stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE BESLISKUNDE BW 35/74 JUNE

A. HORDIJK, P.J. SCHWEITZER & H.C. TIJMS
THE ASYMPTOTIC BEHAVIOUR OF THE MINIMAL TOTAL EXPECTED COST FOR THE DENUMERABLE STATE MARKOV DECISION MODEL

Prepublication

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), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

AMS (MOS) subject classification scheme (1970): 90C40

The Asymptotic Behaviour of the Minimal Total Expected Cost for the Denumerable State Markov Decision Model *)

Arie Hordijk
Mathematisch Centrum, Amsterdam

Paul J. Schweitzer
IBM Thomas J. Watson, Yorktown Heights, New York

Henk Tijms
Mathematisch Centrum, Amsterdam

ABSTRACT

This paper considers the discrete time Markov decision model with a denumerable state space and finite action space. Under certain conditions it is proved that the minimal total expected cost for a planning horizon of n epochs minus n times the minimal long-run average expected cost per unit time has a finite limit as $n \to \infty$ for each initial state.

*) This paper is not for review; it is meant for publication in a journal.

1. INTRODUCTION

This paper considers a discrete time Markov decision model with a denumerable state space and a finite action space. We shall prove that under certain conditions the minimal total expected cost for a planning horizon of n epochs minus n times the minimal long-run average expected cost per unit time has a finite limit for each initial state.

For the finite-state Markov decision model convergence results of this type were established in Bather (1973), Brown (1965), Denardo (1973), Lanery (1967), Lembersky (1973) and Schweitzer (1965 and 1974). The proofs in this paper are based on the papers of Lanery (1967) and Schweitzer (1974).

In section 2 we formulate the model. The convergence result will be proved in section 3. An application of this result to the dynamic inventory model can be found in Hordijk and Tijms (1974).

2. MODEL

We are concerned with a dynamic system which at times $t=1,2,\ldots$ is observed to be in one of a possible number of states. The set of all possible states is assumed to be denumerable and will be denoted by I. After observing the state of the system, an action must be chosen. It is assumed that the set $A(i)$ of possible actions in state i is finite for all i. If the system is in state i at time t and action a is chosen, then, regardless of the history of the system, two things happen: (i) we incur an (expected) cost $c(i,a)$ and (ii) at time $t+1$ the system will be in state j with probability $p_{ij}(a)$. The costs $c(i,a)$ and the transition probabilities $p_{ij}(a)$ are assumed to be known. We suppose that there is a finite number B such that $c(i,a) \geq B$ for all i and a, i.e., the costs $c(i,a)$ are *bounded below*.

Denote by X_t and Δ_t, $t=1,2,\ldots$ the sequences of states and actions. A policy R for controlling the system is any (possibly randomized) rule which for each t specifies which action to take at time t given the current state X_t and the history $(X_1,\Delta_1,\ldots,X_{t-1},\Delta_{t-1})$. A stationary policy f is

a rule that for each i selects an action $f(i) \in A(i)$ such that always action f(i) is taken whenever the system is in state i. Denote by F the class of all stationary policies.

When policy $f \in F$ is used the process $\{X_t\}$ is a Markov chain with stationary transition probabilities $p_{ij}(f) = p_{ij}(f(i))$. Denote by $p_{ij}^{(n)}(f)$ the n-step transition probabilities of this Markov chain, and for $n \geq 1$, let $\pi_{ij}^{(n)} = \{p_{ij}^{(1)}(f) + \ldots + p_{ij}^{(n)}(f)\}/n$. It is well known from Markov chain theory that (see Chung (1960)) the sequence $\{\pi_{ij}^{(n)}(f)\}$ has a limit $\pi_{ij}(f)$ (say) for all $i,j \in I$.

For any $i \in I$ and policy R, let

$$(1) \qquad \phi(i,R) = \liminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} E_R\{c(X_t,\Delta_t) \mid X_1 = i\},$$

where E_R denotes the expectation under policy R. Observe that $\phi(i,R)$ exists ($+\infty$ is admitted) since the costs c(i,a) are bounded below. When the limit in (1) exists $\phi(i,R)$ represents the long-run average expected cost per unit time when the initial state is i and policy R is used. A policy R^* is said to be average cost optimal if $\phi(i,R^*) \leq \phi(i,R)$ for all i and all policies R.

Let $v_0(\cdot)$ be any function such that $\sum p_{ij}(a)v_0(j)$ is finite for all i and a and is bounded below in i and a. For n=1,2,..., define

$$(2) \qquad v_n(i) = \min_{a\in A(i)} \{c(i,a) + \sum_{j\in I} p_{ij}(a)v_{n-1}(j)\}, \quad \text{for } i \in I.$$

Observe that for any $n \geq 1$ the function $v_n(\cdot)$ exists and is bounded below. The quantity $v_n(i)$ can be interpreted as the minimal total expected cost for a planning horizon of n epochs when the initial state is i and a salvage cost of $v_0(j)$ is incurred when the final state is j.

We now introduce a number of assumptions.

Assumption 1. There is a finite number g and a finite function $v(\cdot)$ such that

(i) $\sum_{j\in I} p_{ij}(a)v(j)$ is absolutely convergent for all i and a, and

(3) $$g + v(i) = \min_{a \in A(i)} \{c(i,a) + \sum_{j \in I} p_{ij}(a)v(j)\} \quad \text{for all } i \in I.$$

(ii) $E_R\{v(X_n) \mid X_1 = i\}$ is finite for all i,R and n, and $n^{-1}E_R\{v(X_n) \mid X_1 = i\}$ converges to zero as $n \to \infty$ for all i and R.

Let $F^* = \{f \in F: f(i) \text{ minimizes the right side of (3) for all } i \in I\}$. By the remark following the proof of Theorem 1 in Ross (1968) we have under assumption 1 that $g = \inf_R \phi(i,R)$ for all i and $\phi(i,f) = g$ for all i and $f \in F^*$. That is, the minimal average expected cost is independent of the initial state and equals g, and any policy $f \in F^*$ is average cost optimal. Assumption 1 (ii) will be needed only for these statements.

Assumption 2. The function $v_1(\cdot)-v(\cdot)$ is bounded.

Assumption 3. For any $f \in F$, the Markov chain $\{X_t\}$ is non-dissipative, that is, $\sum_{j \in I} \pi_{ij}(f) = 1$ for all $i \in I$.

Assumption 4. For any $f \in F^*$ holds that each state which is positive recurrent under policy f is aperiodic.

Assumption 5. For any average cost optimal stationary policy the associated Markov chain $\{X_t\}$ has no two disjoint closed sets.

We note that the assumptions 1(ii) and 2 hold when the functions $v_0(\cdot)$ and $v(\cdot)$ are bounded. However, we make these assumptions in view of applications, cf. Hordijk and Tijms (1974).

In the next section we shall prove that under the assumptions 1(i) and 2-4 the sequence $\{v_n(i) - ng - v(i)\}$ has a finite limit for all $i \in I$. Moreover, if in addition the assumptions 1(ii) and 5 hold the limit is independent of $i \in I$.

3. THE ASYMPTOTIC BEHAVIOUR OF THE MINIMAL TOTAL COST

For any $n \geq 1$, let

$$e_n(i) = v_n(i) - ng - v(i) \quad \text{for } i \in I$$

Lemma 1. Suppose that the assumptions 1(i) and 2 are satisfied. Then there is a finite number N such that $|e_n(i)| \le N$ for all $i \in I$ and $n \ge 1$.

Proof. By assumption 2, there is a finite number N such that $e_1(\cdot)$ is bounded by N. Assume now that $|e_k(i)| \le N$ for all i. Observe that together the induction hypothesis and part (i) of assumption 1 imply that $\sum p_{ij}(a)v_k(j)$ converges absolutely for all i and a. Let $f \in F^*$, and let $f_k \in F$ be such that $f_k(i)$ minimizes the right side of (2) with $n = k+1$ for all $i \in I$. It now follows from (2) and (3) that, for all $i \in I$,

$$(4) \qquad e_{k+1}(i) \le \sum_{j\in I} p_{ij}(f)e_k(j), \quad e_{k+1}(i) \ge \sum_{j\in I} p_{ij}(f_k)e_k(j).$$

From these inequalities and the induction hypothesis we get $e_{k+1}(\cdot)$ is bounded by N which completes the proof. □

The following lemma is well known (e.g. p.232 in Royden (1968)).

Lemma 2. For any $n \ge 1$, let $\{a_n(i), i \in I\}$ be a probability distribution. Suppose that $\{a(i), i \in I\}$ is a probability distribution and that $a_n(i)$ converges to $a(i)$ as $n \to \infty$ for all $i \in I$. Then, for any sequence $\{h_n(\cdot)\}$ of bounded functions which converge pointwise to the function $h(\cdot)$ on I,

$$\lim_{n\to\infty} \sum_{j\in I} h_n(j)a_n(j) = \sum_{j\in I} h(j)a(j).$$

Theorem 1. Suppose that the assumptions 1(i), 2 and 3 are satisfied. Let $f \in F^*$, and, for the Markov chain $\{X_t\}$ associated with f, let C be a class of positive recurrent states. Assume that the states of C are aperiodic. Then the sequence $\{e_n(i)\}$ has a finite limit for all $i \in C$, and, moreover, this limit is independent of $i \in C$

Proof. The reasoning of this proof parallels to that in Lanery (1967) and Schweitzer (1965). Fix some state $r \in C$. Let α and β be two limit points of the sequence $\{e_n(r)\}$. By the well-known diagonalization method and the boundedness of the sequences $\{e_n(i), n \ge 1\}$, $i \in I$, we can get two sequences $\{n_k\}$ and $\{m_h\}$ with $n_k \to \infty$ and $m_h \to \infty$ such that, for all $i \in I$, $e_{n_k}(i)$

converges to $\alpha(i)$ (say) as $k \to \infty$ with $\alpha(r) = \alpha$ and $e_{m_h}(i)$ converges to $\beta(i)$ (say) as $h \to \infty$ with $\beta(r) = \beta$. Observe that $\alpha(i)$ and $\beta(i)$ are bounded in $i \in I$. Since r was arbitrarily chosen in C and $\alpha(i)$ is a limit point of $\{e_n(i)\}$, the theorem follows when we have proved that, for some constant c,

$$\alpha(i) = \beta(i) = c \qquad \text{for all } i \in C. \tag{5}$$

To prove this, observe that $e_{n+1}(i) \le \sum p_{ij}(f) e_n(j)$ for all $i \in I$ and $n \ge 1$ (see relation (4)). Applying this inequality repeatedly and using lemma 1, we get

$$e_{n+m}(i) \le \sum_{j \in I} p_{ij}^{(m)}(f) e_n(j) \qquad \text{for all } i \in I \text{ and } n,m \ge 1. \tag{6}$$

Next we observe that from Markov chain theory (see Chung (1960)) it follows that, for all $i,j \in C$, the sequence $\{p_{ij}^{(n)}(f)\}$ has a limit $\pi_j(f)$ (say) which is independent of i. Moreover,

$$\pi_j(f) > 0 \quad \text{for all } j \in C \quad \text{and} \quad \sum_{j \in C} \pi_j(f) = 1. \tag{7}$$

Also, for all $i \in C$ and $n \ge 1$, $\sum p_{ij}^{(n)}(f) = 1$ where the sum is over $j \in C$. We shall now prove that, for all $i \in C$,

$$\beta(i) \le \sum_{j \in C} \alpha(j)\pi_j(f) \quad \text{and} \quad \alpha(i) \le \sum_{j \in C} \beta(j)\pi_j(f). \tag{8}$$

For reasons of symmetry it suffices to prove the first part of (8). To do this, choose for each integer $k \ge 1$ a positive integer h(k) such that $t_k > k$, where $t_k = m_{h(k)} - n_k$. Taking $n = n_k$ and $m = t_k$ in (6), letting $k \to \infty$ and using lemma 2, we get the first part of (8). Substituting the first inequality of (8) into the second one and the second one into the first one, and using the second relation in (7), we have, for all $i \in C$,

$$\beta(i) \le \sum_{j \in C} \beta(j)\pi_j(f) \quad \text{and} \quad \alpha(i) \le \sum_{j \in C} \alpha(j)\pi_j(f). \tag{9}$$

Multiplying both sides of each inequality in (9) by $\pi_i(f)$, summing over $i \in C$ and using (7), we find that the equality signs in (9) hold for all $i \in C$. Together this and (8) imply (5) which completes the proof. □

Lemma 3. Suppose that the assumptions 1, 3 and 5 are satisfied. Assume that $d(\cdot)$ is a bounded function on I such that, for all $i \in I$,

$$(10) \qquad g + v(i) + d(i) = \min_{a \in A(i)} \{c(i,a) + \sum_{j \in I} p_{ij}(a)[v(j) + d(j)]\}.$$

Then, for some constant d, $d(i) = d$ for all $i \in I$.

Proof. The reasoning of this proof is similar to that used to prove Theorem 2.4 in Schweitzer (1969). Choose $f \in F^*$, and let $h \in F$ be such that h(i) minimizes the right side of (10) for all $i \in I$. Since $d(\cdot)$ is bounded it follows from assumption 1 that $n^{-1}E_R\{v(X_n) + d(X_n) \mid X_1 = i\} \to 0$ as $n \to \infty$ for all i and R. Now, by the remark following Theorem 1 in Ross (1968), we have $\phi(i,h) = g$ for all i. Hence policy h is average cost optimal. Since f and h are average cost optimal, we have by the assumptions 3 and 5 that, for any $j \in I$, $\pi_{ij}(f)$ and $\pi_{ij}(h)$ are independent of $i \in I$ and are equal to $\pi_j(f)$ and $\pi_j(h)$ (say), cf. Chung (1960).

By (3) and (10), $d(i) \le \sum p_{ij}(f)d(j)$ for all $i \in I$. Iterating this n times and averaging over n, yields $d(i) \le \sum \pi_{ij}^{(n)}(f)d(j)$ for all $i \in I$ and $n \ge 1$. By assumption 3, $\sum \pi_j(f) = 1$. It now follows from lemma 2 that

$$(11) \qquad d(i) \le \sum_{j \in I} d(j)\pi_j(f) \qquad \text{for all } i \in I.$$

Similarly, using the fact that $d(i) \ge \sum p_{ij}(h)d(j)$ for all i, we get

$$(12) \qquad d(i) \ge \sum_{j \in I} d(j)\pi_j(h) \qquad \text{for all } i \in I.$$

Denote by R(f) and R(h) the set of states that are positive recurrent under policy f and h, respectively. Multiplying both sides of (11) by $\pi_i(f)$, summing over i, and using that $\pi_i(f) > 0$ for $i \in R(f)$, it follows that the equality sign holds in (11) for all $i \in R(f)$. Similarly, the equality sign holds in (12) for all $i \in R(h)$. By assumption 5 we have $R(f) \cap R(h)$

is not empty. Together these facts, (11) and (12) imply the lemma. □

We are now in a position to prove the main result.

Theorem 2. Suppose that the assumptions 1(i) and 2-4 are satisfied. Then the sequence $\{e_n(i)\}$ has a finite limit for all $i \in I$. This limit is independent of $i \in I$ if in addition the assumptions 1(ii) and 5 are satisfied.

Proof. Since the sums in (2) and (3) converge absolutely (cf. lemma 1), it follows from (2) that, for all $i \in I$ and $n \geq 1$,

$$(13) \qquad e_{n+1}(i) = \min_{a \in A(i)} \{b(i,a) + \sum_{j \in I} p_{ij}(a) e_n(j)\},$$

where

$$(14) \qquad b(i,a) = c(i,a) - g + \sum_{j \in I} p_{ij}(a) v(j) - v(i).$$

By assumption 1(i),

$$(15) \qquad \min_{a \in A(i)} b(i,a) = 0 \qquad \text{for all } i \in I.$$

Let $M(i) = \limsup_{n \to \infty} e_n(i)$, and let $m(i) = \liminf_{n \to \infty} e_n(i)$ for $i \in I$. By lemma 1, the functions $M(\cdot)$ and $m(\cdot)$ are bounded. To prove that $m(i) = M(i)$ for all i, we shall first show that

$$(16) \qquad m(i) \geq \min_{a \in A(i)} \{b(i,a) + \sum_{j \in I} p_{ij}(a) m(j)\} \qquad \text{for all } i \in I,$$

$$(17) \qquad M(i) \leq \min_{a \in A(i)} \{b(i,a) + \sum_{j \in I} p_{ij}(a) M(j)\} \qquad \text{for all } i \in I.$$

We only prove (16). The proof of (17) is very similar. To prove (16), fix some state $i_0 \in I$. By the diagonalization method and lemma 1, we can get a sequence $\{n_k\}$ with $n_k \to \infty$ such that the sequence $\{e_{n_k}(i_0)\}$ has the limit $m(i_0)$ and, for all $i \in I$, the sequence $\{e_{n_k - 1}(i)\}$ has a finite limit $\gamma(i)$ (say). Choose $\varepsilon > 0$. Since $A(i_0)$ is finite there is an integer k_0 such that, for all $a \in A(i_0)$ and $k \geq k_0$,

$$(18)\qquad e_{n_k}(i_0) \le m(i_0) + \varepsilon, \quad \sum_{j\in I} p_{i_0 j}(a) e_{n_k - 1}(j) \ge \sum_{j\in I} p_{i_0 j}(a)\gamma(j) - \varepsilon.$$

From these inequalities, (13) and the fact that $\gamma(j) \ge m(j)$ for all j we easily get that $m(i_0) + 2\varepsilon$ is larger than or equal to the right side of (16) with $i = i_0$. This proves (16) since ε and i_0 were chosen arbitrarily.

Let $f \in F$ be such that $f(i)$ minimizes the right side of (16) for all $i \in I$. By (16) and (17), for all $i \in I$,

$$(19)\quad b(i,f(i)) + \sum_{j\in I} p_{ij}(f)m(j) \le m(i) \le \le M(i) \le b(i,f(i)) + \sum_{j\in I} p_{ij}(f)M(j).$$

Multiply both sides of the first inequality in (19) by $\pi_{ii}(f)$ and sum over $i \in I$. We have $\sum \pi_{ii}(f)p_{ij}(f) = \pi_{jj}(f)$ for all j where the sum is over $i \in I$, see Chung (1960). Using this, we get after an interchange of the order of summation,

$$(20)\qquad \sum_{i\in I} \pi_{ii}(f)b(i,f(i)) \le 0.$$

The summation operations used to derive (20) are justified by the boundedness of $m(\cdot)$ and the nonnegativity of $b(\cdot,\cdot)$ (see (15)). Let $R(f)$ be the set of states which are positive recurrent under policy f. Then, by assumption 3, $R(f)$ is not empty. Since $b(\cdot,\cdot)$ is nonnegative and $\pi_{ii}(f) > 0$ for $i \in R(f)$, the inequality (20) implies $b(i,f(i)) = 0$ for all $i \in R(f)$. Hence, by (14),

$$g + v(i) = c(i,f(i)) + \sum_{j\in I} p_{ij}(f)v(j) \qquad \text{for all } i \in R(f),$$

so, $f(i)$ minimizes the right side of (3) for all $i \in R(f)$. Choose $f^* \in F^*$ such that $f^*(i) = f(i)$ for all $i \in R(f)$. Then $R(f)$ is contained in the set of states which are positive recurrent under policy f^*. Now, by theorem 1, $m(i) = M(i)$ for all $i \in R(f)$. To prove $m(i) = M(i)$ for all i, we observe that, by (19),

$$0 \le M(i) - m(i) \le \sum_{j\in I} p_{ij}(f)\{M(j) - m(j)\} \qquad \text{for all } i \in I.$$

Iterate the latter inequality n times and average over n. Letting $n \to \infty$, and using assumption 3 and lemma 2, we get

$$(21) \qquad 0 \le M(i) - m(i) \le \sum_{j \in I} \pi_{ij}(f)\{M(j) - m(j)\} \qquad \text{for all } i \in I.$$

Now, for any $i \in I$, $\pi_{ij}(f) = 0$ when $j \notin R(f)$, cf. Chung (1960). Since $m(j) = M(j)$ for $j \in R(f)$ it now follows from (21) that $m(i) = M(i)$ for all i. This proves the first part of the theorem. To prove the second part, observe that, by (16) and (17),

$$m(i) = \min_{a \in A(i)} \{b(i,a) + \sum_{j \in I} p_{ij}(a)m(j)\} \qquad \text{for all } i \in I.$$

Substituting into this equality the expression for b(i,a) (see (14)), we find that lemma 3 applies with $d(\cdot) = m(\cdot)$. This ends the proof. □

Remark. Suppose that the assumptions 1-5 are satisfied. For any $n \ge 1$, let $f_n \in F$ be such that $f_n(i)$ minimizes the right side of (2) for all i. Assume that, for some $f \in F$, $f_n = f$ for infinitely many values of n. Using theorem 2 and lemma 1, we easily derive from (2) that $f \in F^*$. Hence f is average cost optimal.

REFERENCES

BATHER, J. (1973), Optimal decision procedures for finite Markov chains, Part II: Communicating systems. *Adv. App. Prob.* 5, 521-540.

BROWN, B.W. (1965), On the iterative method of dynamic programming on a finite space discrete time Markov process. *Ann. Math. Statist.* 36, 1279-1285.

CHUNG, K.L. (1960), *Markov chains with stationary transition probabilities.* Springer Verlag, Berlin.

DENARDO, E.V. (1973), A Markov decision problem, in: T.C. Hu & S.M. Robinson, *Mathematical programming,* Academic Press, New York.

HORDIJK, A. & H.C. TIJMS (1974) Convergence results and approximations for optimal (s,S) policies. To appear in *Management Science.*

LANERY, E. (1967), Etude asymptotique des systèmes Markoviens à commande. *Rev. Informat. Recherche Opérationelle* No. 3, 3-56.

LEMBERSKY, M.R. (1973), On maximal rewards and ε-optimal policies in in continuous time Markov decision chains. *Ann. Statist.* 2 , 159-169.

ROSS, S.M. (1968), Arbitrary state Markovian decision processes. *Ann. Math. Statist.* 39, 2118-2122.

ROYDEN, H. (1968), *Real Analysis* (2nd.ed.), MacMillan Company, New York.

SCHWEITZER, P.J. (1965), Perturbation theory and Markovian decision processes. Technical Report No. 15, Operations Research Center, M.I.T., Cambridge, Massachusetts.

SCHWEITZER P.J. (1969), Perturbation theory and undiscounted Markov renewal programming. *Operations Res.* 17, 716-727.

SCHWEITZER, P.J. (1974), Asymptotic convergence of undiscounted value iteration. *To be published.*