stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE BESLISKUNDE (DEPARTMENT OF OPERATIONS RESEARCH)

BW 90/78 AUGUSTUS

A. FEDERGRUEN & P.J. SCHWEITZER

NON-STATIONARY MARKOV DECISION PROBLEMS WITH CONVERGING PARAMETERS

Preprint

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O).

AMS(MOS) subject classification scheme (1970): 90C45

Non-stationary Markov decision problems with converging parameters*)

by

A. Federgruen & P.J. Schweitzer**)

ABSTRACT

This paper considers the solution of Markov Decision Problems, the parameters of which can only be obtained via approximating schemes, or in which it is computationally preferable to approximate the parameters rather than employing exact algorithms for their computation.

Various models are presented in which this situation occurs. Furthermore, it is shown that a modified value-iteration method may be employed both for the discounted and for the undiscounted version of the model, in order to solve the optimality equation and to find optimal policies. In both cases the convergence rate is shown to be geometric.

As a side result, we characterize the asymptotic behaviour of backward products of a geometrically convergent sequence of Markov matrices.

KEY WORDS & PHRASES: *non-stationary; Markov Decision Problems; approximations for parameters; value-iteration method; convergence rates.*

---

*) This report will be submitted for publication elsewhere.

**) Graduate School of Management, University of Rochester, Rochester, N.Y. 14627.

1. INTRODUCTION AND SUMMARY

This paper considers the solution of Markov Decision Problems (MDP's) the parameters of which can only be obtained via approximating schemes or in which it is computationally preferable to approximate the parameters rather than employing exact algorithms for their computation. Let $\Omega = \{1,\ldots,N\}$ denote the state space of the MDP. $K(i)$ represents the finite set of alternatives in state i, which is embedded in a super set $\mathcal{K}(i)$, i.e. $K(i) \subseteq \mathcal{K}(i)$, $i \in \Omega$. q_i^k denotes the one-step expected reward and P_{ij}^k the transition probability to state j when alternative $k \in \mathcal{K}(i)$ is chosen in state i. Note that $P_{ij}^k \geq 0$ and $\Sigma_j P_{ij}^k = 1 (i \in \Omega;\ k \in \mathcal{K}(i))$. Now, suppose that the parameters q_i^k, P_{ij}^k and the sets $K(i)$ $(i \in \Omega;\ k \in \mathcal{K}(i))$ are unknown in advance, but that instead one can compute sequences

(1.1) $\quad \{K(i,n)\}_{n=1}^{\infty} \to K(i);\ i \in \Omega \quad$ where $\quad K(i,n) \subseteq \mathcal{K}(i)$, i.e.

$$K(i,n) = K(i) \quad \text{for all } n \text{ sufficiently large.}$$

(1.2) $$\{q_i^k(n)\}_{n=1}^{\infty} \to q_i^k;\ i \in \Omega;\ k \in \mathcal{K}(i)$$

(1.3) $$\{P_{ij}^k(n)\}_{n=1}^{\infty} \to P_{ij}^k; \quad \text{where}$$

$$P_{ij}^k(n) \geq 0 \quad \text{and} \quad \sum_j P_{ij}^k(n) = 1; \qquad i,j \in \Omega;\ k \in \mathcal{K}(i)$$

This situation occurs in a large number of applications, as is illustrated by the following examples:

EXAMPLE 1. MDP's in which e.g. the one-step rewards q_i^k appear as the optimal values of underlying optimization problems. As an example, consider a resource or inventory system which serves to supply (say) n simultaneous users. At each period of time, one has to decide upon the amount to be withdrawn from the system, as well as upon the optimal way to allocate this amount among the n users. With i representing the inventory level (in the resource system) and k the amount to be withdrawn from the latter, the one-step net benefit q_i^k may be obtained by subtracting a holding cost function h(i) and

a transfer cost T(k) from the net benefit to the entire system that is associated with an optimal allocation of k units among the users. The latter may e.g. be computed by solving a mathematical program so that q_i^k could e.g. have the following structure

$$(1.4) \qquad q_i^k = -h(i) - T(k) + \max\ c(x) \quad \text{s.t. } x \in X,\ f(x) \geq k,\ x \geq 0$$

where x_i $(i = 1,\ldots,n)$ represents the amount allocated to the i-th user, and where the constraints $x \in X$ describe the restrictions imposed by the other resources and by the technological structure. There are various reasons for avoiding the computation of all of the q_i^k $(i \in \Omega,\ k \in K(i))$ *prior* to solving the MDP:

(a) in many applications, exact solution methods for the mathematical program in (1.4) are either non-available or hardly feasible, i.e. one needs or prefers to employ an approximation method, like a Lagrangean technique, a gradient projection method, or a reduced gradient method. Rather than first solving the $\Sigma_{i=1}^{n} \|K(i)\|$ mathematical programs with these approximation methods and next using ε-approximations for the q_i^k when solving the MDP - in case a good stopping criterion for the mathematical programs is at all available - one would prefer to use the approximating schemes for the q_i^k, in a method which *simultaneously* solves the MDP.

(b) For the actions that turn out to be suboptimal which in general represents the vast majority of the total number of $\Sigma_{i=1}^{N} \|K(i)\|$ actions, there is no need to do the computational effort of calculating the associated one-step expected rewards precisely.

In any method which generates approximating schemes for the numbers $\{q_i^k \mid i \in \Omega,\ k \in K(i)\}$ and *simultaneously* solves the MDP, one could stop the schemes associated with those actions that a test procedure detects to be suboptimal.

Suboptimality tests of this kind have been derived in connection with the value-iteration method both for the discounted and for the undiscounted

version of the model. With respect to the former we refer to GRINOLD [12], HASTINGS and MELLO [14], MACQUEEN [19] and PORTEUS [23] and as far as the latter is concerned, a devise for *temporary* elimination of suboptimal actions was proposed by HASTINGS [19], which although originally stated for the unichain case may be applied to the general multichain model (cf. remark 4 in FEDERGRUEN, SCHWEITZER and TIJMS [10]). For the unichain undiscounted case, a test for *permanent* elimination of actions was in addition devised by FEDERGRUEN, SCHWEITZER and TIJMS [10]. All of these elimination procedures can be adapted straightforwardly for the case where rather than applying value-iteration to a MDP with *exact* knowledge of the expected rewards, one would use upper and lower bounds that ultimately converge to the latter.

Note that most of the approximation techniques mentioned above for solving the mathematical programs in (1.4) have the special feature that whenever convergence occurs, the *rate* of convergence is at least *geometric*, where a vector sequence $\{x(n)\}_{n=1}^{\infty}$ is said to *converge to* x^*, *geometrically* if there exist numbers $K > 0$, and $0 \leq \lambda < 1$ such that

$$\|x(n) - x^*\| \leq K\lambda^n, \qquad n = 0,1,\ldots \tag{1.5}$$

(cf. e.g. sections 11.5 and 11.7 in LUENBERGER [18], as well as a recent survey on the subject by GOFFIN [11]; the occurrence or non-occurrence of geometric convergence is independent of the choice of norm on E^N). As examples of the above described model we refer to RUSSEL [24], VERKHOVSKY [31] and VERKHOVSKY and SPIVAK [32].

EXAMPLE 2. MDP's are generally used for describing dynamic systems which have to be controlled on a periodic basis and the design of which is assumed to be given. In many applications, however, one faces the problem of simultaneously having to make a one-time decision with respect to one or more design parameters as well as finding an optimal policy for operating the system, once the construction is complete. Usually both the laws of motion and the operating characteristics of the system are heavily affected by the choice of the design parameters. In mathematical terms, the problem amounts to solving

$$(1.6) \qquad \min_{\alpha \in A} \Big[\sum_{i=1}^{N} p_i(\alpha) V_i(\alpha) + \phi(\alpha) \Big]$$

where α represents a scalar or vector of design parameters, to be chosen out of a set A of feasible choices. $p_i(\alpha)$ represents the probability of starting the operation of the system in state i. In the discounted version of the model, $V_i(\alpha^*)$ would represent the minimal expected total discounted operating costs, when the initial state of the system is i, and $\phi(\alpha^*)$ the design costs when choosing $\alpha = \alpha^*$. Similarly, in the undiscounted version of the model, $V_i(\alpha^*)$ would denote the minimal long run average operating costs when starting in state i, and $\phi(\alpha^*)$ the depreciation and interest costs of the investment that is needed to implement the design parameters α^*. Note that the one-step rewards and transition probabilities in the MDP depend upon α, i.e.

$$(1.7) \qquad q_i^k \stackrel{\text{def}}{\equiv} q_i^k(\alpha); \qquad P_{ij}^k \stackrel{\text{def}}{\equiv} P_{ij}^k(\alpha); \qquad i,j \in \Omega;\ k \in K(i)$$

The optimization problem in (1.6) may be considered as a constrained minimization problem with resepct to α. Note that the optimal value of a MDP is not necessarily differentiable with respect to its parameters, and even if it is, the derivatives are extremely costly to compute.

As a consequence, one will have to confine oneself to *direct search methods* - like the Fibonacci method or the simplex method (cf. MURRAY [21]). Note that each evaluation of the objective function in (1.6) or its gradient with respect to α, requires the solution of a MDP which is extremely expensive. On the other hand, in most direct search methods, one is, at each step of the algorithm merely interested in the *relative* order of the values of the objective function in a number of points, i.e. one can quit calculating the component $V_i(\alpha)$ for some trial point α, as soon as it becomes clear that α is suboptimal. We recall that when solving the MDP via value-iteration, both in the discounted model (cf. MACQUEEN [19], PORTEUS [23]) and in the undiscounted unichain case (cf. ODONI [22]) an upper bound on $V_i(\alpha)$ may be calculated that converges to $V_i(\alpha)$ as the number of iterations tends to infinity. Hence, suboptimality of any point α may be detected after a finite number of steps, after which the search procedure may be continued by starting the evaluation of the objective function in (1.6) for a different choice of α.

The above considerations lead to a proposal for solving the entire problem (1.6) by a single value-iteration scheme in which the parameters $q_i^k(\cdot)$ and $P_{ij}^k(\cdot)$ are adapted in accordance with the search procedure, and ultimately converge to the parameter values corresponding with the optimal value of α.

Note that most direct search methods have the property of locating the optimum at a *geometric* rate, so that in general the approximations for the parameters $q_i^k(\cdot)$ and $P_{ij}^k(\cdot)$ will converge to the desired values at a geometric rate as well (cf. the proposition on p. 130 in LUENBERGER [18]).

For a more detailed description of the proposal method we refer to the appendix.

EXAMPLE 3. Solving nested sequences of (piecewise linear) functional equations where each functional (vector)-equation has the structure of the optimality equation of an undiscounted MDP or Markov Renewal Program.

$$(1.8)\qquad x(0)_i = \max_{k\in K^0(i)} [a_i^k(0) + \sum_j P_{ij}^k\, x(0)_j], \qquad i \in \Omega$$

$$\vdots$$

$$x(m)_i = \max_{k\in K^m(i)} [a_i^k(m) + \sum_j P_{ij}^k\, x(m)_j], \qquad i \in \Omega$$

$$\vdots$$

$$x(r)_i = \max_{k\in K^r(i)} [a_i^k(r) + \sum_j P_{ij}^k\, x(r)_j], \qquad i \in \Omega$$

where $K^r(i) \subseteq \ldots \subseteq K^m(i) \subseteq K^0(i)$ and where the quantities $a_i^k(m)$ and the sets $K^m(i)$ both depend upon $x(0),\ldots,x(m-1)$ i.e. upon the solution of the first m functional equations in the sequence (1.7). A sequence of nested equations of this type occurs e.g. when trying to find the maximal gain rate vector or some of the higher terms in the Laurent series expansion of the maximal total discounted return vector in powers of the interest rate; and accordingly, when trying to locate maximal gain policies or policies that are optimal under more selective (sensitive discounted or average overtaking) optimality criteria (cf. VEINOTT [30], MILLER and VEINOTT [20], DENARDO [5]). For a

more detailed specification of the sequence (1.7) and for a characterization of the solution set, we refer to FEDERGRUEN and SCHWEITZER [9].

In view of the dependence of the sets $K^m(i)$ and the quantities $a_i^k(m)$ on the solution to the previous m equations in (1.7), one conceivable way of solving the m+1-st equation, is by computing these sets and quantities beforehand with the help of an *exact* solution method (Linear Programming, or the Policy Iteration Algorithm, cf. DENARDO [5] and VEINOTT [30]). However, when the state space becomes large, exact solution methods become infeasible, and a successive approximation method is needed to solve the entire system; moreover *exact* decomposition methods like Denardo's LP-method (cf. [5]) may be unstable under numerical errors, since an inexact solution to one LP may render the subsequent LP's infeasible (cf. [9]). Such a successive approximation method was recently obtained by the authors in [9], where a sequence of value-iteration schemes is simultaneously generated in order to solve the entire system of equations (1.7). The schemes that aim at finding a solution to the m+1-st equation, have $a_i^k(m)$ and the sets $K^m(i)$ replaced by approximating sequences $\{a_i^k(m)[n]\}_{n=1}^{\infty}$ and $\{K^m(i)[n]\}_{n=1}^{\infty}$ which are distilled from the schemes that aim at finding a solution to the previous equations, and which have the property of converging to the correct quantities and sets.

All of the schemes involved may be interpreted as value-iteration schemes for undiscounted MDP's, the *parameters of which are replaced by approximating sequences*.

Moreover, here again, the sequences $\{a_i^k(m)[n]\}_{n=1}^{\infty}$ may be constructed in such a way that

$$(1.9) \qquad a_i^k(m)[n] \to a_i^k(m), \qquad \textit{geometrically as } n \to \infty;\ i \in \Omega,\ k \in K^0(i),\ m = 0,\ldots,r.$$

and the successive approximation method can be shown to converge to a solution of the entire system (1.7) at a *geometric* rate as well.

In the classical case where all of the parameters and action sets of the MDP are perfectly known and available, the following value-iteration scheme has proven to be extremely useful, both for the discounted and the undiscounted model:

(1.9) $\qquad v(n+1)_i = Qv(n)_i; \qquad i \in \Omega;\ n = 0,1,\ldots$

where $Qx_i = \max_{k\in K(i)}[q_i^k + \beta \Sigma_{j=1}^N P_{ij}^k x_j]$, $i \in \Omega$; and $0 < \beta \leq 1$. The case $\beta < 1$ corresponds with the discounted model where a discount factor β is applied whereas the case $\beta = 1$ corresponds with the undiscounted model.

The literature on the asymptotic behaviour of (1.9) for the *discounted* model goes back to SHAPLEY [29]. Using contraction mapping arguments, one can show that convergence is guaranteed at a *geometric* rate (cf. DENARDO [4]).

For the undiscounted model, convergence of value-iteration was first studied by WHITE [33], BROWN [2], SCHWEITZER [25] and LANERY [17]. BROWN [2] showed e.g. that $\{v(n) - ng^*\}_{n=1}^{\infty}$ is always bounded, where g^* represents the maximal gain rate vector. In [27] the authors derived the necessary and sufficient condition for $\{v(n) - ng^*\}_{n=1}^{\infty}$ to converge for *all* $v(0) \in E^N$, and in [28] we showed that the rate of convergence is geometric, whenever convergence occurs. Finally SCHWEITZER [26] proposed a data-transformation which enforces convergence of the value-iteration method for every possible choice of $v(0) \in E^N$. For a survey on the subject we refer to [8]. In case only approximations of the parameters and action sets are available, it seems natural to consider the following iterative scheme:

(1.10) $\qquad x(n+1)_i = Q(n)x(n)_i; \qquad i \in \Omega;\ n = 0,1,\ldots$

where

(1.11) $\qquad Q(n)x_i = \max_{k\in K(i,n)} [q_i^k(n) + \beta \sum_{j=1}^{N} P_{ij}^k(n)x_j], \qquad i \in \Omega$

and with $x(0) \in E^N$ arbitrarily chosen. That is, we modify the classical value iteration method merely in the sense that at each iteration, the unknown data of the problem are replaced by their current guesses.

For the *discounted* version of the model, geometric convergence of $\{x(n)\}_{n=1}^{\infty}$ can easily be obtained, as is briefly shown in section 3. No assumptions are made with respect to the chain and periodicity structure or with respect to the type of convergence in (1.2) and (1.3). For the undiscounted version, we henceforth assume:

(H) $\{q_i^k(n)\}_{n=1}^{\infty} \to q_i^k$, *geometrically*; $i \in \Omega$, $k \in K(i)$,

$\{P_{ij}^k(n)\}_{n=1}^{\infty} \to P_{ij}^k$, *geometrically*; $i \in \Omega$, $k \in K(i)$

which was satisfied in all of our examples. In section 4, we describe the asymptotic behaviour of the sequence $\{x(n)\}_{n=1}^{\infty}$ showing the interdependence with the behaviour of the *stationary* scheme $\{v(n)\}_{n=1}^{\infty}$. As a side result we obtain in section 5 the asymptotic behaviour of backwards products of a geometrically convergent sequence of Markov chains. The appendix, finally, specifies our algorithm for the models, mentioned in example 2. First however we give in section 2, some notation and preliminaries.

2. NOTATION AND PRELIMINARIES

A (stationary, pure) policy f is a vector $[f(1),\ldots,f(N)]$ with $f(i) \in K(i)$. With each policy f, we associate a transition probability matrix *(tpm)* $P(f)$; a reward vector $q(f)$ as well as their approximating sequences $\{P(f;n)\}_{n=1}^{\infty}$ and $\{q(f;n)\}_{n=1}^{\infty}$:

(2.1) $P(f)_{ij} = P_{ij}^{f(i)}$; $P(f,n)_{ij} = P_{ij}^{f(i)}(n)$; $i,j \in \Omega$; $n = 1,2,\ldots$

$q(f)_i = q_i^{f(i)}$; $q(f,n)_i = q_i^{f(i)}(n)$; $i \in \Omega$; $n = 1,2,\ldots$

The stochastic matrix $\Pi(f)$ denotes the Cesaro-limit of the sequence $\{P^n(f)\}_{n=1}^{\infty}$, with $P^n(f)$ the n-th power of $P(f)$. We recall

(2.2) $\lim_{n\to\infty} P^n(f) = \Pi(f)$ if and only if $P(f)$ is aperiodic

Let $R(f) = \{j \mid \Pi(f)_{jj} > 0\}$ represent the set of recurrent states under $P(f)$.

In the discounted version of the model, with discount factor $0 < \beta < 1$ we associate with each policy $f \in X_{i=1}^{N} K(i)$, the total discounted return vector

$$v(f,\beta) = \sum_{n=0}^{\infty} \beta^n P^n(f) q(f) = [I - \beta P(f)]^{-1} q(f)$$

The total maximal discounted return vector v^*, defined by

$$v_i^* = \max_{f \in X_i K(i)} v(f,\beta)_i; \qquad i \in \Omega$$

is the unique solution to the functional equation

$$(2.3) \qquad v_i = \max_{k \in K(i)} [q_i^k + \beta \sum_j P_{ij}^k v_j], \qquad i \in \Omega$$

and a policy f is optimal if and only if it achieves the maximum on the right hand side of (2.3) for all $i \in \Omega$ (cf. JEWELL [16]). Note that each one of the Q(n)-operators satisfies the property:

$$(2.4.a) \qquad \beta[x-y]_{min} \leq [Q(n)x - Q(n)y]_{min} \leq [Q(n)x - Q(n)y]_{max} \leq \beta[x-y]_{max}$$

so that

$$(2.4.b) \qquad \|Q(n)x - Q(n)y\| \leq \beta \|x-y\|$$

where for all $x \in E^N$, $x_{max} = \max_i x_i$; $x_{min} = \min_i x_i$ and $\|x\| = \max|x|_i$. In the undiscounted model, we associate with each policy $f \in X_i K(i)$, the gain rate vector

$$g(f) = \Pi(f)q(f) = \lim_{n\to\infty} \frac{1}{n+1} \sum_{\ell=0}^{n} P^{\ell}(f)q(f)$$

Let the maximal gain rate vector be denoted by g^*:

$$(2.5) \qquad g_i^* = \max_{f \in X_i K(i)} g(f)_i, \qquad i \in \Omega.$$

We refer e.g. to [6] for the existence of policies f which attain the N maxima in (2.5) simultaneously. Such policies are called *maximal gain*. In the *undiscounted* model, the following optimality equation arises:

$$(2.6) \qquad v_i + g_i^* = Tv_i; \qquad i \in \Omega$$

where

$$(2.7)\qquad Tx_i = \max_{k\in L(i)} [q_i^k + \sum_j P_{ij}^k x_j]; \qquad i \in \Omega$$

with $L(i) = \{k \in K(i) \mid g_i^* = \sum_j P_{ij}^k g_j^* = \max_{\ell\in K(i)} \sum_j P_{ij}^\ell g_j^*\}, \qquad i \in \Omega$

(c.f. e.g. [6]). Let $V = \{v \in E^N \mid v \text{ satisfies } (2.6)\}$. Fix a solution $v \in V$, and define

$$(2.8)\qquad b(v)_i^k = q_i^k - g_i^* + \sum_j P_{ij}^k v_j - v_i; \qquad i \in \Omega;\ k \in K(i),$$

$$(2.9)\qquad L(i,v) = \{k \in L(i) \mid b(v)_i^k = 0 = \max_{\ell\in L(i)} b(v)_i^\ell\}; \qquad i \in \Omega$$

A policy f is maximal gain if and only if (cf. DENARDO [5])

$$(2.10.a)\qquad f(i) \in L(i), \qquad i \in \Omega;$$

$$(2.10.b)\qquad f(i) \in L(i,v), \qquad i \in R(f)$$

In particular we have for any $v \in V$, that $S^*(v) \overset{\text{def}}{=\!=} X_{i=1}^N L(i,v)$ is a subset of the set of maximal gain policies. Finally, we define for any solution $v \in V$, the operator U(v) by:

$$(2.11)\qquad U(v)x_i = \max_{k\in L(i,v)} [\sum_j P_{ij}^k x_j], \qquad i \in \Omega;\ x \in E^N.$$

3. THE DISCOUNTED MODEL

In this section, we consider the iterative scheme (1.10) with $\beta < 1$

THEOREM 1. $\{x(n)\}_{n=1}^\infty \to v^*$, *geometrically, where* v^* *is the unique solution to* (2.3).

PROOF. Let M be such that $|q_i^k(n)| \le M$ for all $i \in \Omega$, $k \in K(i)$, $n = 1,2,\ldots$ where $M < \infty$ follows from (1.2). Verify that $\|x(n)\| \le M \sum_{\ell=0}^{n-1} \beta^\ell + \beta^n \|x(0)\| \le M(1-\beta)^{-1} + \|x(0)\|$ for all $n \ge 1$, and conclude that $\{x(n)\}_{n=1}^\infty$ is a bounded sequence. Let $\{x(n_k)\}_{k=1}^\infty$ and $\{x(m_k)\}_{k=1}^\infty$ be two convergent subsequences with resp. limit vectors v^o and v^{oo}. It is no restriction to assume that $n_k > m_k$ for all $k \ge 1$.

Apply (2.4.b) repeatedly to conclude that

$$(3.1) \qquad \|x(n_k) - x(m_k)\| \leq \beta^{m_k} \|x(n_k - m_k) - x(0)\| ; \qquad k = 1,2,\ldots$$

Let k tend to infinity, noting that the second factor to the right of (3.1) is bounded, to conclude that $\|v^o - v^{oo}\| = 0$.

Hence $\{x(n)\}_{n=1}^{\infty}$ converges and its limit vector satisfies the optimality equation (2.3), which implies $\lim_{n\to\infty} x(n) = v^*$. Finally to show that the rate of convergence is geometric, replace m_k in (3.1) by a fixed integer m, and let k tend to infinity so as to conclude that

$$(3.2) \qquad \|x(m) - v^*\| \leq \beta^m \|v^* - v(0)\| . \qquad \square$$

The set of all optimal policies can be obtained in the same way as in the stationary model (cf. [8], section 3). In case the parameters q_i^k and P_{ij}^k are approached from below and from above, all of the bounds on v_i^*, stopping criteria for ε-approximations or ε-optimal policies, as well as tests for eliminating suboptimal actions, that were found for the stationary model, can be adapted in a straightforward manner.

4. THE UNDISCOUNTED MODEL

The characterization of the asymptotic behaviour of $\{x(n)\}_{n=1}^{\infty}$ in the *undiscounted* model is more complicated, since the Q-operator loses its contraction-properties (cf. e.g. (2.5.b)) when $\beta = 1$. In fact, the easily verified property $Q(x + c\underline{1}) = Qx + c\underline{1}$ for all $x \in E^N$ and scalars c (with $\underline{1}$ the N-vector of ones), shows that the Q-operator is not a contraction mapping on E^N, nor a J-step contraction mapping for any $J \geq 1$ (cf. DENARDO [4] and [10]). In a sequence of papers (cf. [8], [27], [28]) the authors described the asymptotic behaviour of $\{v(n)\}_{n=1}^{\infty}$ (cf. (1.9)) to which the sequence $\{x(n)\}_{n=1}^{\infty}$ reduces in the *stationary* case where there is perfect knowledge of the parameters and action sets in the model. The following theorem extends the theory to the *non-stationary* case under consideration. First however we need the following definitions.

Let $K > 0$ and $0 \leq \lambda < 1$ be such that (cf. assumption (H)):

$$(4.1)\qquad |q_i^k(n) - q_i^k| \le K\lambda^n; \quad |P_{ij}^k(n) - P_{ij}^k| \le K\lambda^n; \quad n = 0,1,\ldots \quad i,j \in \Omega \text{ and } k \in K(i)$$

For each $n = 1,2,\ldots$ and $\varepsilon \ge 0$, let

$$(4.2)\qquad S(n,\varepsilon) = \{f \in X_{i=1}^N K(i) \mid q(f,n) + P(f,n)x(n) \ge x(n+1) - \varepsilon\underline{1}\}$$

be the set of policies that come within ε of attaining the maxima at the n+1-st iteration of the scheme (1.10). We use $S(n)$ as a shorthand notation for $S(n,0)$.

A *randomized* (stationary)policy f is specified by the tableau $[f_{ik}]$ satisfying $f_{ik} \ge 0$ and $\sum_{k\in K(i)} f_{ik} = 1$ $(i \in \Omega,\ k \in K(i))$ such that f_{ik} denotes the *probability* with which the k-th alternative is chosen when entering state i. Thus, the *pure* policies share the special characteristic of having all of the numbers f_{ik} equal to 0 or 1. Next let $R^* = \{i \in \Omega \mid i \in R(f)$ for some randomized maximal gain policy $f\}$. We recall from lemma 2.1. part (b) of [27] that

$$(4.3)\qquad \{f \mid f \text{ is maximal gain and } R(f) = R^*\} \neq \emptyset.$$

Finally it was pointed out in [27] that the integer

$$(4.4)\qquad J^* = \min\{J \ge 1 \mid P(f)^J \text{ is aperiodic for some randomized maximal gain policy } f \text{ with } R(f) = R^*\}$$

plays a crucial part in the description of the asymptotic behaviour of $\{v(n)\}_{n=1}^{\infty}$, the sequence of iterates in the *stationary* model. In addition, a number of alternative characterizations, as well as a finite algorithm for the computation of J^* were given in [27]. The following theorem shows that most of the results with respect to the asymptotic behaviour of value iteration may be extended to the non-stationary model.

THEOREM 2.

(a) $\{x(n) - ng^*\}_{n=1}^{\infty}$ *is bounded and* $S(n) \subseteq X_i L(i)$ *for all* n *sufficiently large;*

(b) *If* $\lim_{n\to\infty} [x(n) - ng^*]$ *exists, let* v *be this limit vector. Then* $v \in V$ *and* $x(n) - ng^* - v \to 0$ *geometrically as* $n \to \infty$;

(c) *If* $\lim_{n\to\infty} [Q^n y - ng^*]$ *exists for every* $y \in E^N$, *then* $\lim_{n\to\infty} [x(n) - ng^*]$ *exists for every* $x(0) \in E^N$.

(d) *Sufficient conditions for the existence of* $\lim_{n\to\infty} [x(n) - ng^*]$ *for every* $x(0) \in E^N$ *are:*

(1) *every pure (maximal gain) policy* f, *has an aperiodic* tpm.

(2) $J^* = 1$.

(e) $\lim_{n\to\infty} \lim_{m\to\infty} [Q(n+m)\ldots Q(n)x - ng^*]$ *exists for every* $x(0) \in E^N$ *if and only if* $J^* = 1$. *Moreover, if this condition holds*

$$\lim_{n\to\infty} \lim_{m\to\infty} [Q(n+m)\ldots Q(n)x - mg^*] = \lim_{m\to\infty} Q^m x - mg^* \in V$$

where the outer limit is approached geometrically as well.

(f) $\lim_{n\to\infty} [x(nJ^*+r) - (nJ^*+r)g^*]$ *exists for every* $x(0) \in E^N$ *and* $r = 1,\ldots,J^*$; *moreover, the limit is approached geometrically.*

(g) *If* $v = \lim_{n\to\infty} x(n) - ng^*$ *exists for some* $x(0) \in E^N$, *then* $\lim_{n\to\infty} S(n,\varepsilon_n) = S^*(v)$, *provided that* $\{\varepsilon_n\}_{n=1}^{\infty} \downarrow 0$, *where the rate of convergence of* ε_n *is slower than geometric, i.e.* $\lim_{n\to\infty} \varepsilon_n \lambda^{-n} = \infty$ *for all* $0 \le \lambda < 1$. *Take e.g.* $\varepsilon_n = n^{-1}$ *(or the reciprocal of any positive polynomial in* n*).*

PROOF. First fix $v^0 \in V$ and let $e(n) = x(n) - ng^* - v^0$. Let $n_0 \ge 1$ be such that $K(i,n) = K(i)$ for all $n \ge n_0$.

(a) The proof of this part is related to the one given in th. 5.1 of [27]. Fix $f \in S^*(v^0)$ (cf. (2.8)). Then in view of $L(i,v^0) \subseteq L(i)$, $i \in \Omega$:

$$x(n+1)_i - (n+1)g_i^* - v_i^0 \ge q(f;n)_i + \sum_j P(f;n)_{ij}\{x(n)_j - ng_j^* - v_j^0\} - q(f)_i + \sum_j [P(f;n)_{ij} - P(f)_{ij}][ng_j^* + v_j^0] \tag{4.5}$$

i.e.

$$(4.6)\qquad e(n+1)_i \geq -K\lambda^n - NK\lambda^n(n\|g^*\|+\|v^0\|) + \sum_j P(f;n)_{ij}e(n)_j$$

$$\geq -K\lambda^n - NK\lambda^n(n\|g^*\|+\|v^0\|) + e(n)_{min}.$$

By iterating (4.6) n times, we obtain for all $i \in \Omega$:

$$(4.7)\qquad e(n+1)_i \geq -K(1+N\|v^0\|) \sum_{\ell=n_0}^{n} \lambda^{\ell} - KN\|g^*\| \sum_{\ell=n_0}^{n} \ell\lambda^{\ell} + e(n_0)_{min}$$

$$\geq \frac{-K(1+N\|v^0\|)}{(1-\lambda)} - \frac{KN\|g^*\|\lambda}{(1-\lambda)^2} + e(n_0)_{min}$$

To show that $\{e(n)\}_{n=1}^{\infty}$ is bounded from above as well, let

$$a_i^k = \sum_j P_{ij}^k\, g_j^* - g_i^*; \qquad i \in \Omega,\ k \in K(i)$$

and note $\max_{k\in K(i)} a_i^k = 0$. Finally use (4.5) and (2.8) to conclude:

$$(4.8)\qquad e(n+1)_i = \max_{k\in K(i)} \{na_i^k + b(v^0)_i^k + \sum_j P_{ij}^k(n)e(n)_j$$

$$+ \sum_j [P_{ij}^k(n) - P_{ij}^k](v_j^0 + ng_j^*) + q_i^k(n) - q_i^k\}$$

$$\leq \max_{k\in K(i)} \{na_i^k + b(v^0)_i^k\} + e(n)_{max}$$

$$+ NK\lambda^n(\|v^0\| + n\|g^*\|) + K\lambda^n$$

Next use $a_i^k < 0$ for $k \in K(i)\backslash L(i)$, $i \in \Omega$ to conclude that there exists an integer $n_1 \geq n_0$ such that for all $n \geq n_1$ the first term to the right of the inequality (4.8) is achieved for $k \in L(i)$ and hence vanishes (cf. (2.8) and (2.9)). By iterating (4.8) one concludes that for all $n \geq n_1$:

$$(4.9)\qquad e(n+1)_{max} \le e(n_1)_{max} + K(N\|v^0\|+1)\sum_{\ell=n_0}^{n}\lambda^{\ell} + NK\sum_{\ell=n_0}^{n}\ell\lambda^{\ell}\|g^*\|$$

$$\le e(n_1)_{max} + K(N\|v^0\|+1)/(1-\lambda) + NK\|g^*\|\lambda/(1-\lambda)^2$$

(4.9) together with (4.7) prove the first assertion in this part. The fact that $S(n,0) \subseteq X_i L(i)$ for all n sufficiently large, then follows by considering the equality part in (4.8) and by noting, with the help of assumption (H), that both the left side and all of the terms within accolades to its right are bounded, with only the term $[na_i^k]$ as a possible exception.

(b) Subtract $(n+1)g_i^*$ from both sides of the equality (1.10) to get:

$$x(n+1)_i - (n+1)g_i^* = \max_{k\in K(i,n)} [q_i^k - g_i^* + \sum_j P_{ij}^k(n)(x(n) - ng_j^*)$$

$$+ na_i^k + n\sum_j (P_{ij}^k(n) - P_{ij}^k)g_j^*]$$

Next use the second assertion in part (a) to conclude for all n sufficiently large:

$$(4.10)\qquad x(n+1)_i - (n+1)g_i^* = \max_{k\in L(i)} [q_i^k - g_i^* + \sum_j P_{ij}^k(n)(x(n)_j - ng_j^*)$$

$$+ n\sum_j (P_{ij}^k(n) - P_{ij}^k)g_j^*], \qquad i \in \Omega$$

Finally, use assumption (H) and let n tend to infinity, to verify that $v = \lim_{n\to\infty} x(n) - ng^* \in V$.

We next prove the geometric rate of convergence of $\{e(n)\}_{n=1}^{\infty}$ towards 0. It follows from the second assertion in part (a), that for all n sufficiently large,

$$(4.11)\qquad e(n+1)_i = \max_{k\in L(i)} \{(q(n)_i^k - q_i^k) + \sum_j P_{ij}^k(n)e(n)_j$$

$$+ \sum_j (P_{ij}^k(n) - P_{ij}^k)(ng_j^* + v_j) + b(v)_i^k\}$$

Since all other terms in (4.11) approach 0 as $n \to \infty$, the last term must vanish for n sufficiently large, i.e. for large n (4.11) implies, in view of the boundedness of $\{e(n)\}_{n=1}^{\infty}$ and (2.11),

$$\|e(n+1) - U(v)e(n)\| \leq (A_1 + A_2 n)\lambda^n \tag{4.12}$$

for certain positive constants A_1 and A_2. Note as a special case of (2.4.b) that $\|U(v)x - U(v)y\| \leq \|x-y\|$ for all $x,y \in E^N$, and use this inequality to conclude:

$$\begin{aligned} \|e(n+m) - U(v)^m e(n)\| &\leq \sum_{k=1}^{m} \|U(v)^{m-k} e(n+k) - U(v)^{m-k+1} e(n+k-1)\| \\ &\leq \sum_{k=1}^{m} \|e(n+k) - U(v)e(n+k-1)\| \\ &\leq \sum_{\ell=n}^{n+m-1} (A_1 + A_2\ell)\lambda^{\ell} \leq \sum_{\ell=n}^{\infty} (A_1 + A_2\ell)\lambda^{\ell} = \\ &= \{A_1(1-\lambda)^{-1} + nA_2(1-\lambda)^{-1} + \lambda A_2(1-\lambda)^{-2}\}\lambda^n \end{aligned} \tag{4.13}$$

for any $m \geq 1$ and all large n.

Observe that the U(v)-operator has all of the properties of the T-operator, and hence it follows from th. 5.1. part (d) in [27] that there exists a number $\hat{J} \geq 1$ such that for all $x \in E^N$ (in fact, a close inspection of [27] shows that $\hat{J}$ may be taken to be equal to J^*):

$$U(v)^{\infty}x = \lim_{m\to\infty} U(v)^{m\hat{J}}x \quad \text{exists.} \tag{4.14}$$

Moreover, it then follows from th. 6.1. part (b) in [28] that there exists a number $0 \leq \Gamma < 1$ such that for all $x \in E^N$ and $n \geq 1$:

$$\|U(v)^{n\hat{J}}x - U(v)^{\infty}x\| \leq \Gamma^n \|x - U(v)^{\infty}x\| \tag{4.15}$$

Next we replace m by $m\hat{J}$ in (4.13) and let m tend to infinity, to conclude

(4.16) $\quad \|U(v)^{\infty}e(n)\| \leq (A_3 + A_4 n)\lambda^n, \qquad$ for certain $A_3, A_4 > 0$

and all large n. Finally use (4.13), (4.15) and (4.16), with n replaced by n + r and $m = n\hat{J}$ to obtain for all $r = 0,\ldots,\hat{J}-1$ and large n:

(4.17) $$\|e(n(\hat{J}+1) + r)\| \leq \|e(n(\hat{J}+1) + r) - U(v)^{n\hat{J}}e(n+r)\|$$

$$+ \|U(v)^{n\hat{J}}e(n+r) - U(v)^{\infty}e(n+r)\| + \|U(v)^{\infty}e(n+r)\|$$

$$\leq [\lambda^r A_1(1-\lambda)^{-1} + \lambda^r(1-\lambda)^{-1}A_2(n + r + \lambda(1-\lambda)^{-1})]\lambda^n$$

$$+ \|e(n+r) - U(v)^{\infty}e(n+r)\|\Gamma^n + [\lambda^r(A_3 + A_4 r + A_4 n)]\lambda^n$$

Since $n\lambda^n < (\frac{1+\lambda}{2})^n$ for large n, and since $\|e(n+r) - U(v)^{\infty}e(n+r)\| \leq \|e(n+r)\| + \|U(v)^{\infty}e(n+r)\| \leq A_5$ for some constant $A_5 > 0$ (cf. (4.16)) and all $n \geq 1$, this shows that $\|e(n)\|$ goes to zero as least as fast as $A_6\,\rho^n$ where

$$\rho = \sqrt[\hat{J}+1]{\max(\Gamma, \tfrac{1+\lambda}{2})} < 1$$

and $A_6 = A_5 + (A_3 + \hat{J}A_4) + (1-\lambda)^{-1}(A_1 + \hat{J}A_2 + (1-\lambda)^{-1}A_2)$.

(c) Let $y(n) = x(n) - ng^*$. Use (4.10) and the boundedness of $\{y(n)\}_{n=1}^{\infty}$ (cf. (part (a)) to observe that there exist constants $B_1, B_2 > 0$ such that for all n sufficiently large:

(4.18) $$y(n+1) \leq Ty(n) - g^* + (B_1 + B_2 n)\lambda^n$$

$$y(n+1) \geq Ty(n) - g^* - (B_1 + B_2 n)\lambda^n$$

where the T-operator was defined in (2.7). Use the monotonicity property of the T-operator, while iterating the inequalities in (4.18) and conclude for all n sufficiently large and $m \geq 1$:

(4.19) $$\|y(n+m) - [T^m y(n) - mg^*]\| \leq \sum_{\ell=n}^{n+m-1} (B_1 + B_2\ell)\lambda^{\ell}.$$

For each $x \in E^N$, let $L(x) = \lim_{m\to\infty} T^m x - mg^*$. Fix n sufficiently large for (4.19) to hold; in view of the boundedness of $\{y(n)\}_{n=1}^{\infty}$ take a subsequence $\{y(n+m_k)\}_{k=1}^{\infty}$ which converges to a limit point c (say). Replace m by m_k in (4.19) and let k tend to infinity, in order to conclude:

$$(4.20) \qquad \|c - L(y(n))\| \le \lambda^n\{B_1(1-\lambda)^{-1} + B_2 n(1-\lambda)^{-1} + B_2\lambda(1-\lambda)^{-2}\}$$

Hence we obtain for any pair c, c' of limit points of $\{y(n)\}_{n=1}^{\infty}$: $\|c-c'\| \le 2\lambda^n\{B_1(1-\lambda)^{-1} + B_2 n(1-\lambda)^{-1} + B_2\lambda(1-\lambda)^{-2}\}$ and finally let n tend to infinity, to conclude that all limit points of the sequence $\{y(n)\}_{n=1}^{\infty}$ coincide, i.e. $\lim_{n\to\infty} x(n) - ng^*$ exists for every $x(0) \in E^N$.

(d) Part (d) follows by combining part (c) with th. 5.1. part (d) and th. 5.5 (IV) and (V) in [27].

(e) In view of part (d), we first prove the "only if" part. Fix $x \in E^N$. In view of the sequence $\{Q^m x - mg^*\}_{m=1}^{\infty}$ being bounded, let $R = \sup_m \|Q^m x - mg^*\|$. We first show by complete induction with respect to m that for all $n \ge n_0$ and $m \ge 1$:

$$(4.21) \qquad Q(m+n-1)\ldots Q(n)x_i \le Q^m x_i + \sum_{k=n}^{n+m-1} K\lambda^k(NR+1) + NK\sum_{k=0}^{m-1} k\lambda^{n+k}\|g^*\|$$

$$Q(m+n-1)\ldots Q(n)x_i \ge Q^m x_i - \sum_{k=n}^{n+m-1} K\lambda^k(NR+1) - NK\sum_{k=0}^{m-1} k\lambda^{n+k}\|g^*\|$$

Note that for m = 1, $Q(n)x_i = \max_{k\in K(i)} \{q_i^k(n) + \sum_j P_{ij}^k(n)x_j\} \le Qx_i + \|q_i^k - q_i^k(n)\| + \|\sum_j (P_{ij}^k(n) - P_{ij}^k)x_j\| \le Qx_i + K\lambda^n + K\lambda^n NR$, thus proving the first inequality in (4.21) for m = 1, the proof of the second inequality being analogous. Next, assume (4.21) holds for some integer m, and observe that in view of the monotonicity of the Q(m+n)-operator,

$$Q(m+n)\ldots Q(n)x_i \le Q(m+n)(Q^m x)_i + \sum_{k=n}^{n+m-1} k\lambda^k(NR+1) +$$

$$+ NK\sum_{k=0}^{m-1} k\lambda^{n+k}\|g^*\| \le$$

$$\leq Q^{m+1}x_i + \|q_i^k(n+m) - q_i^k\| + \|\sum_j (P_{ij}^k(n+m) - P_{ij}^k)(Q^m x_j - mg_j^*)\|$$

$$+ m\|\sum_j (P_{ij}^k(n+m) - P_{ij}^k)g_j^*\| + \sum_{k=n}^{n+m-1} K\lambda^k(NR+1) + NK \sum_{k=0}^{m-1} k\lambda^{n+k}\|g^*\|$$

$$\leq Q^{m+1}x_i + \sum_{k=n}^{n+m} K\lambda^k(NR+1) + NK \sum_{k=0}^{m} k\lambda^{n+k}\|g^*\|$$

which proves the first inequality in (4.21) for m+1, the proof of the second inequality being analogous.

Now, in case $J^* \geq 2$, it follows from theorem 5.3 in [27] that there exists a vector $y \in E^N$, for which two subsequences $\{Q^{m_k}y - m_k g^*\}_{k=1}^{\infty}$ and $\{Q^{r_k}y - r_k g^*\}_{k=1}^{\infty}$ converge to two distinct limit points c and c' (say). Fix $i^{k=1}$ and suppose $c_i < c_i'$.

In view of (4.21) it then follows that for all k and n sufficiently large:

$$Q(m_k+n-1)\ldots Q(n)x_i - m_k g_i^* \leq 2/3\ c_i + 1/3\ c_i', \quad \text{whereas}$$

$$Q(r_k+n-1)\ldots Q(n)x_i - r_k g_i^* \geq 1/3\ c_i + 2/3\ c_i'$$

thus showing that for all n sufficiently large, $\{Q(m+n-1)\ldots Q(n)y - mg^*\}_{m=1}^{\infty}$ fails to converge. To prove the "if" part of the first assertion, as well as the second assertion, subtract mg_i^* from both sides of both inequalities in (4.21) and let n tend to infinity, invoking part (d).

(f) Fix an integer $J \geq 2$, and observe that

$$(4.22) \qquad Q^J x_i = \max_{\xi \in \tilde{K}(i)} \{\tilde{q}_i^{\xi} + \sum_j \tilde{P}_{ij}^{\xi} x_j\} \qquad \text{where}$$

$$\tilde{K}(i) = \{(f^1,\ldots,f^J) \mid f^1,\ldots,f^J \in X_i K(i)\}$$

$$\tilde{q}_i^{\xi} = q(f^1)_i + P(f^1)q(f^2)_i + \ldots + P(f^1)\ldots P(f^{J-1})q(f^J)_i,$$

$$i \in \Omega,\ \xi = (f^1,\ldots,f^J) \in \tilde{K}(i)$$

$$\tilde{P}_{ij}^{\xi} = P(f^1)\ldots P(f^J)_{ij}; \qquad 1 \leq i,j \leq N \text{ and } \xi = (f^1,\ldots,f^J) \in \tilde{K}(i).$$

Let $\tilde{Q} = Q^J$, and define a related "J-step MDP", denoted by a tilde, with Ω as its state space, $\tilde{K}(i)$ as the (finite) set of alternatives in state $i \in \Omega$, $\tilde{q}_i^{\xi}$ as the one-step expected reward and $\tilde{P}_{ij}^{\xi}$ as the transition probability to state j, when alternative $\xi \in \tilde{K}(i)$ is chosen when entering state i. Observe that for each $n \geq 1$, the operator $\tilde{Q}(n) \stackrel{\text{def}}{=\!=} Q(n+J)\ldots Q(n+1)$ satisfies:

$$(4.23) \qquad \tilde{Q}(n)x_i = \max_{\xi \in \tilde{K}(i,n)} \{\tilde{q}_i^{\xi}(n) + \sum_j \tilde{P}_{ij}^{\xi}(n)x_j\}, \qquad i \in \Omega$$

with $\{\tilde{K}(i,n)\}_{n=1}^{\infty}$, $\{\tilde{q}_i^{\xi}(n)\}_{n=1}^{\infty}$ and $\{\tilde{P}_{ij}^{\xi}(n)\}_{n=1}^{\infty}$ satisfying (1.2) - (1.4) and (H). Part (f) then follows by applying part (d) and th. 5.1 part (d) in [27].

(g) Note that for all $f \in S(n,\varepsilon_n)$ and n sufficiently large (cf. part (a)):

$$(4.24) \qquad q(f)_i - g_i^* + \sum_j P(f)_{ij}v_j - v_i = q(f;n)_i + \sum_j P(f;n)_{ij}x(n)_j$$

$$- x(n+1)_i + B_n \geq -\varepsilon_n + B_n$$

where

$$B_n = [q(f)_i - q(f;n)_i] + \sum_j [P(f)_{ij} - P(f;n)_{ij}][v_j + ng_j^*]$$

$$+ \sum_j P(f;n)_{ij}[v_j - x(n)_j + ng_j^*] - [v_i - x(n+1)_i + (n+1)g_i^*].$$

Note that as both ε_n and B_n tend to 0 as n tends to infinity, it follows that $S(n,\varepsilon_n) \subseteq S^*(v)$ for all n sufficiently large. To show the reversed inclusion note in analogy to (4.24) that for all $f \in S^*(v)$ and n sufficiently large:

$$q(f;n)_i + \sum_j P(f;n)_{ij}x(n)_j - x(n+1)_i = q(f)_i - g_i^* + \sum_j P(f)_{ij}v_j -$$

$$- v_i - B_n = -B_n \geq -\varepsilon_n$$

where the inequality $-B_n \geq -\varepsilon_n$ for sufficiently large n follows from $\{B_n\}_{n=1}^{\infty} \to 0$, geometrically, in view of part (b) and assumption (H), as well

as from the restriction on the sequence $\{\varepsilon_n\}_{n=1}^{\infty}$. □

In the stationary model, it is known (cf. th. 5.3 in [27]) that $J^* = 1$ occurs both as a sufficient and a necessary condition for $\{Q^n x - ng^*\}_{n=1}^{\infty}$ to converge for *all* $x \in E^N$. In the non-stationary model, $J^* = 1$, may fail to be a necessary condition due to irregularities appearing in the first couple of iterations. This is exhibited by example 4 below.

EXAMPLE 4. Let the policy space be a singleton $\{f\}$, $\Omega = \{1,2\}$, $P(f) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $q(f) = \underline{0}$. Note that $J^* = 2$ and $V = \{c\underline{1} \mid c \in E^1\}$. Next define $P(f;1)_{ij} = \delta_{i2}$ $(i,j = 1,2)$ and $q(f;n) = q(f) = 0$. Finally let $P(f;n) = P(f)$ for $n \geq 2$. Note that $Q(1)$ maps E^2 into V, so that $\lim_{n\to\infty} x(n) - ng^*$ exists for every starting point, in spite of $J^* = 2$.

Part (e) of the previous theorem shows how, $J^* = 1$, reappears as a necessary *and* sufficient condition for convergence in the non-stationary model for all possible choices of the scrap value vector. Whereas in the stationary model, convergence of $\{v(n) - ng^*\}_{n=1}^{\infty}$ will always occur for *some* $v(0) \in E^N$, this property may again be lost in the non-stationary model, as is exhibited by example 5 below:

EXAMPLE 5. Consider the MDP, as specified in example 4, merely changing $q(f;1) = [0,1]$. Note that $Q(1)$ maps E^2 into $E^2 \backslash V$ whereas $\{Q^n x - ng^*\}_{n=1}^{\infty} = \{P(f)^n x\}_{n=1}^{\infty}$ fails to converge for all $x \in E^2 \backslash V$. We conclude that $\{x(n) - ng^*\}_{n=1}^{\infty}$ fails to converge for *all* $x(0) \in E^N$.

In case of convergence of $\{x(n) - ng^*\}_{n=1}^{\infty}$, $g(n) = x(n) - x(n-1)$ and $w(n) = nx(n-1) - (n-1)x(n)$ will provide two sequences that converge to g^*, and some $v \in V$. We refer to [8], sections 4 and 5 for various techniques in the stationary model to avoid the numerical difficulty of $\{x(n)\}_{n=1}^{\infty}$ diverging linearly with n, as well as for bounds on the components of g^* and $v \in V$, and tests for permanent and temporary elimination of non-optimal actions. All of these can be extended in a straightforward manner for the non-stationary model, provided upper and lower bounds on the parameter approximations $\{q_i^k(n)\}_{n=1}^{\infty}$ and $\{P_{ij}^k(n)\}_{n=1}^{\infty}$ $(i,j \in \Omega;\ k \in K(i))$ are available.

We conclude this section by observing that the following data-transformation introduced by SCHWEITZER [26], may be used in order to enforce convergence of the value-iteration scheme (1.10) for *all* starting points $x(0) \in E^N$, in case $J^* = 1$ cannot be guaranteed to hold:

$$(4.25) \qquad \tilde{q}_i^k(n) = q_i^k(n), \qquad i \in \Omega,\ k \in K(i)$$

$$\tilde{P}_{ij}^k(n) = \tau(P_{ij}^k(n) - \delta_{ij}) + \delta_{ij}, \qquad i,j \in \Omega,\ k \in K(i)$$

where $0 < \tau < 1$. Note that $\tilde{q}_i^k(n) \to q_i^k$, geometrically ($i \in \Omega$, $k \in K(i)$) whereas $\tilde{P}_{ij}^k(n) \to \tau(P_{ij}^k - \delta_{ij}) + \delta_{ij} \stackrel{\text{def}}{=\!=} \tilde{P}_{ij}^k$, geometrically ($i,j \in \Omega$; $k \in K(i)$). Recall that the transformed MDP with Ω and $X_i K(i)$ as its state - and policy space, and $\{\tilde{q}_i^k \mid i \in \Omega,\ k \in K(i)\}$ and $\{\tilde{P}_{ij}^k \mid i,j \in \Omega \text{ and } k \in K(i)\}$ as the one-step expected rewards and transition probabilities is equivalent to the original one in the sense that it has the same gain rate vector for every policy and $\tilde{V} = \{v \in E^N \mid \tau v \in V\}$ as the set of solutions to the optimality equation (2.6).

Moreover, in the transformed model, (non-stationary) value-iteration is guaranteed to converge since all tpm's $\tilde{P}(f)$, $f \in X_i K(i)$ have all diagonal elements and are aperiodic as a consequence (cf. part (d) of the previous theorem).

Finally, a generalization of this data-transformation (cf. SCHWEITZER [26]) turns every undiscounted Markov Renewal Program (cf. JEWELL [16], DENARDO and FOX [6]) into an equivalent undiscounted MDP in which every policy is aperiodic. As a consequence, Markov Renewal Programs can be solved via non-stationary value-iteration schemes as well (the latter exhibiting all of the nice properties mentioned in theorem 2), whenever only geometric approximations for its parameters and action sets are available.

5. BACKWARDS PRODUCTS OF GEOMETRICALLY CONVERGENT SEQUENCES OF FINITE MARKOV CHAINS

In this final section, we consider backwards products $U(r,k) = P(r+k)\ldots P(r+1)$ of a sequence of finite Markov matrices $P(n)$, where

(5.1) $\quad P(n) \to P(\infty), \quad$ geometrically,

Recent papers (cf. [1] and [3]) have pointed out a number of models in which these backward products occur. Moreover they showed that $\{U(r,k)\}_{k=1}^{\infty}$ converges *geometrically* for all $r \geq 1$, in case $P(\infty)$ is aperiodic and unichained, and whatever the rate of convergence of $\{P(n)\}_{n=1}^{\infty}$ may be. In addition, FEDERGRUEN [7] has pointed out that in this case

$$(5.2) \qquad \lim_{r\to\infty} \lim_{k\to\infty} U(r,k) = \Pi$$

where $\Pi = \lim_{n\to\infty} P(\infty)^n$. Moreover convergence in (5.2) was shown to be at least as fast as the rate of convergence of $\{P(n)\}_{n=1}^{\infty}$ towards $P(\infty)$.

In this section we show as a corollary of theorem 2, that under (5.1) these results may be extended to the *multichain* case, i.e. $\lim_{r\to\infty} \lim_{k\to\infty} U(r,k)$ exists *if and only if* $P(\infty)$ is aperiodic, and the rate of convergence in (5.2) is again geometric. Related results for forward products were recently obtained in [15]. An example in [7] points out that in case $\{P(n)\}_{n=1}^{\infty}$ approaches $P(\infty)$ at a slower than geometric rate, the limit matrix in (5.2) may be different from Π. This shows that in the *multichain* case assumption (5.1) plays a crucial role to ensure convergence of the backwards products $U(r,k)$ to the correct limit matrix.

COROLLARY 3. *Assume* (5.1) *to hold.*

(a) *If* $P(\infty)$ *is aperiodic, then* $\lim_{k\to\infty} U(r,k)$ *exists, for all* $r \geq 1$ *where the rate of convergence is geometric.*

(b) *If* $P(\infty)$ *is aperiodic, then* $\lim_{r\to\infty} \lim_{k\to\infty} U(r,k) = \Pi$ *where the outer limit is approached geometrically as well*

(c) $\lim_{r\to\infty} \lim_{k\to\infty} U(r,k)$ *exists if and only if* $P(\infty)$ *is aperiodic*

(d) *If* $P(\infty)$ *has period* $J \geq 2$, *(i.e.* $P(\infty)^J$ *is aperiodic), then* $\lim_{k\to\infty} U(r,kJ+r)$ *exists for all* $r \geq 1$ *and the rate of convergence is geometric.*

PROOF. Consider the MDP which has a single policy f, state space Ω, $P(f) = P(\infty)$; $P(f;n) = P(n)$ and $q(f) = q(f;n) = 0$, $(n \geq 1)$, and apply the previous theorem to this MDP. Choose $x(0)$ as the j-th unit vector $(1 \leq j \leq N)$ in E^N,

i.e. $x(0)_i = \delta_{ij}$ $(1 \leq i \leq N)$ to establish the assertions for the j-th column of the matrix products. Apply part (b) and (d) of th. 2 to get part (a); part (e) of th. 2 to get part (b) and (c); and part (f) of th. 2 to get part (d). □

REFERENCES

[1] ANTHONISSE, J. & H. TIJMS, *Exponential convergence of products of stochastic matrices*, J.M.A.A. 59 (1977), 360-364.

[2] BROWN, B., *On the iterative method of dynamic programming on a finite state space discrete time Markov Process*, Ann. Math. Stat. 36 (1965), 1279-1285.

[3] CHATTERJEE, S. & E. SENETA, *Towards consensus: some convergence theorems on repeated averaging*, J. Appl. Prob. 14 (1977), 89-97.

[4] DENARDO, E., *Contraction Mappings in the theory underlying dynamic programming*, SIAM Rev. 9 (1967), 165-177.

[5] ———— , *Markov Renewal Programs with small interest rates*, Ann. Math. Stat. 42 (1971), 477-496.

[6] ———— & B. FOX, *Multichain Markov Renewal Programs*, SIAM J. Appl. Math. 16 (1968), 468-487.

[7] FEDERGRUEN, A., *On nonstationary Markov Chains with converging transition matrices*, Math. Center Report BW 84/77 (1977) (to appear in Stoch. Proc. & Appl.).

[8] ———— & P.J. SCHWEITZER, *Discounted and undiscounted value-iteration in Markov Decision Problems: a survey*, Math. Center Report BW 78/77 (1977) (to appear in the Proceedings of the International Conference on Dynamic Programming, Vancouver, 1977, to be published by Academic Press).

[9] ———— & ———— , *Successive approximation methods for solving nested functional equations in Markov Decision Theory*, (1977) (forthcoming).

[10] ———— & ———— & H. TIJMS, *Contraction Mappings underlying undiscounted Markov Decision Problems*, Math. Center Report BW 72/77 (1977) (to appear in J. Math. Anal. Appl.).

[11] GOFFIN, J., *On convergence rates of subgradient optimization methods*, McGill University working paper, no. 76-34 (1976).

[12] GRINOLD, R., *Elimination of suboptimal actions in Markov decision problems*, Op. Res. 21 (1973), 848-851.

[13] HASTINGS, N., *A test for nonoptimal actions in undiscounted finite Markov Decision Chains*, Man. Sci. 23 (1976), 87-92.

[14] ———— & J. MELLO, *Tests for suboptimal actions in discounted Markov Programming*, Man. Sci. 19 (1973), 1019-1022.

[15] HUANG, C., D. ISAACSON & B. VINOGRADE, *The rate of convergence of certain nonhomogeneous Markov Chains*, Z. Wahrscheinlichkeitstheorie 35, (1976) 141-146.

[16] JEWELL, W., *Markov Renewal Programming*, Op. Res. 11 (1963), 938-971.

[17] LANERY, E., *Etude asymptotique des Systèmes Markoviens à commande*, Rev. Inf. Rech. Op. 1 (1967), 3-56.

[18] LUENBERGER, D., *Introduction to linear and nonlinear programming*, Addison-Wesley Publ. Comp., Reading Massachusetts, (1973).

[19] MACQUEEN, J., *A test for suboptimal actions in Markovian Decision Problems*, Op. Res. 15 (1967), 559-561.

[20] MILLER, B. & A. VEINOTT Jr., *Discrete Dynamic Programming with a small interest rate*, Ann. Math. Stat. 40 (1969), 366-370.

[21] MURRAY, W., *Numerical methods for unconstrained optimization*, Academic Press, New York (1972).

[22] ODONI, A., *On finding the maximal gain for Markov Decision Processes*, O.R. 17 (1969), 857-860.

[23] PORTEUS, E., *Some bounds for discounted sequential decision processes*, Man. Sci. 18 (1971), 7-11.

[24] RUSSEL, C., *An optimal policy for operating a multi-purpose reservoir*, Op. Res. 20 (1972), 1181-1189.

[25] SCHWEITZER, P.J., *Perturbation Theory and Markovian Decision Processes*, Ph.D. dissertation, M.I.T. Op. Res. Center Report 15 (1965).

[25a] SCHWEITZER, P.J., *Perturbation theory and finite Markov Chains*, J. Appl. Prob. 5 (1968), 401-413.

[26] ———————— , *Iterative solution of the functional equations for undiscounted Markov Renewal Programming*, J.M.A.A. 34 (1971), 495-501.

[27] ———————— & A. FEDERGRUEN, *The asymptotic behaviour of undiscounted value iteration in Markov Decision Problems*, Math. of O.R. 2 (1978), 360-381.

[28] ———————— & ———————— , *Geometric convergence of value-iteration in multichain Markov Decision Problems*, Math. Center Report BW 80/77 (1977) (to appear in Adv. Appl. Prob.)

[29] SHAPLEY, L., *Stochastic Games*, Proc. Nat. Acad. of Sci. USA 39 (1953), 1095-1100.

[30] VEINOTT Jr, A., *Discrete dynamic programming with sensitive discount optimality criteria*, Ann. Math. Stat. 40 (1969), 1635-1660.

[31] VERKHOVSKY, B., *Smoothing system design and parametric Markovian Programming, in Markov Decision Theory*, Mathematical Centre Tract 93 (1977) (Proceedings of the advanced seminar on Markov Decision Theory, Amsterdam, 1976).

[32] ———————— & V. SPIVAK, *Water Systems optimal design and controlled stochastic processes*, Ekonomika 1, Mathematicheskie Metody, VIII (1972), 966-972.

[33] WHITE, D., *Dynamic Programming, Markov Chains, and the method of successive approximations*, J.M.A.A. 6 (1963), 373-376.

APPENDIX

In this appendix we describe the algorithm, we propose for solving the models mentioned in example 2. Assuming that the functions $q_i^k(\alpha)$ and $P_{ij}^k(\alpha)$ and $\phi(\alpha)$ are continuous in α (cf. (1.6) and (1.7)), the function to be minimized in (1.6) is guaranteed to be continuous in α in the discounted version, whereas in the undiscounted version some additional requirements on the chain structure of the tpm's of the policies in $X_i K(i)$ have to be imposed (continuity is e.g. guaranteed in the unichain case; cf. SCHWEITZER [25a]). In the absence of these requirements on the chain structure, $V_i(\alpha)$ can still be shown to be piecewise continuous, with a finite number of discontinuities, and an obvious modification of the below described algorithm can be employed:

<u>step 0</u>: Initialize MIN := $+\infty$ and $x \in E^N$. Fix $\alpha^{best} := \alpha^{new} := \alpha^{old} \in [\alpha_0, \alpha_1]$ and $\varepsilon > 0$

<u>step 1</u>: $x := \min_{k \in K(i)} [q_i^k(\alpha^{new}) + \beta \sum_j P_{ij}^k(\alpha^{new}) x_j]$, $\quad i \in \Omega$
and compute lower and upper bounds on $V(\alpha^{new})$ as a function of x:
$L(\alpha^{new}) \leq V(\alpha^{new}) \leq U(\alpha^{new})$

<u>step 2</u>: "If" $L(\alpha^{new}) + \phi(\alpha^{new}) > MIN$, "then" {$\alpha^{new}$ is suboptimal; $\alpha^{old} := \alpha^{new}$ and choose α^{new} according to a specifically chosen unconstrained search procedure; go to step 5}

<u>step 3</u>: "If" $U(\alpha^{new}) - L(\alpha^{new}) < \varepsilon$, "then" {$MIN := L(\alpha^{new}) + \phi(\alpha^{new})$; $\alpha^{best} := \alpha^{new}$, i.e. α^{new} is the best parameter choice so far; $\alpha^{old} := \alpha^{new}$; choose α^{new} according to the unconstrained search procedure; go to step 5}

<u>step 4</u>: go back to step 1, and execute the next iteration

<u>step 5</u>: "if" $\| \alpha^{old} - \alpha^{new} \| < \varepsilon$ "then" go to "END" "else" go back to step 1, and execute the next iteration with the adapted parameters

"END" Use α^{best} as an ε-optimal parameter choice, and $\frac{1}{2}(U(\alpha^{best}) + L(\alpha^{best}))$ as an ε-approximation of the value of the entire problem.