COUNTABLE STATE MARKOVIAN DECISION PROCESSES UNDER THE DOEBLIN CONDITIONS

Kadota, Yoshinobu
Department of Mathematics, Faculty of Education, Wakayama University | Department of Mathematics, Faculty of Science, Kyushu University

https://doi.org/10.5109/13133

出版情報：統計数理研究. 18 (3/4), pp.85-94, 1979-03. Research Association of Statistical Sciences
バージョン：
権利関係：

COUNTABLE STATE MARKOVIAN DECISION PROCESSES UNDER THE DOEBLIN CONDITIONS

By

Yoshinobu KADOTA*

(Received November 20, 1977. Revised September 10, 1978)

1. Introduction.

We consider a discrete time stationary Markovian decision process with a countable state space, finite action sets and bounded immediate rewards. Our concern is with the non-discounted case. If the state space is finite, Blackwell [1], Miller and Veinott [10] and Veinott [11] have made studies on the problem. The purpose of this paper is to extend some results of them to the countable state space.

Section 2 provides the requisite notation and definitions. Section 3 deals with the Laurent series expansion obtained by Miller and Veinott [10]. We assume that the transition probability function associated with each stationary policy satisfies the Doeblin condition. Then, we show that the expansion is extended to our case. Section 4 deals with n-discount optimality criteria for $n=-1$ and 0 in the sense of Veinott [11], where -1-discount optimality coincides with average optimality. We propose a uniformity of the Doeblin conditions and call it Condition (UD). Under (UD) we obtain the following results: There is a 0-discount optimal stationary policy, which is a limit of ρ-discount optimal stationary policies as ρ tends to 0. This policy gives a solution of the functional equation. The policy improvement procedure and the successive approximation method can be effectively used for our case. In Section 5 a sufficient condition for (UD) is introduced. The condition is more intuitive than (UD) and not equivalent to the simultaneous Doeblin condition of Hordijk [6].

Recently, Taylor [12] and Hodijk and Sladky [8] extended the Laurent expansion to the (countably) infinite state space. Hordijk [7] investigated the same problem as the present paper. But, in the literatures [7], [8] and [12] the conditions for the transition probability functions seem rather restrictive in contrast with those in the present paper. Our conditions are so week that the values of the average rewards are allowed to depend on an initial state.

* Department of Mathematics, Faculty of Science, Kyushu University, Fukuoka. Now at Department of Mathematics, Faculty of Education, Wakayama University, Wakayama.

2. Preliminaries.

We are concerned with a system which is observed at times $n=1, 2, \cdots$. At each time the system is found in one of possible states. Let S denote the space of possible states and assume S to be countable. For any state $i\in S$, let A_i be a finite set of possible actions. If the system is observed in state i at time n, an action a is chosen from A_i. Then, we receive an immediate reward $r(i, a)$ and the system moves to a new state j at time $n+1$ according to the conditional probability $p(\{j\}\,|\,i, a)$ depending only on i, a and j. We assume $\sum_{j\in S}p(\{j\}\,|\,i, a)=1$ and that there is a positive number M such that $|r(i, a)|\leqq M$, for all $i\in S$ and $a\in A_i$. The process specified by the four objects S, A, p, r is called a stationary Markovian decision process.

Let $F=\times_{i\in S}A_i$. A (non-randomized Markovian) *policy* is a sequence $\pi=(f_1, f_2, \cdots)$ of members f_n of F. If the policy $\pi=(f_1, f_2, \cdots)$ is used, the action chosen in state i at time n is determined by $f_n(i)$, the i-th component of f_n. Let denote the *stationary* policy $(f, f, \cdots)$ by $f^{(\infty)}$ or f, and the policy $(g, f, f, \cdots)$ by $(gf^{(\infty)})$.

For any $f\in F$, let $r(f)$ be the column vector of countable-infinite dimension, whose i-th component is $r(i, f(i))$. Let $P(f)$ be the Markovian matrix of countable-infinite dimension, whose (i, j)-th element is defined by $p_f(i, j)=p(\{j\}\,|\,i, f(i))$, i.e., $P(f)_{ij}=p_f(i, j)$. Denote by $\mathcal{F}$ the family of all subsets of S. Then for any $P(f)$ we have the associated transition probability function p_f on $S\times\mathcal{F}$ by $p_f(i, E)=\sum_{j\in E}p_f(i, j)$. For $\pi=(f_1, f_2, \cdots)$, let $P_n(\pi)=P(f_1)P(f_2)\cdots P(f_n)$ if $n\geqq 1$ and $P_0(\pi)=I$, the identity motrix. If $\pi=f$, we denote $P_n(f)=P(f)^n$ and $p_f^{(n)}(i, E)=\sum_{j\in E}P(f)_{ij}^n$ for i, E.

Let $\beta=(1+\rho)^{-1}$ be the discount factor where $0<\rho\leqq\infty$. We suppress the dependence of β on ρ in the sequel. When a policy $\pi=(f_1, f_2, \cdots)$ is used, the vector of expected total β-discounted rewards starting from each state is given by

$$V_\rho(\pi)=\sum_{n=0}^{\infty}\beta^{n+1}P_n(\pi)r(f_{n+1}) \qquad \text{for} \quad 0\leqq\beta<1\,.$$

Similarly, the vector of expected average rewards per unit time is given by

$$x(\pi)=\liminf_n\,(n+1)^{-1}\sum_{k=0}^{n}P_k(\pi)r(f_{k+1})\,.$$

Let vectors X, Y be of same type. Then, we say that $X\geqq Y$ if each component of X is at least as large as the corresponding component of Y, and $X>Y$ if $X\geqq Y$ and $X\neq Y$. A policy π^* is called *ρ-discount optimal* if $V_\rho(\pi^*)\geqq V_\rho(\pi)$ for all π, and *average optimal* if $x(\pi^*)\geqq x(\pi)$ for all π. Following Veinott [11] we call π^* *n-discount optimal* for $n=-1, 0$, if

$$\liminf_{\rho\to 0+}\,\rho^{-n}(V_\rho(\pi^*)-V_\rho(\pi))\geqq 0 \qquad \text{for all} \quad \pi\,.$$

Our problem is to find those optimal policies. Blackwell [2] has shown that there is a ρ-discount optimal stationary policy for each $\rho>0$. But as is seen in Derman [3], there is no average optimal (stationary) policy in general. We shall need the following condition throughout the paper.

CONDITION (D). For each $f\in F$, p_f satisfies the Doeblin condition.

According to Hordijk [6], a transition probability function p satisfies the Doeblin condition if and only if there are a finite set K, a positive integer l and a positive number c such that

$$p^{(l)}(i, K)\geqq c \quad \text{for all } i\in S. \tag{1}$$

Following Doob [4, Chapter V, § 5], we summarize the ergodic theorem under the Doeblin condition. For p under the condition, let $\{E_a\}_{a=1}^{N}$ denote the finitely many ergodic classes. For any a, let $\{{}_aC_\alpha\}_{\alpha=0}^{d_a-1}$ denote the finitely many non-empty cyclic classes of E_a. Then for any $i\in {}_aC_\alpha$ and $E\in\mathcal{F}$, $q_{a,\alpha}(E)=\lim_{n\to\infty} p^{(nd_a)}(i, E)$ converges exponentially fast uniformly in i, E, and turns out to be a probability measure on ${}_aC_\alpha$ independent of i. For any $i\in S$, let the limits of non-decreasing sequences, $\lambda(i, E_a)=\lim_{n\to\infty} p^{(n)}(i, E_a)$ and $\lambda(i, {}_aC_\alpha)=\lim_{n\to\infty} p^{(nd_a)}(i, {}_aC_\alpha)$. Then they also converge exponentially fast uniformly in i. (Because, using $\lambda(i, E_a)=\sum_{\alpha=0}^{d_a-1}\lambda(i, {}_aC_\alpha)$, we have

$$\begin{aligned} 0&\leqq\lambda(i, {}_aC_\alpha)-p^{(nd_a)}(i, {}_aC_\alpha)\leqq\lambda(i, E_a)-p^{(nd_a)}(i, E_a)\\ &\leqq\sum_{b=1}^{N}\{\lambda(i, E_b)-p^{(nd_a)}(i, E_b)\}\leqq 1-p^{(nd_a)}(i, \bigcup_{b=1}^{N}E_b). \end{aligned} \tag{2}$$

Apply Theorem 5.6 of Doob [4, 207p] to (2).) Let m be an integer satisfying $0\leqq m\leqq d_a-1$. Then, the preceding facts imply that there are constants A and γ with $0\leqq\gamma<1$ such that

$$|p^{(nd_a+m)}(i, {}_aC_\alpha\cap E)-\lambda(i, {}_aC_{\alpha-m})q_{a,\alpha}(E)|\leqq A\gamma^{nd_a+m}, \tag{3}$$

for all integers $n\geqq 0$, $i\in S$ and $E\in\mathcal{F}$, where ${}_aC_{\alpha-m}={}_aC_\beta$ if $\beta=\alpha-m \pmod{d_a}$. We can choose A, $\gamma(<1)$, so large that they are independent of a, α and m, since N and $\{d_a\}$ are finite, Let denote $P^*=\lim_{n\to\infty}(n+1)^{-1}\sum_{k=0}^{n}P^k$, the Cesàro limit. Clearly, we have $P^*=PP^*=P^*P=P^*P^*$ and $p^*(i, S)=1$ for all $i\in S$. For any i, E, P^* is represented in the form

$$p^*(i, E)=\sum_{a=1}^{N}\lambda(i, E_a)\{d_a^{-1}\sum_{\alpha=0}^{d_a-1}q_{a,\alpha}(E)\}. \tag{4}$$

Let

$$s_n(i, E)=\sum_{k=0}^{n}(p^{(k)}(i, E)-p^*(i, E)). \tag{5}$$

Then by (3) and (4), there is a constant B such that

$$|s_n(i, E)|\leqq B \quad \text{for all } i\in S, E\in\mathcal{F} \text{ and } n. \tag{6}$$

In fact B can be chosen less than or equal to $9dNA(1-\gamma)^{-1}$, where d is the least common multiple of $\{d_a\}_{a=1}^{N}$.

Note that $x(f)=P(f)^*r(f)$ for any $f\in F$ under Condition (D).

3. Laurent series expansion of $V_\rho(f)$.

For any $f\in F$, let

$$H_\rho(f)=\sum_{n=0}^{\infty}\beta^{n+1}(P(f)^n-P(f)^*)\,.$$

To obtain the expansion of $V_\rho(f)$, we shall show the existence of $H(f)=\lim_{\rho\to 0+}H_\rho(f)$.

The norm of a bounded vector $X=(X_j)$ is defined by $\|X\|=\sup_{j\in S}|X_j|$. For a matrix $A=(a_{ij})$ such that $\sup_{i\in S}\sum_{j\in S}|a_{ij}|<\infty$, let $A(i, E)=\sum_{j\in E}a_{ij}$. The norm of A is defined by $\|A\|=\sup_{i\in S}\sup_{E\in\mathcal{F}}|A(i, E)|$.

In this section we freely drop the notation f for brevity when no ambiguity results. We shall need two preliminary lemmas.

LEMMA 1. *For any $f\in F$, let $h_\rho(f)=V_\rho(f)-\rho^{-1}x(f)$. Then under Condition (D), there is a bonstant $C=C_f$ such that $\|h_\rho(f)\|\leqq C$ for all $\rho>0$.*

PROOF. Using $(1-\beta)^{-1}=\sum_{n=0}^{\infty}\beta^n$, we obtain

$$h_\rho=\sum_{n=0}^{\infty}\beta^{n+1}(P^n-P^*)r=(1-\beta)\sum_{n=0}^{\infty}\beta^{n+1}\{\sum_{k=0}^{n}(P^k-P^*)r\}\,. \tag{7}$$

From (6) we obtain $\|\sum_{k=0}^{n}(P^k-P^*)r\|\leqq 2BM$. By taking $C=2BM$, (7) establishes the lemma.

LEMMA 2. *Suppose Condition (D) is satisfied. For any $f\in F$, let denote by d the least common multiple of $\{d_a\}_{a=1}^{N}$ which follow from the ergodic theorem. Take any integer $0\leqq m\leqq d-1$. Then,*

$$H_f^{(m)}(i, E)=\lim_{n\to\infty}s_{nd+m}(i, E\,;\, f)$$

converges uniformly in $i\in S$ and $E\in\mathcal{F}$, where s_n is defined by (5) associated with $f\in F$. Furthermore, $H_f^{(m)}(i, E)$ is bounded uniformly in i and E.

PROOF. Without loss of generality we may assume that the state i is transient, i.e., $i\in T=S-\bigcup_{a=1}^{N}E_a$. In fact, if $i\in E_a$ for some a, the equalities (9), (10) below will be reduced to more simple forms. For any $i\in T$, $E\in\mathcal{F}$, it is clear that

$$\sum_{u=0}^{d_a-1}\{\sum_{\alpha=0}^{d_a-1}\lambda(i, {}_aC_{\alpha-u})q_{a,\alpha}(E)-p^*(i, E\cap E_a)\}=0\,. \tag{8}$$

Let $b(k, u, \alpha)=p^{(kd_a+u)}(i, E\cap{}_aC_\alpha)-\lambda(i, {}_aC_{\alpha-u})q_{a,\alpha}(E)$, and let l be an integer $0\leqq l\leqq d_a-1$. From (8) we obtain

$$s_{nd_a+l}(i, E\cap E_a)=\sum_{\alpha=0}^{d_a-1}\sum_{u=0}^{d_a-1}\sum_{k=0}^{n-1}b(k, u, \alpha)+\sum_{u=0}^{l}\sum_{\alpha=0}^{d_a-1}b(n, u, \alpha)+c(l)\,, \tag{9}$$

where

$$c(l)=\sum_{u=0}^{l}\{\sum_{\alpha=0}^{d_a-1}\lambda(i, {}_aC_{\alpha-u})q_{a,\alpha}(E)-p^*(i, E\cap E_a)\}\,.$$

From (3), $\sum_{k=0}^{\infty}|b(k, u, \alpha)|\leqq A\gamma^u(1-\gamma^{d_a})^{-1}$ for any u, α. On the other hand, $c(l)$ is a constant independent of n. Thus, the right hand side of (9) converges absolutely and uniformly in i, E, as $n\to\infty$. Let $nd+m=n_ad_a+m_a$ for every a, where $0\leqq m_a\leqq d_a-1$. Then we have

$$s_{nd+m}(i, E)=\sum_{a=1}^{N}s_{n_ad_a+m_a}(i, E\cap E_a)+\sum_{k=0}^{nd+m}p^{(k)}(i, E\cap T)\,, \tag{10}$$

since $p^*(i, T)=0$ for all $i\in S$. From Theorem 5.6 of Doob [4], $\sum_{k=0}^{\infty}p^{(k)}(i, E\cap T)\leqq$

$A(1-\gamma)^{-1}$. By the way, m_a does not vary as $n\to\infty$ in (10), since d is a multiple for all d_a. Thus, taking $l=m_a$ and $n=n_a$ in (9), (10) converges uniformly in i, E as $n\to\infty$. The boundedness of $H_f^{(m)}(i, E)$ is now immediate. The proof is complete.

For any $f\in F$, take d as in Lemma 2. For i, E, let

$$H_f(i, E)=d^{-1}\sum_{m=0}^{d-1}H_f^{(m)}(i, E).$$

The associated matrix $H(f)$ with H_f is defined by $H(f)_{ij}=H_f(i, \{j\})$.

THEOREM 1. *Suppose Condition (D) is satisfied. Then for any $f\in F$,*

(a) *$H_{\rho, f}(i, E)$ converges to $H_f(i, E)$ uniformly in i, E as $\rho\to 0$, where $H_{\rho, f}(i, E)=\sum_{j\in E}H_\rho(f)_{ij}$.*

(b) *$H_f(i, E)=\sum_{j\in E}H(f)_{ij}$ for any i, E and $\|H(f)\|\leqq B_f$.*

(c) *$h_\rho(f)$ converges to $h(f)=H(f)r(f)$ uniformly in i, and $\|h(f)\|\leqq C_f$.*

PROOF. For part (a), express $H_\rho(i, E)$ in the form similar to (7) and exchange the terms d-th termwise. Using $1-\beta^d=(1-\beta)(\sum_{k=0}^{d-1}\beta^k)$, we obtain

$$H_\rho(i, E)=(\sum_{k=0}^{d-1}\beta^k)^{-1}\sum_{m=0}^{d-1}\beta^{m+1}\{(1-\beta^d)\sum_{n=0}^{\infty}(\beta^d)^n s_{nd+m}(i, E)\}. \tag{11}$$

By Lemma 2, the term put in brackets { } in (11) converges uniformly in i, E as $\beta\to 1$ to the Abel sum which is equal to $H^{(m)}(i, E)$. This implies (a) immediately.

For any $E\in\mathcal{F}$, let r be the characteristic function of E. Then we have from Lemma 1 that $|H_\rho(i, E)|\leqq B$, so that from (a), $|H(i, E)|\leqq B$ for any i, E. Thus for part (b), it sufficies to show the equality. Let $E_0=\{j\in S;\ H_{ij}\geqq 0\}$. It is easy to see that $\sum_{j\in E_0}H_{ij}\leqq B$. Let $\{K_n\}$ be an increasing sequence of finite sets such that $\lim_{n\to\infty}K_n=E$. Using (a) we can exchange the limits $\lim_{\rho\to 0}\lim_{n\to\infty}H_\rho(i, K_n\cap E_0)$. Thus, we have $H(i, E\cap E_0)=\sum_{j\in E\cap E_0}H_{ij}$. Similarly, for $E_1=S-E_0$ we have $H(i, E\cap E_1)=\sum_{j\in E\cap E_1}H_{ij}$, converging absolutely. Again from (a), $H(i, E)=\sum_{n=0}^{1}H(i, E\cap E_n)$. Those three equalities establish (b).

From (b), Hr is well-defined. Using (a), we have $H(i, S)=0$ for all i. Let $r'=\|r\|+r\ (\geqq 0)$. Then, we obtain $h_\rho-Hr=(H_\rho-H)r'$. Using (a), (b) and Lemma 1, part (c) follows at once, which completes the proof.

If S is finite, following Lemma 3(a) and Theorem 2(a) can be obtained by Blackwell [1], and the other results of Lemma 3 and of Theorem 2 obtained by Miller and Veinott [10]. When S is not finite, Taylor [12] and Hordijk and Sladky [8] have obtained Theorem 2(b) under more restricted conditions than Condition (D). Using Theorem 1, the proofs of Lemma 3 are easy then omitted. (See [10] for (d).) Theorem 2 directly follows from Lemma 3.

LEMMA 3. *Suppose Condition (D) is satisfied. Let $f\in F$.*

(a) *Denote $H_0=H$. Then for $0\leqq\rho\leqq\infty$, H_ρ uniquely satisfies $(I-\beta P)H_\rho=H_\rho(I-\beta P)=\beta(I-P^*)$ and $P^*H_\rho=H_\rho P^*=0$.*

(b) *Let $M_\rho=\sum_{n=0}^{\infty}\beta^{n+1}P^n$, then $P^*M_\rho=\rho^{-1}P^*$ and $M_\rho=P^*M_\rho+H_\rho$.*

(c) *Let $L_\rho=\sum_{n=0}^{\infty}\rho^n(-1)^nH^n$, then L_ρ uniquely satisfies $(I+\rho H)L_\rho=L_\rho(I+\rho H)=I$.*

(d) *It holds that $H_\rho=HL_\rho=L_\rho H$.*

THEOREM 2. *Suppose Condition (D) is satisfied. Take any $f\in F$. Then*

(a) *$x(f)$ is the bounded unique solution of $(I-P(f))x=0$, $P(f)^*x=P(f)^*r(f)$, and*

$h(f)$ is the bounded unique solution of $(I-P(f))y=r(f)-x(f)$, $P(f)^*y=0$.

(b) *For $0<\rho<(2\|H(f)\|)^{-1}$, it holds that*

$$V_\rho(f)=\sum_{n=-1}^{\infty} y_n(f)\rho^n, \tag{12}$$

where $y_{-1}(f)=x(f)$ and $y_n(f)=(-1)^n H(f)^{n+1}r(f)$, $n=0, 1, \cdots$.

The (12) is called by Miller and Veinott [10] the Laurent series expansion of $V_\rho(f)$.

Note that in the class of stationary policies -1-discount optimality is equivalent to average optimality.

4. Applications to -1- and 0-discount optimality criteria.

In this section we shall obtain several results with respect to -1- and 0-discount optimality criteria with the aid of the expansion given by Theorem 2. We shall need the following condition which assures a uniformity of Condition (D).

CONDITION (UD). Suppose (D) is satisfied. In addition, we can choose $B=B_f$ in (6) bounded on F, i.e., $\sup_{f\in F}\{B_f\}=D<\infty$.

If S is finite, then Condition (UD) holds. We shall give a more investigation of (UD) in Section 5.

We say a sequence $\{f_n\}\subset F$ converges to $f\in F$, writing $\lim_{n\to\infty} f_n=f$, if for any $i\in S$ there is a positive integer $N=N_i$ such that $f_n(i)=f(i)$ for all $n\geqq N$.

The following lemma will play a fundamental role in our arguments.

LEMMA 4. *Suppose Condition (UD) is satisfied.*

(a) *It holds that $\sup_{f\in F}\{\|h(f)\|\}\leqq 2DM$.*

(b) *Let $\varepsilon(\rho, f)=\sum_{n=1}^{\infty} y_n(f)\rho^n$. Then, $\varepsilon(\rho, f)$ converges uniformly in $f\in F$ (and in $i\in S$ to 0-vector) as ρ tends to 0.*

(c) *If $\lim_{n\to\infty} f_n=f$, then $\lim_{n\to\infty} x(f_n)=x(f)$ and $\lim_{n\to\infty} h(f_n)=h(f)$.*

PROOF. By Theorem 1(b), part (a) is clear. Then using (12), part (b) is obtained. Thus it suffices to show (c). For any i, E and $k=0, 1, \cdots$, $p_g^{(k)}(i, E)$ is continuous in g and so is $Q_m(i, E; g)=(m+1)^{-1}\sum_{k=0}^{m} p_g^{(k)}(i, E)$. From (UD), $Q_m(i, E; g)$ converges to $p_g^*(i, E)$ uniformly in g as $m\to\infty$, then $p_g^*(i, E)$ is continuous in g. Take such $\{f_n\}$ with f as stated in the lemma. For any $\varepsilon>0$ there is a finite set K such that $p_f^*(i, S-K)<\varepsilon/4$. We can take $N=N_i$ so large that $\sum_{j\in K}|P(f_n)_{ij}^*-P(f)_{ij}^*|<\varepsilon/4$ for all $n\geqq N$, and that $r(j, f_n(j))=r(j, f(j))$ for all $j\in K$, $n\geqq N$, using the continuity of $p_g^*(i, E)$ in g and the finiteness of K. For $g\in F$ denote $x(g)=(x(g)_i)_{i\in S}$. Since $x(g)=P(g)^*r(g)$, we have $|x(f_n)_i-x(f)_i|\leqq\varepsilon\|r\|$ for $n\geqq N$. This establishes the first part of (c).

From the boundedness of $\|h(f_n)\|$, we can get a subsequence $\{f_{n_k}\}$ of $\{f_n\}$ such that $\lim_{k\to\infty} h(f_{n_k})_i=h_i^*$ for all $i\in S$. Let $h^*=(h_i^*)$ the column vector, then $\|h^*\|\leqq 2DM$. For every f_{n_k}, consider the latter two equations of Theorem 2(a) substituting $y=h(f_{n_k})$, and let $k\to\infty$. Then we easily obtain that $(I-P(f))h^*=r(f)-x(f)$ and $P(f)^*h^*=0$. Since the equations have a unique solution, we obtain $h^*=h(f)$. The $\{f_{n_k}\}$ is arbitrary, then $\lim_{n\to\infty} h(f_n)=h(f)$, completing the proof.

Let $V_\rho^*=\sup_\pi\{V_\rho(\pi)\}$ and $x^*=\sup_\pi\{x(\pi)\}$. Following Theorems 3, 4 and Corollary

1 are the extensions of Blackwell [1] to the countable state space. Theorem 4 was originaly obtained by Howard [9].

THEOREM 3. *Under Condition (UD), there is a 0-discount optimal stationary policy.*

PROOF. For any decreasing sequence $\{\rho_n\}$ with limit point 0, let $\{f_n\}$ be a sequence such that $V^*_{\rho_n}=V_{\rho_n}(f_n)$, obtained by Bleckwell [2]. Since F is compact, there is a convergent subsequence $\{f_{n_k}\}$ of $\{f_n\}$ with limit point $f\in F$. It suffices to show that f is 0-discount optimal. We may rewrite f_k for f_{n_k} and ρ_k for ρ_{n_k}. Using (12) and Lemma 4, we have $x(f)\geqq x(f')$ for any $f'\in F$. Let $f'=f_k$, then we have again from (12) and Lemma 4 that

$$\lim_{k\to\infty}(V^*_{\rho_k}-V_{\rho_k}(f))=0\,. \tag{13}$$

Take any decreasing sequence $\{\sigma_n\}$ such that $\lim_{n\to\infty}(V^*_{\rho_n}-V_{\sigma_n}(f))\geqq 0$. In the same manner as $\{\rho_n\}$, we obtain subsequences $\{\sigma_k\}$ and $\{g_k\}$ with a limit $g\in F$. Then we have $x(f)=x(g)$ and $h(f)=h(g)$, so that, $\lim_{k\to\infty}(V_{\sigma_k}(f)-V_{\sigma_k}(g))=0$. Thus, using the corresponding equality of (13) for $\{\sigma_k\}$, g, we have $\lim_{k\to\infty}(V_{\sigma_k}(f)-V^*_{\sigma_k})=0$. Since $\{\sigma_k\}$ is arbitrary, the proof is complete.

Note that Theorem 3 remains true for randomized non-Markovian policies, since the existence theorem of Blackwell [2] used in the above proof is valid for those policies.

COROLLARY 1. *Under Condition (UD), $f\in F$ is 0-discount optimal if and only if $x(f)\geqq x(g)$ for all $g\in F$ and $h(f)\geqq h(g)$ whenever $x(f)=x(g)$.*

The proof is easy and left to the reader.

For $i\in S$ and $a\in A_i$, we denote by $p(i;\,a)$ the row vector whose j-th component is $p_a(i,\,j)$. For $f\in F$ let

$$u_f(i,\,a)=p(i;\,a)x(f)-x(f)_i\,,$$

$$v_f(i,\,a)=(r(i,\,a)+p(i;\,a)h(f))-(x(f)_i+h(f)_i)\,.$$

The following theorem gives the policy improvement method with respect to the average reward.

THEOREM 4. *Take $f\in F$ and denote*

$$G_1(i,\,f)=\{a\in A_i;\ u_f(i,\,a)>0\},\ G_2(i,\,f)=\{a\in A_i;\ u_f(i,\,a)=0\ \ and\ \ v_f(i,\,a)>0\},$$

and $G(i,\,f)=G_1(i,\,f)\cup G_2(i,\,f)$. Under Condition (UD), let $g\in F$ such that $g(i)\in G(i,\,f)$ for some i and that $g(i)=f(i)$ whenever $g(i)\notin G(i,\,f)$. Then it holds that $x(g)\geqq x(f)$. More precisely,

(a) *if there is an $i\in S$ such that $g(i)\in G_1(i,\,f)$, then $x(g)>x(f)$;*

(b) *if $g(i)\notin G_1(i,\,f)$ for all $i\in S$ and*

(i) *if there is an $i\in S$ such that $g(i)\in G_2(i,\,f)$ and that i is a (positive) recurrent state of Markov chain p_g, then $x(g)>x(f)$;*

(ii) *if all $i\in S$ such that $g(i)\in G_2(i,\,f)$ are transient states of p_g, then $x(g)=x(f)$.*

PROOF. For every positive integer n and $i\in S$, let $G^n_1(i,\,f)=\{a\in A_i;\ u_f(i,\,a)>1/n\}$, $G^n_2(i,\,f)=\{a\in A_i;\ u_f(i,\,a)=0$ and $v_f(i,\,a)>1/n\}$ and $G^n(i,\,f)=G^n_1(i,\,f)\cup G^n_2(i,\,f)$.

We define $g_n \in F$ by $g_n(i)=g(i)$ if $g(i) \in G^n(i, f)$ and $g_n(i)=f(i)$ if $g(i) \notin G^n(i, f)$. It is clear that $\lim_{n\to\infty} g_n=g$. Using (12), we obtain a positive number ρ_n for each n such that $V_\rho(g_n f^{(\infty)}) \geqq V_\rho(f)$ for all $0<\rho<\rho_n$. Then by Theorem 8(a) of Blackwell [2] we have $V_\rho(g_n) \geqq V_\rho(g_n f^{(\infty)}) \geqq V_\rho(f)$ for $0<\rho<\rho_n$. Multiply the inequalities by ρ and let $\rho \to 0$. Thus we have form (12) that $x(g_n) \geqq x(f)$, so that from Lemma 4(c), $x(g) \geqq x(f)$. To establish (a), let $g(i_0) \in G_1(i_0, f)$. Then there is a positive integer N such that $g_n(i_0)=g(i_0) \in G_1^N(i_0, f)$ for all $n \geqq N$. Therefore we have from (12) that

$$V_\rho(g_n)_{i_0} \geqq V_\rho(g_n f^{(\infty)})_{i_0} > V_\rho(f)_{i_0} + \{\rho N(1+\rho)\}^{-1} + B(f, g, \rho)_{i_0} \quad \text{for } 0<\rho<\rho_n, \tag{14}$$

where $B(f, g, \rho)_{i_0}$ is bounded uniformly in f, g and ρ. Using (14) similar to the case $x(g) \geqq x(f)$, we obtain $x(g)_{i_0} \geqq x(f)_{i_0}+1/N$, so that, $x(g) > x(f)$.

Next we prove (b). Since $u_f(i, g(i))=0$ for all $i \in S$, we have $x(f)=P(g)x(f)=P(g)^* x(f)$. Then,

$$x(g)_i - x(f)_i = \sum_{j\in S} p_g^*(i, j) v_f(j, g(j)) . \tag{15}$$

Note from Theorem 2(a) that $v_f(j, f(j))=0$ for all $j \in S$. Thus $v_f(j, g(j))$ is positive or 0 according to $g(j) \in G_2(j, f)$ or $g(j)=f(j)$. If i is recurrent of p_g, then i is positive recurrent, i.e., $p_g^*(i, i)>0$. If j is transient of p_g, then $p_g^*(i, j)=0$ for all $i \in S$. Thus (15) shows (i) and (ii) of (b), completing the proof.

The proof of (b) is essentially similar to that of Lemma 1 in Derman [3]. Note that (b) is verified only by using Condition (D).

The following corollary shows the relation between 0-discount optimality and the functional equation in the average reward criterion.

COROLLARY 2. *If $f \in F$ is average optimal, it satisfies $x(f)=\max_{g\in F}\{P(g)x(f)\}$. Furthermore, if f is 0-discount optimal, then it satisfies*

$$x(f)+h(f)=\max_{g\in F_1}\{r(g)+P(g)h(f)\}, \tag{16}$$

where $F_1=\{g \in F ;\ P(g)x(f)=x(f)\}$.

PROOF. If f is average optimal, it follows from Theorem 4(a) that $G_1(i, f)=\phi$ for all i. This implies the first assertion of the corollary at once. Next, let f be 0-discount optimal. Suppose (16) is not true. Since $x(f)+h(f)=r(f)+P(f)h(f)$, there are $g \in F_1$ and $i_0 \in S$ such that $g(i_0) \in G_2(i_0, f)$ and that $g(i)=f(i)$ for all $i \neq i_0$. Using Theorem 4 we have $x(g) \geqq x(f)$, so that, $x(g)=x(f)$. Then it will be easy to see that $h(g) \geqq h(f)$, using (12) and Theorem 8(a) of Blackwell [2]. Take $\delta>0$ such that $v_f(i_0, g(i_0)) \geqq \delta$. Then we have from (12) that

$$V_\rho(gf^{(\infty)})_{i_0} - \rho^{-1}x(f)_{i_0} \geqq \beta\{\delta + h(f)_{i_0} + \sum_{n=1}^{\infty} p(i, g(i)) y_n(f)\rho^n\} .$$

Since $V_\rho(g) \geqq V_\rho(gf^{(\infty)})$, letting $\rho \to 0$ yields $h(g)_{i_0} \geqq h(f)_{i_0}+\delta$, so that, $h(g) > h(f)$. But $x(g)=x(f)$ and $h(g)>h(f)$ contradict Lemma 4, because f is 0-discount optimal. Thus we obtain (16), completing the proof.

Corollary 1 means that the policy improvement procedure of Theorem 4 will terminate if a 0-discount optimal stationary policy is obtained. Conversely, if x^* is a constant vector, (16) coincides with the functional equation of Theorem 1 in Derman

[3]. Then if the procedure terminates, the obtained stationary policy is average optimal. This is not true if x^* is not constant as is seen in Example 2 of Blackwell [1].

The following theorem gives the approximation method to obtain the values $x^*=(x_i^*)_{i\in S}$.

THEOREM 5. *Suppose Condition (UD) is satisfied. Construct a sequence* $\{w_n\}$ *of vectors by* $w_0(i)=0$ *for all* $i\in S$ *and by*

$$w_n(i)=\max_{a\in A_i}\{r(i,a)+p(i;a)w_{n-1}\} \qquad \text{for} \quad n=1,2,\cdots, \quad \text{and} \quad i\in S.$$

Then it follows that $\lim_{n\to\infty} w_n(i)/n=x_i^*$.

PROOF. According to Hordijk [5], if there are constants ρ_0 and M which satisfy

$$\|\rho V_\rho^*-\sigma V_\sigma^*\|\leqq|\rho-\sigma|\cdot M \tag{17}$$

for all ρ, σ such that $0<\rho$, $\sigma<\rho_0$, then it holds that $\lim_{n\to\infty} w_n/n=\lim_{\rho\to 0+}\rho V_\rho^*$. We have from Theorem 3 and Corollary 1 that $\lim_{\rho\to 0+}\rho V_\rho^*=x^*$. Therefore to establish the theorem, it suffices to show (17). Using the existence theorem in Blackwell [2] and the properties of maxima, we obtain

$$\|\rho V_\rho^*-\sigma V_\sigma^*\|\leqq\sup_{f\in F}\|\rho V_\rho(f)-\sigma V_\sigma(f)\|. \tag{18}$$

Therefore, if we show instead of (17) that

$$\|\rho V_\rho(f)-\sigma V_\sigma(f)\|\leqq|\rho-\sigma|\cdot M \qquad \text{for all} \quad f, \tag{19}$$

(18) implies (17) directly. The verification of (19) is easy, using (12), $\|y_n(f)\|\leqq(2D)^{n+1}M$ and $\sum_{n=0}^{\infty}(n+1)z^n=(1-z)^{-2}$. Thus we complete the proof.

5. Comments on Condition (UD).

In [6] Hordijk introduced a condition which also gives a uniformity of the Doeblin conditions and he called it the condition sim D. The sim D assures the existence of a average optimal stationary policy with some additional assumptions. Condition (UD) does not imply the sim D. A counterexample will be easily made, because (UD) needs the existence of a positive integer n such that $\inf_{f\in F} p_f^{(n)}(i,\bigcup_{a=1}^{N}E_a)>0$ for all $i\in S$, whereas the sim D does not need necessarily. We don't know whether the converse is true or not.

At the end of this paper, we shall give a sufficient condition for (UD). Condition (LD) below is more intuitive than (UD). (LD) may be similar to the sim D apparently but they are not equivalent.

LEMMA 5. *Suppose a transition prabability function* p *satisfies the Doeblin condition. Then in the class of D-tuples* (K, l, c) *which satisfy* (1), *there is a D-tuble such that* (a) $K\subset\bigcup_{a=1}^{N}E_a$ *and* (b) *for every* a, α, $K\cap_a C_\alpha$ *consists of only one state.*

PROOF. For any D-tuple (K, l, c), check the ergodic theorem for Cases (a), (b), (c) and (d) in Doob [4]. Part (a) of the lemma follows from Theorem 5.6 of [4]. Note that for any $j\in K\cap_a C_\alpha$ and all $i\in_a C_\alpha$, there are a positive integer n and a positive

number δ such that $\tilde{p}^{(nd_\alpha)}(i, j) \geqq \delta$, where $\tilde{p}$ is defined by [4]. On the other hand, it holds that $p^{(md_\alpha)}(i, K \cap {}_\alpha C_\alpha) \geqq c$ for all $i \in {}_\alpha C_\alpha$ and m with $md_\alpha \geqq l$. Those two facts imply the lemma. The precise proof is lengthy and left to the reader.

It is clear that the number of elements of K obtained by Lemma 5 is the least among the class of finite sets of D-tuples.

CONDITION (LD). Suppose Condition (D) is satisfied. For each p_f, take the D-tuple (K_f, l_f, c_f) following Lemma 5. Then we can choose l_f and c_f independent of $f \in F$, i.e., $p_f^{(l)}(i, K_f) \geqq c$ for all $f \in F$ and $i \in S$.

THEOREM 6. *Condition (LD) implies Condition (UD).*

POOF. From (LD), the $\rho(<1)$ of Theorem 5.6 in [4] is independent of $f \in F$. For any f and α, a, $p_f^{(jd_\alpha)}(i, j) \geqq c$ for all $i \in {}_\alpha C_\alpha$, where $\{j\} = K \cap {}_\alpha C_\alpha$. Therefore, the ρ in Case (b) of [4] is independent of f, a and α for all noncyclic Markov chains $\{p_f^{(d_\alpha)}\}$. The precise proof is left to the reader.

Acknowledgement

The author wishes to express his sincere appreciation to Professor N. Furukawa whose advices and encouragements helped make this paper possible. The author also wishes to thank Dr. S. Iwamoto for his useful comments.

References

[1] BLACKWELL, D. (1962). *Discrete dynamic programming*, Ann. Math. Statist. **33**, 719-726.
[2] BLACKWELL, D. (1965). *Discounted dynamic programming*. Ann. Math. Statist. **36**, 226-235.
[3] DERMAN, C. (1966). *Denumerable state Markovian decision process—average cost criterion*. Ann. Math. Statist. **37**, 1545-1554.
[4] DOOB, J.L. (1953). Stochastic Processes. Wiley, New York. Chapter V, § 5.
[5] HORDIJK, A. (1974). *On the convergence of the average expected return in dynamic programming*. J. Math. Anal. **46**, 542-544.
[6] HORDIJK, A. (1974). Dynamic Programming and Markov Potential Theory. Math. Centre Tracts, **51**, Amsterdam.
[7] HORDIJK, A. (1976). *Regenerative Markov decision models*. Math. Programming Study **6**. North-Holland, Amsterdam. 49-72.
[8] HORDIJK, A. and SLADKY, K. (1977). *Sensitive optimality criteria in countable state dynamic programming*. Math. Operations Res. **2**, 1-14.
[9] HOWARD, R.A. (1960). Dynamic Programming and Markov Processes. Technology Press, Cambridge, Massachusetts.
[10] MILLER, B.L. and VEINOTT, A.F., JR. (1969). *Discrete dynamic programming with a small interest rate*. Ann. Math. Statist. **40**, 366-370.
[11] VEINOTT, A.F., JR. (1969). *Discrete dynamic programming with sensitive discount optimality criteria*. Ann. Math. Statist. **40**, 1635-1660.
[12] TAYLOR, H.M. (1976). *A Laurent series for the resolvent of a strongly continuous stochastic semi-group*. Math. Programming Study **6**. North-Holland, Amsterdam. 258-263.