Stochastic Processes and their Applications 9 (1979) 223–235

DENUMERABLE STATE SEMI-MARKOV DECISION PROCESSES WITH UNBOUNDED COSTS, AVERAGE COST CRITERION

A. FEDERGRUEN
Mathematisch Centrum, Amsterdam, The Netherlands

A. HORDIJK
Rijksuniversiteit Leiden, Leiden, The Netherlands

H.C. TIJMS
Vrije Universiteit, Amsterdam, The Netherlands

Received 27 September 1978
Revised 11 June 1979

This paper establishes a rather complete optimality theory for the average cost semi-Markov decision model with a denumerable state space, compact metric action sets and unbounded one-step costs for the case where the underlying Markov chains have a single ergodic set. Under a condition which, roughly speaking, requires the existence of a finite set such that the supremum over all stationary policies of the expected time and the total expected absolute cost incurred until the first return to this set are finite for any starting state, we shall verify the existence of a finite solution to the average costs optimality equation and the existence of an average cost optimal stationary policy.

Semi-Markov decision processes	unbounded one-step costs
denumerable state space	average costs
optimal stationary policy	optimality equation

1. Introduction

We are concerned with a dynamic system which at decision epochs beginning with epoch 0 is observed to be in one of the states of a *denumerable* state space I and subsequently is controlled by choosing an action. For any state $i \in I$, the set $A(i)$ denotes the set of pure actions available in state i. If at any decision epoch the system is in state i and action $a \in A(i)$ is taken, then, regardless of the history of the system, the following happens:

(i) an immediate cost $c(i, a)$ is incurred,

(ii) the time until the next decision epoch and the state at the next decision epoch are random with joint probability distribution function $Q(\cdot, \cdot | i, a)$.

For any $i \in I$ and $a \in I$, let

$$p_{ij}(a) = Q(\infty, j | i, a) \quad \text{for } j \in I$$

and

$$\tau(i, a) = \sum_{j \in I} \int_0^\infty tQ(\mathrm{d}t, j|i, a).$$

i.e. $p_{ij}(a)$ denotes the probability that the next state will be j and $\tau(i, a)$ denotes the unconditional mean time until the next decision epoch when action a is taken in state i. Observe that $\sum_{j \in I} p_{ij}(a) = 1$ for all i, a. We make the following assumption.

Assumption 1. (a) For any $i \in I$, the set $A(i)$ is a compact metric set,

(b) for any $i \in I$, both $c(i, a)$, $p_{ij}(a)$ for any $j \in I$ and $\tau(i, a)$ are continuous on $A(i)$,

(c) there is a number $\varepsilon > 0$ such that $\tau(i, a) \geq \varepsilon$ for all $i \in I$ and $a \in A(i)$.

We now introduce some familiar notions. For $n = 0, 1, \ldots$, denote by X_n and a_n the state and the action at the nth decision epoch (the 0th decision epoch is at epoch 0). A policy π for controlling the system is any measurable rule which for each n specifies which action to choose at the nth decision epoch given the current state X_n and the sequence $(X_0, a_0, \ldots, X_{n-1}, a_{n-1})$ of past states and actions where the actions chosen may be randomised. A policy π is called *memoryless* when the actions chosen are independent of the history of the system except for the present state. Define $\mathcal{R}$ as the class of all stochastic matrices $P = (p_{ij})$, $i, j \in I$ such that for any $i \in I$ the elements of the ith row of P can be represented by

$$p_{ij} = \int_{A(i)} p_{ij}(a)\pi_i(\mathrm{d}a) \quad \text{for all } j \in I \tag{1}$$

for some probability distribution $\pi_i\{\cdot\}$ on $A(i)$. Then any memoryless policy π can be represented by some sequence $(P_1, P_2, \ldots)$ in $\mathcal{R}$ such that the ith row of P_n gives the probability distribution of the state at the nth decision epoch when the current state at the $(n-1)$st decision epoch is i and policy π is used. Define $F = X_{i \in I} A(i)$. Observe that, under Assumption 1(a), F is a compact metric set in the product topology. For any $f \in F$, let $P(f)$ be the stochastic matrix whose (i, j)th element is $p_{ij}(f(i))$, $i, j \in I$ and for $n = 1, 2, \ldots$ denote by the stochastic matrix $P^n(f) = (p_{ij}^n(f))$ the n-fold matrix product of $P(f)$ with itself. A memoryless policy $\pi = (P_1, P_2, \ldots)$ is called *stationary* when $P_n = P(f)$ for all $n \geq 1$ for some $f \in F$. This policy which always prescribes to take the single action $f(i) \in A(i)$ whenever the system is in state i will be denoted by $f^{(\infty)}$. Observe that under the stationary policy $f^{(\infty)}$ the process $\{X_n, n \geq 0\}$ is a Markov chain with one-step transition probability matrix $P(f)$.

For $n = 0, 1, \ldots$, denote by τ_n the time between the nth and $(n+1)$st decision epoch. A policy π^* is said to be (*strongly*) *average cost optimal* when $\limsup_{n \to \infty} \phi_n(i, \pi^*)$ is less than or equal to $\limsup_{n \to \infty} \phi_n(i, \pi)$ ($\liminf_{n \to \infty} \phi_n(i, \pi)$) for any $i \in I$ and policy π where $\phi_n(i, \pi)$ is defined by

$$\phi_n(i, \pi) = \frac{E_\pi\left\{\sum_{k=0}^{n} c(X_k, a_k) \,\middle|\, X_0 = i\right\}}{E_\pi\left\{\sum_{k=0}^{n} \tau_k \,\middle|\, X_0 = i\right\}}, \quad n = 0, 1, \ldots \tag{2}$$

with E_π is the expectation under policy π. We here assume that this quantity is well-defined for any $i \in I$ and policy π as is the case under the additional Assumption 2(a) to be stated below.

It is well-known that an average cost optimal policy may not exist and even an example has been given in [7] in which an average cost policy exists but any average cost optimal policy is nonstationary. It is remarkable in this example, that besides uniformly bounded $c(i, a)$ and $\tau(i, a)$, any stochastic matrix $P \in \mathscr{R}$ is irreducible and positive recurrent. In general we can only state that for fixed initial state we may restrict ourselves to the class of memoryless policies. More precisely, by a slight generalization of the proof of Theorem 2 in [2], we have the known result that for any fixed $i_0 \in I$ and policy π_0 a memorlyless policy π_M can be found such that for any $k \in I$, Borel set $B \subseteq A(k)$ and $n \geqslant 0$,

$$\mathbf{P}_{\pi_M}\{X_n = k, a_n \in B \mid X_0 = i_0\} = \mathbf{P}_{\pi_0}\{X_n = k, a_n \in B \mid X_0 = i_0\}. \tag{3}$$

Another general result says that if the average cost optimality equation has a solution satisfying some regularity condition then any stationary policy generated by the optimality equation is strongly average cost optimal, cf. [14].

In this paper we shall establish for the average cost optimality criterion a rather complete theory for the denumerable state semi-Markov decision model with unbounded one-step costs for the case where the underlying Markov chains have only a single ergodic set. This theory both extends considerably and unifies the finite state space model and the special cases of the denumerable state space model so far studied in the literature, cf. [1, 3, 4, 9, 10, 13, 14, 17–19]. This paper exhibits the existence of a finite solution to the average cost optimality equation and the existence of a strongly average cost optimal stationary policy under a condition which, roughly speaking, requires the unichainedness of the stochastic matrices $P(f)$, $f \in F$ and the existence of a *finite* set K such that the supremum over the stationary policies of both the expected time and the total expected absolute costs incurred until the first return to this set K are finite for any starting state. This assumption considerably weakens the usual conditions requiring that the set K is a singleton or assuming that both the one-step costs and the mean recurrence times are uniformly bounded. The latter asumptions are seldom met in typical applications as in inventory and queueing theory.

In Section 2 we will give the essence of our analysis by first establishing relationships between the original decision processes and the embedded decision processes on the finite set K. Next in Section 3 we will prove both the average cost optimality equation and the existence of a strongly average cost optimal stationary policy by using proof techniques developed in [4, 9, 14, 18]. We conclude this section be remarking that extensions of the theory presented for the unichained case to the important case of 'communicating Markov decision chains' (cf. [1, 9]) will require different proof techniques as possibly linear programming or fixed point methods.

2. Analysis of embedded decision processes

We first need some notation. For any $A \subset I$, define

$$N(A) = \inf\{n \geq 1 \mid X_n \in A\}$$

where $N(A) = \infty$ if $X_n \notin A$ for all $n \geq 1$, i.e. $N(A)$ denotes the number of transitions until the first return to the set A. Also for any $A \subset I$ and $f \in F$, define for $i, j \in I$ and $n \geq 1$ the taboo probability

$$_A p_{ij}^n(f) = \mathbf{P}_{f^{(\infty)}}\{X_n = j, X_k \notin A \text{ for } 1 \leq k \leq n-1 \mid X_0 = i\}. \tag{4}$$

Observe that

$$E_{f^{(\infty)}}\{N(A) \mid X_0 = i\} = 1 + \sum_{n=1}^{\infty} \sum_{j \notin A} {}_A p_{ij}^n(f)). \tag{5}$$

We now introduce our main assumption.

Assumption 2. (a) There is a finite set K such that for any $i \in I$ the quantities $u^*(i)$ and $y^*(i)$ are finite where

$$\sup_{f \in F} E_{f^{(\infty)}}\left\{\sum_{k=0}^{N(K)-1} \tau_k \mid X_0 = i\right\} = u^*(i) \quad \text{for all } i \in I \tag{6}$$

and

$$\sup_{f \in F} E_{f^{(\infty)}}\left\{\sum_{k=0}^{N(K)-1} |c(X_k, a_k)| \,\middle|\, X_0 = i\right\} = y^*(i) \quad \text{for all } i \in I, \tag{7}$$

(b) for any $f \in F$, the stochastic matrix $P(f)$ has no two disjoint closed sets.

In words, Assumption 2(a) requires the existence of a finite set K such that the supremum over all stationary policies of both the expected time and the total expected absolute cost incurred until the first return to the set K are finite for any starting state. To satisfy Assumption 2(a) in applications, it may be necessary to exclude in certain states those actions which are far from being 'optimal', e.g. in an $M/M/c$ queueing system with a controllable number of operating servers consider only policies under which all c servers will be operating when the queue size exceeds some given large value. By other arguments based on the specific form of the application, it is usually not difficult to show that any other policy can be improved in average costs by a policy belonging the class of policies considered. We remark that Assumption 2(a) is satisfied with bounded functions u^* and y^* when the quantities $c(i, a)$ and $\tau(i, a)$ are uniformly bounded and any of the recurrence conditions on the set $(P(f), f \in F)$ given in [6] holds, cf. also [5].

We shall now first verify as key result that under the Assumptions 1 and 2 for any $f \in F$ a state $s_f \in K$ exists such that under policy $f^{(\infty)}$ the expected time and the total expected absolute cost incurred until the first return to the state s_f are bounded by $u^*(i) + c$ and $y^*(i) + c$ respectively for any starting state i for some constant c independent of $f \in F$. We shall need the following lemma whose proof is standard.

Lemma 1. *Let A be any subset of I. Then, for any $i \in I$ and $f \in F$*

$$E_{f^{(\infty)}}\left\{\sum_{k=0}^{N(A)-1} \tau(X_k, a_k) \mid X_0 = i\right\} = E_{f^{(\infty)}}\left\{\sum_{k=0}^{N(A)-1} \tau_k \mid X_0 = i\right\}$$

$$= \tau(i, f(i)) + \sum_{n=1}^{\infty} \sum_{j \notin A} \tau(j, f(j)) {}_A p_{ij}^n(f). \tag{8}$$

By Lemma 1, we may replace τ_k by $\tau(X_k, a_k)$ in (6). This result will be essentially used in the analysis hereafter. It now follows from Lemma 1 and (5)–(6) that under the Assumptions 1(c) and 2(a),

$$E_{f^{(\infty)}}\{N(K) \mid X_0 = i\} \leqq \frac{u^*(i)}{\varepsilon} \quad \text{for any } f \in F \text{ and } i \in I. \tag{9}$$

It is our conjecture that (9) implies tightness of the collection of the stationary probability distributions of the stochastic matrices $P(f)$, $f \in F$.

Under Assumption 2, define for any $f \in F$

$$q_{ij}(f) = \sum_{n=1}^{\infty} {}_K p_{ij}^n(f), \quad i \in I, \ j \in K, \tag{10}$$

i.e. $q_{ij}(f)$ is the probability that at the first return to the set K the transition occurs into state j starting from state i and using policy $f^{(\infty)}$. Observe that, by (9),

$$\sum_{j \in K} q_{ij}(f) = 1 \quad \text{for all } i \in I. \tag{11}$$

For any $f \in F$, define for $i \in I$ and $j \in K$ the (possibly infinite) quantity

$$\nu_{ij}(f) = \text{expected number of returns to the set } K \text{ until the first transition into state } j \text{ occurs starting from state } i \text{ and using policy } f^{(\infty)}. \tag{12}$$

Theorem 1. *Suppose that the Assumptions 1 and 2 hold. Then*

(a) *for any $f \in F$, the finite stochastic matrix $(q_{ij}(f))$, $i, j \in K$ has no two disjoint closed sets,*

(b) *for any $i \in I$ and $j \in K$, the probability $q_{ij}(f)$ is continuous on F,*

(c) *there is a finite number B such that for any $f \in F$ a state $s_f \in K$ exists for which $\nu_{is_f}(f) \leqq B$ for all $i \in I$.*

Proof. (a) Fix $f \in F$. Let $K_1 \subseteq K$ and $K_2 \subseteq K$ be any two non-empty sets that are closed under the stochastic matrix $Q(f) = (q_{ij}(f))$, $i, j \in K$. To prove that $K_1 \cap K_2$ is not empty, define for $r = 1, 2$ the set

$$I_r = \{j \in I \mid p_{ij}^n(f) > 0 \text{ for some } i \in K_r \text{ and } n \geqq 1\}.$$

It is immediate that both sets I_1 and I_2 are closed under $P(f)$ and hence $I_1 \cap I_2 \neq \emptyset$. Choose now $t \in I_1 \cap I_2$. Since $t \in I_1$, it follows that

$$p_{st}^m(f) > 0 \quad \text{for some } s \in K_1 \text{ and } m \geq 1. \tag{13}$$

Using (9), $t \in I_2$ and the fact that K_2 is closed under $Q(f)$ it is readily verified by contradiction that

$$p_{tu}^n(f) > 0 \quad \text{for some } u \in K_2 \text{ and } n \geq 1. \tag{14}$$

By (13) and (14), $p_{su}^{m+n}(f) > 0$. This implies that $u \in K_2$ can be reached from $s \in K_1$ under $Q(f)$. Since K_1 is closed under $Q(f)$, it follows that $u \in K_1$ so that $K_1 \cap K_2 \neq \emptyset$.

(b) By Assumption 1, we have that F is a compact metric set on which $p_{ij}(f)$ is continuous for any $i, j \in I$. Using this fact and the relation

$$_K p_{ij}^n(f) = \sum_{h \in K} p_{ih}(f) \, _K p_{hj}^{n-1}(f) \quad \text{for } n = 2, 3, \ldots$$

it follows by induction that $_K p_{ij}^n(f)$ is continuous on F for any $n \geq 1$ and $i, j \in I$. Hence $q_{ij}(f)$ is continuous on F if the sum (10) converges uniformly on F. To prove this, fix $s \in I$ and observe that, by (9),

$$\sum_{n=0}^{\infty} \mathbf{P}_{f^\infty}\{N(K) > n \,|\, X_0 = s\} \leq \frac{u^*(s)}{\varepsilon} \quad \text{for all } f \in F. \tag{15}$$

Choose now $0 < \delta < 1$. Then there is an integer M such that

$$\mathbf{P}_{f^\infty}\{N(K) > M \,|\, X_0 = s\} \leq \delta \quad \text{for all } f \in F. \tag{16}$$

To prove this, assume the contrary. Using the fact that $\mathbf{P}_{f^\infty}\{N(K) > n \,|\, X_0 = s\}$ is non-increasing in n, we then get a contradiction with (15). Now, by (16) we have for any $j \in K$

$$\sum_{n=M+1}^{\infty} {}_K p_{sj}^n(f) \leq \mathbf{P}_{f^\infty}\{N(K) > M \,|\, X_0 = s\} \leq \delta \quad \text{for all } f \in F$$

which proves the desired result since $\delta > 0$ was chosen arbitrarily.

(c) By the finiteness of K and the assertions (a) and (b) of the theorem, this assertion is an immediate consequence of Theorem 2.6 in [5] or Theorem 4 in [6].

The following theorem will play a crucial role in the analysis in the next section.

Theorem 2. *Suppose that the Assumptions* 1 *and* 2 *hold. Then there is a finite number* c *such that for any* $f \in F$ *a state* $s_f \in K$ *exists for which*

$$E_{f^\infty}\left\{ \sum_{k=0}^{N(\{s_f\})-1} \tau(X_k, a_k) \,|\, X_0 = i \right\} \leq u^*(i) + c \quad \textit{for all } i \in I \tag{17}$$

and

$$E_{f^\infty}\left\{ \sum_{k=0}^{N(\{s_f\})-1} |c(X_k, a_k)| \,|\, X_0 = i \right\} \leq y^*(i) + c \quad \textit{for all } i \in I. \tag{18}$$

Proof. By Theorem 2 we can choose a finite number B and for any $f \in F$ a state $s_f \in K$ such that

$$\nu_{is_f}(f) \leq B \quad \text{for all } i \in I \text{ and } f \in F. \tag{19}$$

We shall now verify (17). The proof of (18) is identical. Fix now $f \in F$. We introduce the following notation. For any $i \in I$ and $j \in K$, define $\hat{q}_{ij}^1(f) = q_{ij}(f)$ and, for $n = 2, 3, \ldots$, let

$$\hat{q}_{ij}^n(f) = \sum_{\substack{k \in K \\ k \neq s_f}} q_{ik}(f) \hat{q}_{kj}^{n-1}(f) \quad \text{for } i \in I \text{ and } j \in K.$$

Observe that $\hat{q}_{ij}^n(f)$ is the probability that during the first $n-1$ returns to the set K no transition occurs into state s_f and that at the nth return to the set K a transition occurs into state j starting from state i and using policy $f^{(\infty)}$. We have

$$\nu_{is_f} = 1 + \sum_{n=1}^{\infty} \sum_{\substack{j \in K \\ j \neq s_f}} \hat{q}_{ij}^n(f) \quad \text{for all } i \in I. \tag{20}$$

Define $\nu_0 = 0$ and, for $n \geq 1$, $\nu_n = \inf\{m > \nu_{n-1} \mid X_m \in K\}$. Also, define $\delta_0 = 1$ and, for any $k \geq 1$, $\delta_k = 1$ if $X_m \neq s_f$ for $1 \leq m \leq k$ and $\delta_k = 0$ otherwise. Denote by $T(i, f)$ the first expression in (8) with $A = K$. Then using the first equality in (8) and (6), we find

$$\begin{aligned}
E_{f^{(\infty)}}\left\{ \sum_{k=0}^{N(\{s_f\})-1} \tau(X_k, a_k) \mid X_0 = i \right\} &= E_{f^{(\infty)}}\left\{ \sum_{k=0}^{\infty} \delta_k \tau(X_k, a_k) \mid X_0 = i \right\} \\
&= E_{f^{(\infty)}}\left\{ \sum_{n=1}^{\infty} \sum_{k=\nu_{n-1}}^{\nu_n - 1} \delta_k \tau(X_k, a_k) \mid X_0 = i \right\} \\
&= T(i, f) + \sum_{n=2}^{\infty} E_{f^{(\infty)}}\left\{ \sum_{k=\nu_{n-1}}^{\nu_n - 1} \delta_k \tau(X_k, a_k) \mid X_0 = i \right\} \\
&= T(i, f) + \sum_{n=2}^{\infty} \sum_{j \neq s_f} \hat{q}_{ij}^{n-1}(f) T(j, f) \\
&\leq u^*(i) + \max_{j \in K} u^*(j) \sum_{n=2}^{\infty} \sum_{j \neq s_f} \hat{q}_{ij}^{n-1}(f) \quad \text{for all } i \in I.
\end{aligned}$$

Invoking (19) and (20), we now get the desired result.

We need the following results from positive dynamic programming (cf. [9, 16]).

Lemma 2. *Consider the positive dynamic program* $(S, D(s), q(t \mid s, a), r(s, a))$ *where the state space* S *is denumerable, the action set* $D(s)$ *is a compact metric set for any* $s \in S$ *and the immediate return* $r(s, a)$ *is non-negative for all* $s \in S$ *and* $a \in D(s)$. *Also assume that for any* $s \in S$ *both* $r(s, a)$ *and the one-step transition probability* $q(t \mid s, a)$

for any $t \in S$ *are continuous on* $D(s)$. *For any policy* π, *define* $V(s, \pi) = E_\pi\{\sum_{n=0}^{\infty} r(X_n, a_n) \mid X_0 = s\}$, $s \in S$ *where* X_n *and* a_n *denote the state and the action at the nth decision epoch. Let* $V(s) = \sup_\pi V(s, \pi)$, $s \in S$. *Then*

$$\sup_{f \in F} V(s, f^{(\infty)}) = V(s) \quad \textit{for all } s \in S, \tag{21}$$

$$V(s) = \sup_{a \in D(s)} \left\{ r(s, a) + \sum_{t \in S} V(t) q(t \mid s, a) \right\} \quad \textit{for all } s \in S. \tag{22}$$

Proof. We need some notation. For any integer $M \geq 1$, let $r^M(s, a) = \min(r(s, a), M)$ for all s, a. For any $0 < \alpha < 1$, $s \in S$ and policy π, define

$$V_\alpha(s, \pi) = E_\pi\left\{ \sum_{n=0}^{\infty} \alpha^n r(X_n, a_n) \mid X_0 = s \right\} \quad \text{and}$$

$$V_\alpha^M(s, \pi) = E_\pi\left\{ \sum_{n=0}^{\infty} \alpha^n r^M(X_n, a_n) \mid X_0 = s \right\}.$$

Using the non-negativity of $r(s, a)$ we have by the monotone convergence theorem

$$\lim_{M \to \infty} V_\alpha^M(s, \pi) = V_\alpha(s, \pi) \quad \text{for any } 0 < \alpha < 1,\ s \in S \text{ and policy } \pi, \tag{23}$$

and, by a Tauberian theorem,

$$\lim_{\alpha \to 1} V_\alpha(s, \pi) = V(s, \pi) \quad \text{for any } s \in S \text{ and policy } \pi. \tag{24}$$

Letting $V_\alpha^M(s) = \sup_\pi V_\alpha^M(s, \pi)$, $s \in I$, it is well-known from discounted dynamic programming (e.g. cf. [9, 12]) that for any $0 < \alpha < 1$ and $M \geq 1$

$$V_\alpha^M(s) = \max_{a \in D(s)} \left\{ r(s, a) + \alpha \sum_{t \in S} V_\alpha^M(t) q(t \mid s, a) \right\} \quad \text{for all } s \in S \tag{25}$$

and

$$\sup_{f \in F} V_\alpha^M(s, f^{(\infty)}) = \sup_\pi V_\alpha^M(s, \pi) \quad \text{for all } s \in S. \tag{26}$$

Using the fact that $\lim_{n \to \infty} \sup_x g_n(x) = \sup_x \lim_{n \to \infty} g_n(x)$ for any non-decreasing sequence of functions $\{g_n\}$, we obtain from (23) and (26) that $\sup_{f \in F} V_\alpha(s, f^{(\infty)}) = \sup_\pi V_\alpha(s, \pi)$ for all $s \in S$ and $0 < \alpha < 1$. Next, by letting $\alpha \to 1$ in this relation and using (24) we get (21). The optimality equation (22) follows by the same reasoning from (23)–(25) by first letting $M \to \infty$ and next letting $\alpha \to 1$.

By defining an appropriate Markov decision model with an absorbing state ∞, the next theorem is an easy consequence of Assumption 2(a) and Lemma 2.

Theorem 3. *Suppose that the Assumptions* 1 *and* 2(a) *hold. Then*

$$u^*(i) = \sup_{a \in A(i)} \left\{ \tau(i, a) + \sum_{j \in K} p_{ij}(a) u^*(j) \right\} \quad \text{for all } i \in I, \tag{27}$$

$$y^*(i) = \sup_{a \in A(i)} \left\{ |c(i, a)| + \sum_{j \in K} p_{ij}(a) y^*(j) \right\} \quad \text{for all } i \in I. \tag{28}$$

By this theorem, we have that Assumption 2(a) is equivalent to the condition requiring the existence of a finite set K and a finite non-negative function $y(i)$, $i \in I$ such that

$$|c(i, a)| + \tau(i, a) + \sum_{j \in K} p_{ij}(a) y(j) \leq y(i) \quad \text{for all } i \in I \text{ and } a \in A(i). \tag{29}$$

The condition (29) with K equal to a singleton was first studied in [9] where this condition was called a Liapunov condition, cf. also [8, 10] for further investigations on Liapunov conditions.

For any $P \in \mathcal{R}$, define the substochastic matrix $\hat{P} = (\hat{p}_{ij})$ by

$$\hat{p}_{ij} = \begin{cases} p_{ij} & \text{for } i \in I, j \notin K, \\ 0 & \text{for } i \in I, j \in K. \end{cases} \tag{30}$$

Then, by Theorem 3,

$$\hat{P} u^* \leq u^* \quad \text{and} \quad \hat{P} y^* \leq y^* \qquad \text{for any } P \in \mathcal{R}. \tag{31}$$

Lemma 3. *Suppose that the Assumptions* 1 *and* 2(a) *hold. Then, for any sequence* $(P_1, P_2, \ldots)$ *of stochastic matrices in* $\mathcal{R}$,

$$P_1 \cdots P_n y^*(i) \leq y^*(i) + \max_{j \in K} y^*(j) + \sum_{k=1}^{n-1} \sum_{h \in K} \hat{P}_{k+1} \cdots \hat{P}_n y^*(h)$$

$$\leq y^*(i) + n \max_{j \in K} y^*(j) \qquad \text{for all } n \geq 1 \text{ and } i \in I. \tag{32}$$

The same inequalities apply when y^* *is replaced by* u^*.

Proof. By a last exit decomposition, we have for any $n \geq 1$, $i \in I$ and $j \notin K$,

$$(P_1 \cdots P_n)_{ij} = (\hat{P}_1 \cdots \hat{P}_n)_{ij} + \sum_{k=1}^{n-1} \sum_{h \in K} (P_1 \cdots P_k)_{ih} (\hat{P}_{k+1} \cdots \hat{P}_n)_{hj}$$

By this relation and a repeated application of (31), we get (32).

We can now conclude by Lemma 3 and (3) that for any policy π the quantity $\phi_n(i, \pi)$ in (2) is well-defined and finite.

3. The average cost optimality equation

We first analyse a discounted cost function to derive the average cost optimality equation by a technique developed in [14, 18]. For any $\beta > 0$ and policy π, let

$$V_\beta(i, \pi) = E_\pi\left\{ \sum_{n=0}^{\infty} e^{-\beta \sum_{k=0}^{n-1} \tau(X_k, a_k)} c(X_n, a_n) | X_0 = i \right\}, \quad i \in I$$

and, for any $\beta > 0$, let

$$V_\beta(i) = \inf_\pi V_\beta(i, \pi), \quad i \in I.$$

Using Lemma 3 and (3), it is straightforward to verify that for any $\beta > 0$ the quantity $V_\beta(i, \pi)$ is well-defined for any i, π and that, for constant c_β,

$$|V_\beta(i)| \leq c_\beta y^*(i) \quad \text{for all } i \in I. \tag{33}$$

We now make the following assumption.

Assumption 3. For any $i \in I$, $\sum_{j \in I} p_{ij}(a)\{u^*(j) + y^*(j)\}$ is continuous on $A(i)$.

Note that, by Assumption 1 and a convergence theorem of Scheffé, the sequence of probability distributions $\{(p_{ij}(a_n), j \in I), n \geq 1\}$ converges setwise to the probability distribution $(p_{ij}(a), j \in I)$ as $a_n \to a$. By (33) and the convergence theorem in [15, p. 232], it follows that under this additional assumption 3 $\sum_j p_{ij}(a) V_\beta(j)$ is continuous on $A(i)$ for any $i \in I$. Using this result, a minor modification of the proof of Theorem 6.1 in [14] shows that for any $\beta > 0$ (cf. also [12])

$$V_\beta(i) = \min_{a \in A(i)} \left\{ c(i, a) + e^{-\beta \tau(i,a)} \sum_{j \in I} p_{ij}(a) V_\beta(j) \right\} \quad \text{for all } i \in I. \tag{34}$$

Moreover, let $f_\beta^{(\infty)}$ be any stationary policy such that the action $f_\beta(i)$ minimizes the right side of (34) for all $i \in I$, then

$$V_\beta(i, f_\beta^{(\infty)}) = V_\beta(i) \quad \text{for all } i \in I \tag{35}$$

as may be easily verified by iterating repeatedly the equality

$$V_\beta(i) = c(i, f_\beta(i)) + e^{-\beta \tau(i, f_\beta(i))} \sum_{j \in I} p_{ij}(f_\beta) V_\beta(j), \quad i \in I$$

and using (33) and Lemma 3.

Lemma 4. *Suppose that the Assumptions* 1–3 *hold. Then there are finite numbers* β^*, $\gamma > 0$ *such that for any* $f \in F$ *a state* $s_f \in K$ exists for which

$$|\beta V_\beta(s_f, f^{(\infty)})| \leq \gamma \quad \textit{for all} \quad 0 < \beta < \beta^*$$

and, for any $i \in I$,

$$|V_\beta(i, f^{(\infty)}) - V_\beta(s_f, f^{(\infty)})| \leq \gamma(u^*(i) + y^*(i)) \quad \textit{for all } 0 < \beta < \beta^*.$$

Proof. By Theorem 2, we can choose a constant c and for any $f \in F$ a state $s_f \in K$ such that (17)–(18) hold. Fix now $\beta > 0$ and $f \in F$. We have

$$V_\beta(i, f^{(\infty)}) = E_{f^{(\infty)}}\left\{ \sum_{n=0}^{N(\{s_f\})-1} e^{-\beta \sum_{k=0}^{n-1} \tau(X_k, a_k)} c(X_n, a_n) \mid X_0 = i \right\}$$
$$+ E_{f^{(\infty)}}\{e^{-\beta \sum_{k=0}^{N(\{s_f\})-1} \tau(X_k, a_k)} \mid X_0 = i\} V_\beta(s_f, f^{(\infty)}) \quad \text{for all } i \in I. \tag{36}$$

Taking $i = s_f$ in (36) and using (18) and Assumption 1(c), we derive from (36) that

$$|V_\beta(s_f, f^{(\infty)})| \leq \frac{(y^*(s_f) + c)}{1 - e^{-\beta\varepsilon}}. \tag{37}$$

From (36), (17), (18) and the inequality $1 - e^{-x} \leq x$ for $x \geq 0$, we easily derive

$$|V_\beta(i, f^{(\infty)}) - V_\beta(s_f, f^{(\infty)})| \leq y^*(i) + c + (u^*(i) + c)|\beta V_\beta(s_f, f^{(\infty)})| \quad \text{for all } i \in I. \tag{38}$$

Together (37), (38) and the finiteness of the set K imply the lemma.

We are now in a position to verify the average cost optimality equation.

Theorem 4. *Suppose that the Assumptions* 1–3 *hold. Then there is a constant* g *and a function* $v(i)$, $i \in I$ *with*

$$\sup_{i \in I} \frac{|v(i)|}{u^*(i) + y^*(i)} < \infty \tag{39}$$

satisfying the average cost optimality equation

$$v(i) = \min_{a \in A(i)} \left\{ c(i, a) - g\tau(i, a) + \sum_{j \in I} p_{ij}(a) v(j) \right\} \quad \text{for all } i \in I. \tag{40}$$

Proof. Following [14, 18], fix some state $s \in I$. By (35) and Lemma 4 we can choose finite numbers β^*, $c > 0$ such that for all $0 < \beta < \beta^*$ we have

$$|\beta V_\beta(s)| \leq c \quad \text{and} \quad |V_\beta(i) - V_\beta(s)| \leq c(u^*(i) + y^*(i))$$

for all $i \in I$. Using Assumption 1(a), the diagonalization method and the convergence theorem in [15, p. 232], it is now standard (e.g. cf. [14, p. 146]) to derive from (34) the desired results.

The Assumptions 1–3 are satisfied in the example in [7] for which any average cost optimal policy is nonstationary. Hence an additional assumption is required to guarantee that a stationary policy $f^{(\infty)}$ such that the action $f(i)$ minimizes the right side of (40) for all $i \in I$ is average cost optimal, cf. also [13]. We now state the following condition.

Assumption 4. For any $f \in F$, $\lim_{n \to \infty} \hat{P}^n(f)(u^* + y^*) = 0$ where $\hat{P}^n(f)$ denotes the n-fold matrix product of the substochastic matrix $\hat{P}(f)$ with itself.

Lemma 5. *Suppose that the Assumptions* 1, 2(a), 3 *and* 4 *hold. Then, for any sequence* $(P_1, P_2, \ldots)$ *of stochastic matrices in* $\mathcal{R}$,

$$\lim_{n\to\infty} \frac{1}{n} P_1 \cdots P_n (u^* + y^*)(i) = 0 \quad \text{for all } i \in I.$$

Proof. Define $x_0(i) = u^*(i) + y^*(i)$, $i \in I$ and for $n \geq 1$ define x_n recursively by

$$x_n(i) = \sup_{a \in A(i)} \sum_{j \in K} p_{ij}(a) x_{n-1}(j), \quad i \in I. \tag{41}$$

By the same arguments as in the proof of Lemma 3.7 in [10], we find by Assumption 4 that $x_n(i)$ monotonically decreases to 0 as $n \to \infty$ for any $i \in I$. Now, let $(P_1, P_2, \ldots)$ be any sequence of stochastic matrices in $\mathcal{R}$. By (41), $\hat{P} x_{n-1} \leq x_n$ for all $P \in \mathcal{R}$ and $n \geq 1$ and so $\hat{P}_{k+1} \cdots \hat{P}_n x_0 \leq x_{n-k}$ for any $n \geq 1$ and $k < n$. Using this inequality and Lemma 3, we find

$$P_1 \cdots P_n x_0(i) \leq x_0(i) + \max_{j \in K} x_o(j) + \sum_{k=1}^{n-1} \sum_{h \in K} x_{n-k}(h) \quad \text{for all } n \geq 1 \text{ and } i \in I.$$

Together this inequality, the finitenesss of K and $\lim_{n\to\infty} x_n(i) = 0$ for all i imply the lemma.

We now state our final result.

Theorem 5. *Suppose that the Assumptions* 1–4 *hold. Let* $\{g, v(i) \mid i \in I\}$ *be any finite solution to the average cost optimality equation* (40) *such that* (39) *holds. Choose any stationary policy* $f^{(\infty)}$ *such that the action* $f(i)$ *minimizes the right side of* (40) *for all* $i \in I$. *Then policy* $f^{(\infty)}$ *is strongly average cost optimal and* g *is uniquely determined by* $g = \lim_{n\to\infty} \phi_n(i, f^{(\infty)})$ *for all* $i \in I$.

Proof. Using (39) we have by Lemma 5 and (3) that

$$\lim_{n\to\infty} \frac{1}{n} E_\pi\{|v(X_n)| \mid X_0 = i\} = 0 \quad \text{for any } i \in I \text{ and policy } \pi.$$

Now, by observing that we can replace τ_k by $\tau(X_k, a_k)$ in (2) a repetition of the well-known proof of Theorem 7.6 in [14] gives the desired result.

Remark. Suppose the Assumptions 1–4 hold and let $\{g, v(i) \mid i \in I\}$ be a solution to (40) such that (39) holds. Then the function $v(i)$, $i \in I$ is uniquely determined up to an additive constant under the regularity condition to be stated below. Therefore note first that, by (17), for any $f \in F$ the stochastic matrix $P(f)$ has a unique stationary probability distribution $\{\pi_j(f), j \in I\}$. Suppose now that for any strongly average cost optimal stationary policy the total expected cost incurred until the first return to the finite set K is finite for any starting state when the one-step costs in state i are given

by $u^*(i)+y^*(i)$. Then, using a standard ergodic theorem, we have that for any strongly average cost optimal stationary policy $f^{(\infty)}$ the Cesaro limit of the sequence $\{P^n(f)(u^*+y^*)(i), n \geq 1\}$ equals the finite number $\sum_j \pi_j(f)(u^*(j)+y^*(j))$ for any $i \in I$. Next a minor modification on the proof of Lemma 3 in [11] shows that the function $v(i)$, $i \in I$ is uniquely determined up to an additive constant.

We finally note that it was pointed out by Professor M. Schäl that in Assumption 1 the continuity of the one-step costs $c(s, a)$ can be weakened to lower semi-continuity.

References

[1] J. Bather, Optimal decision procedures for finite Markov chains, Parts I, II and III, Advances in Appl. Probability 5 (1973) 328–339, 521–540, 541–553.

[2] C. Derman and R. Strauch, A note on memoryless rules for controlling sequential control processes, Ann. Math. Statist. 37 (1966) 276–278.

[3] C. Derman, Denumerable state Markovian decision processes-average cost criterion, Ann. Math. Statist. 37 (1966) 1545–1554.

[4] A. Federgruen and H.C. Tijms, The optimality equation in average cost denumerable state semi-Markov decision problems, recurrency conditions and algorithms, J. Appl. Probability 15 (1978) 356–373.

[5] A. Federgruen, A. Hordijk and H.C. Tijms, Recurrence conditions in denumerable state Markov decision processes, in: M.L. Puterman, Ed., Dynamic Programming and its Applications (Academic Press, New York, 1978).

[6] A. Federgruen, A. Hordijk and H.C. Tijms, A note on simultaneous recurrence conditions on a set of denumerable stochastic matrices, J. Appl. Probability 15 (1978) 842–847.

[7] L. Fisher and S.M. Ross, An example in denumerable decision processes, Ann. Math. Statist. 39 (1968) 674–675.

[8] K. van Hee, A. Hordijk and J. van der Wal, Successive approximations for convergent dynamic programming, in: H.C. Tijms and J. Wessels, Eds., Markov Decision Theory, Mathematical Centre Tract No. 93 (Mathematisch Centrum, Amsterdam 1977).

[9] A. Hordijk, Dynamic Programming and Markov Potential Theory, Mathematical Centre Tract No. 51 (Mathematisch Centrum, Amsterdam, 1974).

[10] A. Hordijk, Regenerative Markov decision models, in: R.J.B. Wets, Ed., Mathematical Programming Study 6 (North-Holland, Amsterdam, 1976) 49–72.

[11] A. Hordijk, P.J. Schweitzer and H.C. Tijms, The asymptotic behaviour of the minial total expected cost for the denumerable state Markov decision model, J. Appl. Probability 12 (1975) 298–305.

[12] S.A. Lipman, On dynamic programming with unbounded rewards, Management Sci. 21 (1975) 717–731.

[13] D.R. Robinson, Markov decision chains with unbounded costs and applications to the control of queues, Advances in Appl. Probability 8 (1976) 159–176.

[14] S.M. Ross, Applied Probability Models with Optimization Applications (Holden-Day, San Francisco, 1970).

[15] H.L. Royden, Real Analysis (MacMillan, New York, 2nd ed., 1968).

[16] M. Schäl, Ein verallgemeinertes stationäres entscheidungs modell der dynamische optimierung, in: Methods of Operations Research 10 (Anton Hain, Meisenheim, 1970) 145–162.

[17] M. Schäl, On negative dynamic programming with irreducible Markov chains and the average cost criterion, in: Dynamische Optimierung 98 (Bonner Mathematische Schriften, Bonn, 1977) 93–97.

[18] H.M. Taylor, Markovian sequential replacement processes, Ann. Math. Statist. 36 (1965) 1677–1694.

[19] J. Wijngaard, Stationary Markovian decision problems and perturbation theory of quasi-compact linear operators, Math. Op. Res. 2 (1977) 91–102.