3 (cf. [9] for details). However, in (4.1.1) it suffices to construct $w_i^{(k,l)}$'s only for $k = l + p - i, \ldots, p - i + 1$; moreover, taking into account the specific structure of matrix $M(f)$ (cf. (4.1.3)–(4.1.6)) the resulting equations can be simplified.

Introducing $F_i \equiv \mathop{\mathsf{X}}_{j \in I_i(f)} F(j)$, first observe that $w_{p+1}^{(k,l)}$ (for $k = 0, \ldots, l - 1$) can be found as a solution of (3.3.1), (3.3.1′), (3.3.2) (where we replace $r(f)$, $Q(f)$, $F^{(k,l)}$ by $r_{p+1}(f)$, $Q_{p+1,p+1}(f)$, $F_{p+1}^{(k,l)}$, respectively). Similarly, $w_i^{(k,l)}$ (for $k = l + p - i, \ldots$

..., 0) can be calculated successively for $i = p, p-1, \ldots, 1$ by employing the results of

Corollary 4.2. Let for $j = i+1, \ldots, p+1$ $w_j^{(k,l)}$ (with $k = l+p-j, \ldots, 0$) be known. Then $w_i^{(k,l)}$ (for $k = l+p-i, \ldots, 0$) can be computed (e.g. by the policy iteration algorithm used in the proof of Theorem 2.1 of [9]) as a solution of equations (considered for $k = l+p-i, \ldots, 0$)

$$\max_{f \in F_i^{(k+1,l)}} \varphi_i^{(k,l)}(f) = 0 \tag{4.2.1}$$

where (we set $w_i^{(l+p-i+1,l)} = 0$)

$$\varphi_i^{(k,l)}(f) = (Q_{ii}(f) - I)\, w_i^{(k,l)} - w_i^{(k+1,l)} + \sum_{j=i+1}^{l+p-k} Q_{ij}(f)\, w_j^{(k,l)} \tag{4.2.2}$$

$$(\text{for } k = p+l-i, \ldots, l)$$

$$\varphi_i^{(k,l)}(f) = (Q_{ii}(f) - I)\, w_i^{(k,l)} - w_i^{(k+1,l)} + \tag{4.2.2'}$$

$$+ \sum_{j=i+1}^{p+1} Q_{ij}(f)\, w_j^{(k,l)} + \binom{l-1}{k} r_i(f) \quad (\text{for } k = l-1, \ldots, 0)$$

and $F_i^{(k,l)}$ are defined recursively by

$$F_i^{(k,l)} = \{f \in F_i^{(k+1,l)} : \varphi_i^{(k,l)}(f) = 0\} \quad \text{with} \quad F_i^{(l+p-i+1,l)} \equiv F_i\,. \tag{4.2.3}$$

Furthermore, for $i = p+1, p, \ldots, 1$ and $m = 1, 2, \ldots$

$$F_i^{(k,l)} = F_i^{(k+m,l+m)} \supset \ldots \supset F_i^{(0,l)} = F_i^{(m,l+m)} \supset \ldots \supset F_i^{(0,l+m)} \tag{4.2.4}$$

and $w_i^{(k,l)}$ (not determined uniquely by (4.2.1)–(4.2.3) for $k = 0, \ldots, p-i$) can be selected such that

$$w_i^{(k+m,l+m)} = \sum_{j=0}^{m} \binom{m}{j} w_i^{(k+j,l)}\,. \tag{4.2.4'}$$

Proof. To show that $w_i^{(k,l)}$ in (4.1.1) can be found as a solution of (4.2.1)–(4.2.3), we adapt general procedures for finding coefficients $w_i^{(k)}$ of the bounding polynomial in (2.8) suggested in the proof of Theorem 3.1 of [9]. In particular, it suffices to select (by policy iterations) $w_i^{(k)}$'s in (3.3), (3.3') of [9] such that (3.3.5), (3.3.5'), (3.3.6) of [9] are fulfilled. So we only need to adapt these results to (1.8) and employ the specific structure of matrix $M(f)$ identified by (4.1.3)–(4.1.6).

To this order, considering class $Q_{ii}(\hat{f})$ with respect to matrix $M(\hat{f})$, by (4.1), (4.1.3) for $\tilde{v}_i(\hat{f})$ (index of $Q_{ii}(\hat{f})$) we get $\tilde{v}_i(\hat{f}) = v_i(\hat{f}) + l = l+p-i+1$. Then (3.3), (3.3') of [9] can be written (for $k = p+l-i, \ldots, 0$) as

$$\varphi_i^{(k,l)}(f) = (M_{ii}(f) - I)\, w_i^{(k,l)} - w_i^{(k+1,l)} + \sum_{j=i+1}^{l+p-k} M_{ij}(f)\, w_j^{(k,l)}\,. \tag{4.2.5}$$

Inserting from (4.1.1), (4.1.4'), (4.1.4'') into (4.2.5) and recalling (4.1.8'), we im-

mediately get (4.2.2), (4.2.2′); so (4.2.1)–(4.2.3) follow directly from (3.3.5), (3.3.5′) and (3.3.6) of [9].

(4.2.4), (4.2.4′) can be verified by induction on $i = p+1, \ldots, 1$. First observe that for $i = p+1$ (4.2.4), (4.2.4′) have been already established in Corollary 3.3 (cf. (3.3.3), (3.3.3′) where we replace $w^{(k,l)}, F^{(k,l)}$ by $w_{p+1}^{(k+1,l)}, F_{p+1}^{(k,l)}$, respectively). Now let us assume that (4.2.4), (4.2.4′) hold for any $j = i+1, \ldots, p+1$ and notice that, instead of (4.2.2), only (4.2.2′) can be considered for $k = p+l-i, \ldots, 0$ if we set $w_i^{(k,l)} = 0$ for any $k > p+l-i$. Then (similarly as in the proof of Corollary 3.4) from (4.2.2′) we get

$$\sum_{j=0}^{m}\binom{m}{j}\varphi_i^{(k+j,l)}(f) = (Q_{ii}(f) - I)\sum_{j=0}^{m}\binom{m}{j}w_i^{(k+j,l)} - \tag{4.2.6}$$

$$-\sum_{j=0}^{m}\binom{m}{j}w_i^{(k+1+j,l)} + h_i^{(k+m,l+m)}(f)$$

where (by (3.3.4) and the induction assumption)

$$h_i^{(k+m,l+m)}(f) = \sum_{t=i+1}^{p+1} Q_{it}(f)\sum_{j=0}^{m}\binom{m}{j}w_t^{(k+j,l)} + \tag{4.2.6′}$$

$$+\sum_{j=0}^{m}\binom{l-1}{k+j}\binom{m}{j}r_i(f) = \sum_{t=i+1}^{p+1} Q_{it}(f)\, w_t^{(k+m,l+m)} + \binom{l+m-1}{k+m}r_i(f)\,.$$

The proof can be completed by considering (4.2.1)–(4.2.6) for $k \doteq k+m$, $l \doteq l+m$ and comparing (4.2.2′) written for $k \doteq k+m$, $l \doteq l+m$ with (4.2.6), (4.2.6′) (cf. also arguments used in the proof of Corollary 3.3). □

Similarly as in the transient case, by employing the results of Theorem 4.1 and Corollary 4.2, we can easily establish existence of a family of sensitive averaging optimality criteria for which optimal policies can be found in the class of stationary policies.

Theorem 4.3. Let (4.1), (4.2) hold. Then to any $m = 1, 2, \ldots$ there exists stationary policy $\pi_{(m)}^* \equiv (f_{(m)}^*)$ such that for all $l = 1, \ldots, m$; $i = 1, \ldots, p, p+1$ and any policy $\pi \equiv (f_t)$

$$\liminf_{t\to\infty} t^{i-p-1}\left[v_i^{(l)}(\pi_{(m)}^*; t) - v_i^{(l)}(\pi; t)\right] \geqq 0\,. \tag{4.3.1}$$

Proof. The proof is strictly similar to that of Theorem 3.4. Choosing $f_{(m)}^* \in F^{(0,m)}$ by (4.1.2), (4.2.4) we get for $i = 1, \ldots, p, p+1$

$$\lim_{n\to\infty} n^{i-p-1}\left[v_i^{(l)}(n) - v_i^{(l)}(n; f_{(m)}^*)\right] = 0\,. \tag{4.3.2}$$

As $v^{(l)}(t)$ are calculated by dynamic programming recursion (1.6)

$$v_i^{(l)}(t) \geqq v_i^{(l)}(\pi; t) \quad \text{for any} \quad \pi \equiv (f_t) \quad \text{and each} \quad t = 0, 1, \ldots\,. \tag{4.3.3}$$

However, $\pi^*_{(m)} \equiv (f^*_{(m)})$ is stationary, so (cf. (1.5*)) $v^{(l)}_i(t+1; f^*_{(m)}) = v^{(l)}_i(\pi^*_{(m)}; t)$, and (4.3.1) follows immediately by (4.3.2) and (4.3.3). □

Remark 4.4. By (4.3.1) we immediately conclude that to any $l = 1, 2, \ldots$ there exists stationary policy $\pi^*_{(l)}$ such that

$$\liminf_{t\to\infty} t^{-p}\left[v^{(l)}(\pi^*_{(l)}; t) - v^{(l)}(\pi; t)\right] \geqq 0 \tag{4.4.1}$$

for an arbitrary policy $\pi \equiv (f_t)$. In particular, if $Q(f)$ is stochastic for any $f \in F$, then (cf. (4.1), (4.2)) $p = 1$, $Q(f) = Q_{11}(f)$ and (4.4.1) reads

$$\liminf_{t\to\infty} t^{-1}\left[v^{(l)}(\pi^*_{(l)}; t) - v^{(l)}(\pi; t)\right] \geqq 0\,. \tag{4.4.2}$$

Recall that (4.4.2) was also obtained by different methods in [7] (cf. Theorem 3.5 and Lemma 3.4 of [7]). Moreover, if $l = 1$ then $t^{-1} v^{(1)}(\pi; t)$ are bounded, and (4.4.2) reduces to the well-known result on the existence of a stationary average optimal policy in a classical Markov decision chain; i.e. we get

$$\lim_{t\to\infty} t^{-1} v^{(1)}(\pi^*_{(1)}; t) \geqq \limsup_{t\to\infty} t^{-1} v^{(1)}(\pi; t)\,. \tag{4.4.2'}$$

5. MODELS WITH EXPONENTIALLY GROWING UTILITIES

To present a complete analysis of the considered dynamic programming model, it only remains to discuss the case with $\sigma \equiv \max_{f\in F} \sigma(f) > 1$. Our analysis will again employ the "uniform" decomposition of the set $\{Q(f), f \in F\}$ according to (4.1); however, in this section (instead of (4.2)) we assume that

$$\max_{f\in F} \sigma_i(f) \equiv \sigma_i = \sigma > 1 \quad \text{for any} \quad i = 1, \ldots, p \tag{5.1}$$

$$\sigma_{p+1}(f) < \sigma \quad \text{for any} \quad f \in F\,. \tag{5.1'}$$

Similarly as in Section 4, $v^{(l)}_i(n)$ denotes a subvector of $v^{(l)}(n)$ (calculated by dynamic programming recursion (1.6)) whose components correspond to those of submatrix $Q_{ii}(f)$ in (4.1). Comparing with the results for transient and normalized cases, if (5.1) is assumed l-order cumulative rewards given by $v^{(l)}(n)$ grow exponentially as it is indicated in

Theorem 5.1. Let (4.1) and (5.1), (5.1') hold. Then for any (fixed) $l = 1, 2, \ldots$ there exists vectors $c \ll 0$, $d \gg 0$ such that for all $i = 1, \ldots, p$

$$n^{p-i} c_i \leqq \sigma^{-n} v^{(l)}_i(n) \leqq n^{p-i} d_i \tag{5.1.1}$$

$$\lim_{n\to\infty} \sigma^{-n} v^{(l)}_{p+1}(n) = 0\,. \tag{5.1.2}$$

Proof. In virtue of (4.1), (5.1), (5.1') considering the class $Q_{ii}(\hat{f})$ with respect

to matrix $M(\hat{f})$ given by (1.7), for $\tilde{v}_i(\hat{f})$ (index of $Q_{ii}(\hat{f})$) we get $\tilde{v}_i(\hat{f}) = v_i(\hat{f}) = p - i + 1$. So (5.1.1), (5.1.2) follow immediately by (1.9) and the bounds established in Proposition 2. □

Remark 5.2. Comparing the proof of Theorem 5.1 with analogous procedures for normalized and transient dynamic programming, we can see that in Theorem 5.1 no polynomial bounds were employed for establishing the estimates of $v_i^{(1)}(n)$. By (5.1.1) we immediately get for an arbitrary policy and each $i = 1, \ldots, p$

$$\lim_{t\to\infty} t^{i-p-1}\sigma^{-t}\, v_i^{(1)}(\pi; t) = 0\,. \tag{5.2.1}$$

In particular, if $p = 1$ $\{\sigma^{-t}\, v^{(1)}(\pi; t),\ t = 0, 1, \ldots\}$ is bounded but, as it is shown in the following Example, there need not exist any stationary policy $\pi^* \equiv (f^*)$ such that for an arbitrary policy $\pi \equiv (f_t)$

$$\liminf_{t\to\infty} \sigma^{-t}\left[v^{(1)}(\pi^*; t) - v^{(1)}(\pi; t)\right] \geqq 0\,. \tag{5.2.2}$$

Example. Let $F \equiv \{f_1, f_2\}$ with $r(f_1) = r(f_2) = [1\ 0]'$ and

$$Q(f_1) = \begin{bmatrix} 5 & 5 \\ 8 & 2 \end{bmatrix}, \quad Q(f_2) = \begin{bmatrix} 2 & 8 \\ 8 & 2 \end{bmatrix} \quad \text{(obviously } \sigma = \sigma(f_1) = \sigma(f_2) = 10)\,.$$

Denoting $\pi_1 \equiv (f_1)$, $\pi_2 \equiv (f_2)$ and recalling dynamic programming recursion (1.6), obviously, $v^{(1)}(n)$ are obtained using a nonstationary policy selecting alternatively f_1 and f_2. Furthermore, by an easy calculation we get

$$\lim_{t\to\infty} \sigma^{-t}\, v^{(1)}(t) = [0.073\ \ 0.073]'\,, \quad \text{however}\,,$$

$$\lim_{t\to\infty} \sigma^{-t}\, v^{(1)}(\pi_2; t) \ll \lim_{t\to\infty} \sigma^{-t}\, v^{(1)}(\pi_1; t) = [0.068\ \ 0.068]'\,;$$

so no stationary policy satisfying (5.2.2) exists.

6. CONCLUSION

In the present paper we have established a family of optimality criteria for cumulative rewards in multiplicative Markov decision chains having a nice property that an optimal policy can be selected in the class of stationary policies. We have slightly extended known results for transient dynamic programming; however, the heart of the paper consists in the results for "normalized" dynamic programming models. On the contrary to the approaches used in the literature (cf. [2, 3, 6, 7, 10, 11]), our analysis heavily employs the "backward recursions" of dynamic programming together with the polynomial bounds on the respective utility vector (cf. [8], [9]). Furthermore, we have also demonstrated that for the remaining case with $\sigma > 1$ standard optimality criteria need not guarantee existence of a stationary policy.

(Received June 24, 1981.)

REFERENCES

[1] F. R. Gantmakher: Teoriya matric. Second edition, Nauka, Moskva 1966.

[2] J. Flynn: Conditions for equivalence of optimality criteria in dynamic programming. Ann. Statist. *4* (1976), 5, 936—953.

[3] A. Hordijk, K. Sladký: Sensitive optimality criteria in countable state dynamic programming. Mathem. Oper. Res. *2* (1977), 1, 1—14.

[4] R. A. Howard, J. E. Matheson: Risk-sensitive Markov decision processes. Manag. Sci. *18* (1972), 7, 357—369.

[5] P. Mandl: Controlled Markov chains (in Czech). Kybernetika *6* (1969), Supplement, 1—74.

[6] U. G. Rothblum: Multiplicative Markov Decision Chains. PhD Dissertation, Dept. Oper. Res., Stanford U., Stanford, Calif. 1974.

[7] K. Sladký: On the set of optimal controls for Markov chains with rewards. Kybernetika *10* (1974), 4, 350—367.

[8] K. Sladký: Bounds on discrete dynamic programming recursions I — Models with non-negative matrices. Kybernetika *16* (1980), 6, 526—547.

[9] K. Sladký: Bounds on discrete dynamic programming recursions II — Polynomial bounds on problems with block-triangular structure. Kybernetika *17* (1981), 4, 310—328.

[10] A. F. Veinott, Jr.: Discrete dynamic programming with sensitive optimality criteria (preliminary report). Ann. Math. Statist. *39* (1968), 4, 1372.

[11] A. F. Veinott, Jr.: Discrete dynamic programming with sensitive discount optimality criteria. Ann. Math. Statist. *40* (1969), 5, 1635—1660.

Ing. Karel Sladký, CSc., Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation — Czechoslovak Academy of Sciences), Pod vodárenskou věží 4, 182 08 Praha 8. Czechoslovakia.