A LIMIT THEOREM FOR MARKOV CHAINS AND HORDIJK-TIJMS ITERATIVE METHOD

By

Yoshinobu KADOTA

(Received Oct. 29, 1975)

§ 1. Introduction and summary

A discrete-time stationary Markov chain with a countable state space is naturally represented by the Markov matrix P , whose (i, j) -element denotes the probability from a state i to a next state j . As is well known, there exists the Cesàro limit of sequence $\{P^n; n=0, 1, 2, \dots\}$, i.e. the limit of average probabilities per unit time $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k$ (say P^*). Let $Q_n = \sum_{k=0}^{\infty} c_{n,k} P^k$. If $c_{n,k} = \frac{1}{n}$ for $0 \leq k \leq n-1$ and $c_{n,k} = 0$ for all $k \geq n$, then $\lim_{n \rightarrow \infty} Q_n$ is reduced to the Cesàro limit P^* .

The first purpose of this paper is to give a limit theorem for $\{Q_n\}$, i.e., a general sufficient condition for $(c_{n,k})$ under which Q_n converges to P^* as n tends to infinity. This is done in Section 2 and the condition given there contains the case when Q_n is of the form $\frac{1}{n} \sum_{k=0}^{n-1} P^k$ as particular case. The second purpose is to apply the limit theorem to the modified iterative method of Hordijk-Tijms [5] in dynamic programming, and show that their condition can be made more loose.

A Markov decision process (which will be abbreviated as "MDP" henceforth) is given by four tuple $\mathcal{M} = (S, F, \{P(\cdot)\}, \{r(\cdot)\})$ in Section 3. S is a finite state space. F is a finite set consisting of mappings from S to a preassigned finite action space. For every $f \in F$, $P(f)$ and $r(f)$ are an S -dimensional Markov matrix and an S -dimensional column vector, respectively.

We shall be concerned with MDP with the average return criterion. Blackwell [1] proved that in MDP of above sense there always exists an optimal strategy which is stationary, and Howard [6] gave a policy-iteration algorithm for finding the optimal strategy and the optimal return. On the other hand, Hordijk-Tijms [5] proposed an iterative method and showed that the method enables us

to get an approximation to the optimal return under the condition that the optimal return is independent of initial state. Compared with the method of Howard [6], their method has an advantage that we need not solve a system of equations of average return and relative values at each stage of iterations.

This paper consists of four sections. In Section 2 we shall give a new limit theorem for Markov chains. In Section 3 we shall give formulations of MDPs and some preliminary lemmas. In Section 4, we shall apply our limit theorem for Markov chains to the iterative method given by Hordijk and Tijms and show that the method is valid even under some more loose condition than the condition of them. Their condition is that the optimal return should be independent of initial state. Our condition is introduced by Assumption H in Section 4. Assumption H includes the condition of Hordijk-Tijms [5] as a special case. Theorem 4.2 is a modified version of Corollary 4.3 in Brown [2], and Theorem 4.3 is a modified one of Corollary 4.4 in [2].

§2. A limit theorem for Markov chains

In this section we shall give a limit theorem for Markov chains, which is an extension of the mean ergodic limit theorem in the usual sense. The extended limit theorem will play a fundamental role in the arguments in Section 4.

Let S be a finite or denumerable-infinite state space, and P a transition matrix of a Markov chain with the state space S . For any $i, j \in S$, the (i, j) -element of P , which means the one-step transition probability from i to j , is denoted by p_{ij} . We assume that $p_{ij} \geq 0$ for all $i, j \in S$ and $\sum_{j \in S} p_{ij} = 1$ for all $i \in S$. It is well known that there exists the Cesàro limit of the powers of P , $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k = P^*$, where the limit is taken element-wise. In the subsequent arguments it will be shown that P^* is defined as a limit other than the Cesàro limit.

Let $\mathfrak{T}$ be the family of all linear transformations from the space of real vectors of denumerable-infinite dimension to itself. We shall identify a linear transformation $T \in \mathfrak{T}$ and the infinite-dimensional matrix $(c_{n,k})$ defined by T . $\mathfrak{T}_c$ is defined as the family of all elements $T = (c_{n,k})$ of $\mathfrak{T}$ such that for any convergent sequence $\{s_k\}$ of real numbers, $\sum_{k=0}^{\infty} c_{n,k} s_k$ converges to some limit as $n \rightarrow \infty$.

Let $t_n = \sum_{k=0}^{\infty} c_{n,k} s_k$. Then $\mathfrak{T}_r$ is defined as the family of all elements of $\mathfrak{T}_c$ such that

$\lim_{n \rightarrow \infty} t_n = \lim_{k \rightarrow \infty} s_k$. Any element of $\mathfrak{T}_r$ is said to be *regular*.

The following propositions have been given by Hardy [4].

PROPOSITION 2.1. *In order that T should belong to $\mathfrak{T}_c$, it is necessary and sufficient that*

- (i) $\sum_{k=0}^{\infty} |c_{n,k}| < H$, where H is independent of n ,
- (ii) $c_{n,k} \rightarrow \delta_k$ for each k as $n \rightarrow \infty$

and

- (iii) $c_n = \sum_{k=0}^{\infty} c_{n,k} \rightarrow \delta$ as $n \rightarrow \infty$. In these circumstances it holds that $\sum_{k=0}^{\infty} \delta_k$ is absolutely convergent, and

$$t_n \longrightarrow t = \delta s + \sum_{k=0}^{\infty} \delta_k (s_k - s) = s(\delta - \sum_{k=0}^{\infty} \delta_k) + \sum_{k=0}^{\infty} \delta_k s_k$$

as $n \rightarrow \infty$, whenever $s_k \rightarrow s$.

PROPOSITION 2.2. *In order that T should belong to $\mathfrak{T}_r$ (i.e. that T should be regular), it is necessary and sufficient that the conditions of Proposition 2.1 should be satisfied, that $\delta_k = 0$ for each k , and that $\delta = 1$.*

We now prepare the following

LEMMA 2.1. *Let $T = (c_{n,k})$ be a regular transformation such that $\{c_{n,k}\}$ be non-increasing in k for each n . Let $\{s_k^{(l)}\}_{k=0,1,\dots}$ be a convergent sequence of real numbers for $l=0, 1, \dots, d-1$. Let $s^{(l)} = \lim_{k \rightarrow \infty} s_k^{(l)}$ for $l=0, 1, \dots, d-1$, let $t_{md+l} = s_m^{(l)}$ for $m=0, 1, \dots$ and $l=0, 1, \dots, d-1$ and let $u_n = \sum_{k=0}^{\infty} c_{n,k} t_k$. Then it holds that*

$$(2.1) \quad u_n \longrightarrow \frac{1}{d} \sum_{l=0}^{d-1} s^{(l)} \quad \text{as } n \longrightarrow \infty.$$

PROOF. Take an arbitrary integer $l (0 \leq l \leq d-1)$. Then we have

$$\begin{aligned} u_{nd+l} &= \sum_{k=0}^{\infty} c_{nd+l,k} t_k \\ &= \sum_{m=0}^{d-1} \sum_{k=0}^{\infty} c_{nd+l,kd+m} s_k^{(m)} \\ &= \sum_{m=0}^{d-1} \sum_{k=0}^{\infty} a_{n,k}^{(m)} s_k^{(m)}, \end{aligned}$$

where $a_{n,k}^{(m)} = c_{nd+l, kd+m}$.

It is clear that for each m , $(a_{n,k}^{(m)})$ satisfies the conditions (i) and (ii) with $\delta_k = 0$ in Proposition 2.1. Next we shall prove that $(a_{n,k}^{(m)})$ satisfies the condition (iii) with $\delta = \frac{1}{d}$ in Proposition 2.1. It is easily shown that

$$(2.2) \quad \sum_{m=0}^{d-1} \sum_{k=0}^{\infty} a_{n,k}^{(m)} = \sum_{k=0}^{\infty} c_{nd+l, k} = c_{nd+l}.$$

Since $\{c_{nd+l, k}\}$ is non-increasing in k , we have

$$(2.3) \quad \begin{aligned} \sum_{k=0}^{\infty} a_{n,k}^{(0)} - a_{n,0}^{(0)} &\leq \sum_{k=0}^{\infty} a_{n,k}^{(d-1)} \leq \sum_{k=0}^{\infty} a_{n,k}^{(d-2)} \leq \dots \\ &\dots \leq \sum_{k=0}^{\infty} a_{n,k}^{(1)} \leq \sum_{k=0}^{\infty} a_{n,k}^{(0)}. \end{aligned}$$

From (2.2) and (2.3) we get

$$(2.4) \quad d \left(\sum_{k=0}^{\infty} a_{n,k}^{(0)} \right) - da_{n,0}^{(0)} \leq c_{nd+l} \leq d \left(\sum_{k=0}^{\infty} a_{n,k}^{(0)} \right).$$

Taking the supremum and the infimum as n tends to infinity on the left-hand side and on the right-hand side of (2.4), respectively, we have from Proposition 2.2 that

$$\overline{\lim}_{n \rightarrow \infty} d \sum_{k=0}^{\infty} a_{n,k}^{(0)} \leq 1 \leq \underline{\lim}_{n \rightarrow \infty} d \sum_{k=0}^{\infty} a_{n,k}^{(0)}.$$

The inequalities imply that

$$(2.5) \quad \lim_{n \rightarrow \infty} \sum_{k=0}^{\infty} a_{n,k}^{(0)} = \frac{1}{d}.$$

From (2.3) and (2.5)

$$\lim_{n \rightarrow \infty} \sum_{k=0}^{\infty} a_{n,k}^{(m)} = \frac{1}{d} \quad \text{for } m=0, 1, \dots, d-1.$$

Thus we have seen that $(a_{n,k}^{(m)})$ satisfies the condition (iii) with $\delta = \frac{1}{d}$. Hence $(a_{n,k}^{(m)})$ belongs to $\mathfrak{T}_c$. From the last statement of Proposition 2.1, we have

$$(2.6) \quad \lim_{n \rightarrow \infty} u_{nd+l} = \frac{1}{d} \sum_{m=0}^{d-1} s^{(m)}.$$

Since l is arbitrary, (2.6) implies that (2.1) holds. This completes the proof. In the subsequent arguments P^0 means the identity matrix I .

THEOREM 2.1. *Let $T=(c_{n,k})$ be a regular transformation having the same property as the one in Lemma 2.1. Then it holds that*

$$(2.7) \quad \lim_{n \rightarrow \infty} \sum_{k=0}^{\infty} c_{n,k} P^k = P^*.$$

PROOF. For each $i, j \in S$ we shall check the behavior of $p_{ij}^{(n)}$ as $n \rightarrow \infty$. According to Chung [3] there occur the following cases. For symbols and terms not defined here, references are made to [3].

- I. If j is nonrecurrent, then $\lim_{n \rightarrow \infty} p_{ij}^{(n)} = 0$ for all $i \in S$.
- II. If j is recurrent with period d_j and mean recurrence time m_{jj} , and
 - (a) if i belongs to a different essential class, then $p_{ij}^{(n)} = 0$ for every n ;
 - (b) if i belongs to the same class as j so that $i \in C_l(j)$, then $\lim_{n \rightarrow \infty} p_{ij}^{(nd_j+l)} = \frac{d_j}{m_{jj}}$ and $p_{ij}^{(n)} = 0$ if $n \not\equiv l \pmod{d_j}$;
 - (c) if i is inessential, then for every l ,

$$\lim_{n \rightarrow \infty} p_{ij}^{(nd_j+l)} = f_{ij}^*(l) \frac{d_j}{m_{jj}}.$$

On the other hand, for every i and j $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} p_{ij}^{(k)} = \frac{f_{ij}^*}{m_{jj}}$ provided that we define $m_{jj} = \infty$ if j is nonrecurrent.

We shall show element-wise that (2.7) is true, according to the above classification. In cases I and II (a), as is seen above, $p_{ij}^{(n)}$ converges to 0 as $n \rightarrow \infty$. Therefore the limit in the left-hand side of (2.7) is equal to 0, since T is regular. It is clear that the right-hand side is equal to 0.

In cases II (b) and (c), $\{p_{ij}^{(n)}\}$ consists of d_j convergent sequences $\{p_{ij}^{(nd_j+l)}\}_{n=0,1,\dots}$ ($0 \leq l \leq d_j - 1$). Moreover it holds in these cases that $\sum_{l=0}^{d_j-1} f_{ij}^*(l) = f_{ij}^*$ for every $i, j \in S$, and particularly in case II (b) it holds that $f_{ij}^* = 1$. Lemma 2.1 applied to $s_n^{(l)} = p_{ij}^{(nd_j+l)}$ implies that (2.7) holds in both cases.

§3. Formulations of Markov decision processes and some preliminaries

We consider a system with a finite state space $S = \{1, 2, \dots, S\}$. At each time n we observe a current state $i \in S$ of the system and then choose an action a from

a finite set $A(i)$ preassigned to us corresponding to i . Then the system gives an immediate reward $r(i, a)$ to us, and moves to a new state $j \in S$ with probability $p_{ij}(a)$, which is independent of both time n and past history.

Let $F = \prod_{i=1}^S A(i)$. Taking any f from F , f can be written as $f = (f(1), f(2), \dots, f(S))$ where $f(i)$ belongs to $A(i)$ for each i . Any element f of F determines an S -dimensional Markov matrix $P(f)$ whose (i, j) -element is given by $p_{ij}(f(i))$ and an S -dimensional column vector $r(f)$ whose i -th element is given by $r(i, f(i))$.

In terms of the above notation a MDP is given by $\mathcal{M} = (S, F, \{P(\cdot)\}, \{r(\cdot)\})$.

A *strategy* is a sequence $\pi = (f_1, f_2, \dots)$ consisting of elements $f_n \in F$. If $f_n = f$ for all $n \geq 1$ we call the strategy *stationary* and denote it by f^∞ or f . For any $\pi = (f_1, f_2, \dots)$ and any n , let

$$P(f_1 f_2 \cdots f_n) = P(f_1)P(f_2) \cdots P(f_n),$$

where in the case when $n=0$, $P(f_1 f_2 \cdots f_n)$ is interpreted as the identity matrix I . The i -th element of $P(f_1 f_2 \cdots f_n) r(f_{n+1})$ represents the expected return received at the n -th time by $\pi = (f_1, f_2, \dots)$ starting from an initial state i .

We shall be concerned with the optimization under the average return criterion. For any strategy π we define a long run *average expected return* per unit time by

$$x(\pi) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} P(f_1 f_2 \cdots f_k) r(f_{k+1}).$$

The i -th element of $x(\pi)$ corresponds to an initial state i . If π is stationary, then $x(\pi)$ is denoted by $x(f)$ and it is clear that $x(f) = P^*(f) r(f)$, where $P^*(f) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} P(f)^k$.

Let $x^* = \sup_{\pi} \{x(\pi)\}$, where the supremum is taken element-wise. A strategy π is said to be *optimal*, if $x(\pi) = x^*$. We call x^* an *optimal (expected average) return*. It is well-known by Blackwell [1] that there exists a stationary strategy which is optimal in the sense of our average criterion.

We now define sequences $\{\gamma_n\}$, $\{y_n\}$ and $\{\hat{y}_n\}$ according to Hordijk-Tijms [5]. Take an arbitrary sequence $\{\alpha_n; n=1, 2, \dots\}$ of real numbers. Define a sequence $\{\gamma_n\}$ by

$$(3.1) \quad \gamma_0 = 0 \quad \text{and} \quad \gamma_n = 1 + \alpha_n \gamma_{n-1} \quad \text{for} \quad n=1, 2, \dots.$$

Let y_0 be any S -dimensional column vector, and define $\{y_n\}$ successively by

$$(3.2) \quad y_n = \max_{f \in F} \{r(f) + \alpha_n P(f) y_{n-1}\} \quad n = 1, 2, \dots$$

For any $f \in F$ we put

$$(3.3) \quad \hat{r}(f) = r(f) - x^*.$$

Replacing r by $\hat{r}$ in our MDP, we define another MDP $\hat{\mathcal{M}} = (S, F, \{P(\cdot)\}, \{\hat{r}(\cdot)\})$. We denote y_n peculiar to $\hat{\mathcal{M}}$ by $\hat{y}_n$; i.e.,

$$\hat{y}_n = \max_{f \in F} \{\hat{r}(f) + \alpha_n P(f) \hat{y}_{n-1}\} \quad n = 1, 2, \dots$$

where $\hat{y}_0 = y_0$.

If X and Y are vectors of same type, we say $X \leq Y$ if each element of X is equal or less than the corresponding element of Y . The following lemma gives the relation between y_n and $\hat{y}_n$.

LEMMA 3.1. *It holds that $\hat{y}_n \geq y_n - \gamma_n x^*$ for $n = 0, 1, \dots$.*

PROOF. The lemma can be proved by induction on n , by the use of the properties of maxima and of the well-known result that $P(f)x^* \leq x^*$ for all $f \in F$.

In order to study the asymptotic behavior of $\hat{y}_n$, we shall prepare the following two lemmas.

LEMMA 3.2. *Let $\hat{x}^*$ be the optimal return of the MDP $\hat{\mathcal{M}}$. Then $\hat{x}^* = 0$.*

PROOF. Let $\hat{x}(f)$ be the expected average return per unit time received from a stationary strategy $f \in F$. For any $f \in F$, using the relation $P^*(f)^2 = P^*(f)$ we get

$$(3.4) \quad \hat{x}(f) = P^*(f)\hat{r}(f) = P^*(f)\{r(f) - x^*\} = P^*(f)\{x(f) - x^*\} \leq 0,$$

which implies that $\hat{x}^* \leq 0$. It is clear from (3.4) that the equality $\hat{x}^* = 0$ is realized by f such that $x(f) = x^*$.

For any α such that $0 < \alpha < 1$ and any strategy $\pi = (f_1, f_2, \dots)$, let

$$\hat{V}_\alpha(\pi) = \sum_{k=0}^{\infty} \alpha^k P(f_1 f_2 \dots f_k) \hat{r}(f_{k+1}),$$

which means the expected (total) discounted return received from π in the MDP

$\mathcal{M}$ with a discount factor α . The maximal expected discounted return $\hat{V}_\alpha$ is then given by $\hat{V}_\alpha = \sup_\pi \hat{V}_\alpha(\pi)$. There is a stationary optimal strategy $f \in F$ which satisfies $\hat{V}_\alpha(f^\infty) = \hat{V}_\alpha$ for all α close enough to 1 (see [1]). We get from Lemma 3.2 and the Laurent series expansion for α near 1 given by Miller-Veinott [7] that

$$(3.5) \quad \hat{V}_\alpha = \hat{V}_\alpha(f^\infty) = (1 + \rho) \sum_{k=0}^{\infty} (-\rho)^k H(f)^{k+1} \hat{r}(f),$$

where $H(f) = (I - P(f) + P^*(f))^{-1} - P^*(f)$ and $\rho = \frac{1-\alpha}{\alpha}$.

The following lemma can be proved along the same line as Lemma 1 in [5] by the use of (3.5).

LEMMA 3.3. *There exist a number α_0 and a constant B such that $0 < \alpha_0 < 1$ and*

$$\|\hat{V}_\alpha - \hat{V}_\beta\| \leq |\alpha - \beta| \cdot B \quad \text{for all } \alpha, \beta \in (\alpha_0, 1),$$

where $\|\cdot\|$ is defined by $\|X\| = \max_{1 \leq i \leq s} |X_i|$ for any real vector $X = (X_i)_{i=1,2,\dots,s}$.

Let $\{\alpha_n\}$ be such that (i) $0 < \alpha_n < 1$ for all $n \geq 2$; (ii) $\alpha_n \rightarrow 1$ as $n \rightarrow \infty$; (iii) $\alpha_2 \alpha_3 \cdots \alpha_n \rightarrow 0$ as $n \rightarrow \infty$ and (iv) $\sum_{j=2}^n (\alpha_n \alpha_{n-1} \cdots \alpha_j) |\alpha_j - \alpha_{j-1}| \rightarrow 0$ as $n \rightarrow \infty$.

Then we have from (i), (ii) and (3.1) that

$$(3.6) \quad \gamma_n \geq 0 \quad \text{and} \quad \gamma_n \rightarrow \infty \quad \text{as} \quad n \rightarrow \infty.$$

We have, as an example of such sequences, a sequence $\{\alpha_n\}$ such that $\alpha_n = 1 - \frac{1}{n^b}$ for any $\frac{1}{2} \leq b \leq 1$, which appeared in [5].

In all subsequent arguments $\{\alpha_n\}$ is so taken as to satisfy all conditions (i) $\sim$ (iv).

It is easily seen from Lemma 3.3 that the limit of $\hat{V}_\alpha$ as α tends increasingly to 1 does exist as finite. Denote the limit of $\hat{V}_\alpha$ by $\hat{V}_\infty$. The following lemma can be proved by the same method as Theorem 1 in [5].

LEMMA 3.4. *It holds that $\lim_{n \rightarrow \infty} \hat{y}_n = \hat{V}_\infty$.*

§ 4. Application to Hordijk-Tijms iterative method

The following theorem shows that for any stationary strategy $f \in F$, the long run average return weighted by the products of $\{\alpha_n\}$ is equal to the usual average return per unit time.

THEOREM 4.1. *For any stationary strategy $f \in F$, it holds that*

$$(4.1) \quad x(f) = \lim_{n \rightarrow \infty} \frac{1}{\gamma_n} \left(r(f) + \sum_{k=1}^{n-1} \alpha_n \alpha_{n-1} \cdots \alpha_{n+1-k} P(f)^k r(f) \right).$$

PROOF. For each n , define $c_{n,k}$ by

$$c_{n,k} = \begin{cases} \frac{1}{\gamma_n} & \text{for } k=0 \\ \frac{1}{\gamma_n} \alpha_n \alpha_{n-1} \cdots \alpha_{n+1-k} & \text{for } 1 \leq k \leq n-1 \\ 0 & \text{for } k \geq n, \end{cases}$$

where $c_{0,k}=0$ for all $k \geq 0$ and $c_{1,0}=1$ and $c_{1,k}=0$ for all $k \geq 1$. Then it turns out that the matrix $(c_{n,k})$ is regular and $\{c_{n,k}\}$ is non-increasing in k for each n . Hence we obtain (4.1) as an immediate consequence of Theorem 2.1.

Note that in proving Theorem 4.1 we don't need the conditions (iii) and (iv) for $\{\alpha_n\}$.

The following theorem suggests an algorithm for finding an approximate value of the optimal return.

THEOREM 4.2. *It holds that $\lim_{n \rightarrow \infty} \frac{y_n}{\gamma_n} = x^*$.*

PROOF. Let f be a stationary strategy which is optimal. Define $\{Z_n(f)\}$ iteratively by

$$(4.2) \quad Z_n(f) = r(f) + \alpha_n P(f) Z_{n-1}(f) \quad \text{for } n=1, 2, \dots,$$

where $Z_0(f) = y_0$. Using Theorem 4.1 we get

$$x^* = x(f) = \lim_{n \rightarrow \infty} \frac{1}{\gamma_n} \left(r(f) + \sum_{k=0}^{n-1} \alpha_n \alpha_{n-1} \cdots \alpha_{n+1-k} P(f)^k r(f) \right) = \lim_{n \rightarrow \infty} \frac{Z_n(f)}{\gamma_n}.$$

It is clear from (3.2) and (4.2) that $y_n \geq Z_n(f)$. Hence we have

$$(4.3) \quad \lim_{n \rightarrow \infty} \frac{y_n}{\gamma_n} \geq \lim_{n \rightarrow \infty} \frac{Z_n(f)}{\gamma_n} = x^*.$$

On the other hand, we get from Lemmas 3.1 and 3.4 and (3.6) that

$$(4.4) \quad 0 = \overline{\lim}_{n \rightarrow \infty} \frac{\hat{y}_n}{\gamma_n} \geq \overline{\lim}_{n \rightarrow \infty} \left(\frac{y_n}{\gamma_n} - x^* \right) = \overline{\lim}_{n \rightarrow \infty} \frac{y_n}{\gamma_n} - x^*.$$

Comparing (4.3) and (4.4), the theorem is proved.

For each n , define Γ_n by

$$(4.5) \quad \begin{aligned} \Gamma_n &= \{g \in F; y_n = \max_{f \in F} \{r(f) + \alpha_n P(f)y_{n-1}\} \\ &= r(g) + \alpha_n P(g)y_{n-1}\} \quad n = 1, 2, \dots \end{aligned}$$

THEOREM 4.3. *There exists a positive integer N such that for any $n \geq N$ and any $f \in \Gamma_n$ it holds that $P(f)x^* = x^*$.*

PROOF. Choose sufficient large N so that if $f \in \Gamma_n$ for some $n \geq N$, f belongs to Γ_m for infinitely many m . This can be done since the number of stationary strategies is finite. Take any integer $n \geq N$ and any $f \in \Gamma_n$. Then we may suppose from the property of N that $f \in \Gamma_{n_k}$ for $k = 1, 2, \dots$. From the definition of Γ_n

$$y_{n_k} = r(f) + \alpha_{n_k} P(f)y_{n_k-1}.$$

Dividing both sides by γ_{n_k} we get

$$\frac{y_{n_k}}{\gamma_{n_k}} = \frac{r(f)}{\gamma_{n_k}} + \frac{\alpha_{n_k} \gamma_{n_k-1}}{\gamma_{n_k}} P(f) \frac{y_{n_k-1}}{\gamma_{n_k-1}}.$$

Letting $k \rightarrow \infty$, from the condition (ii) for $\{\alpha_n\}$ and (3.6) and from Theorem 4.2 we get $x^* = P(f)x^*$. This completes the proof.

Throughout the subsequent arguments we shall impose the following assumption.

ASSUMPTION H. If $f \in F$ satisfies the equation $P^*(f)(x(f) - x^*) = 0$, then f is stationary optimal, i.e., $x(f) = x^*$.

Assumption H is fulfilled, for example, in the following two cases. Case 1: x^* is a constant vector (Case 1 is the same situation as in Hordijk-Tijms [5]).

Case 2: for any $f \in F$ for which $P^*(f)(x(f) - x^*) = 0$, the Markov chain $P(f)$ does not have any nonrecurrent state.

It is seen from the following consideration that Assumption H is satisfied under Case 2: take any $f \in F$ for which $P^*(f)(x(f) - x^*) = 0$. Under Case 2 then it follows from the finiteness of state space that all states are positive recurrent, i.e. $p_{ii}^*(f) > 0$ for all $i \in S$. Since it is true that $x^* \geq x(f)$, $x^* - x(f) = 0$.

We denote Γ_n associated with $\hat{\mathcal{M}}$ by $\hat{\Gamma}_n$.

LEMMA 4.1. *Let Assumption H be satisfied. Then there exists a positive integer N such that for any $n \geq N$, every $f \in \hat{\Gamma}_n$ is optimal in $\mathcal{M}$ and moreover $P(f)x^* = x^*$.*

PROOF. In the manner similar to Theorem 4.3 we get $\hat{y}_{n_k} = \hat{r}(f) + \alpha_{n_k} P(f) \hat{y}_{n_k-1}$ for any $f \in \hat{\Gamma}_n$ and for any $n \geq N$. From Lemma 3.4 letting $k \rightarrow \infty$ we get $\hat{V}_\infty = \hat{r}(f) + P(f) \hat{V}_\infty$. By iterating n -times, we get

$$\hat{V}_\infty = \sum_{k=0}^n P(f)^k \hat{r}(f) + P(f)^{n+1} \hat{V}_\infty.$$

Dividing both sides by $n+1$ and letting $n \rightarrow \infty$, we get $0 = P^*(f) \hat{r}(f)$. Hence we have $P^*(f)(x(f) - x^*) = 0$. Then Assumption H completes the proof.

A policy π is a sequence $(\dots, f_n, \dots, f_2, f_1)$ of members of F .

THEOREM 4.4. *Under Assumption H, it holds that*

$$(4.6) \quad \lim_{n \rightarrow \infty} \{y_n - \gamma_n x^*\} = \hat{V}_\infty.$$

PROOF. Let $\tau = (\dots, g_n, \dots, g_2, g_1)$ be a policy such that $g_n \in \hat{\Gamma}_n$ for $n = 1, 2, \dots$. Take N which is assured in Lemma 4.1. Then we get for $n \geq N$,

$$\begin{aligned} \hat{y}_n &= \hat{r}(g_n) + \sum_{k=1}^n \alpha_n \alpha_{n-1} \dots \alpha_k P(g_n g_{n-1} \dots g_k) \hat{r}(g_{k-1}) \\ &= \left[r(g_n) + \sum_{k=1}^n \alpha_n \alpha_{n-1} \dots \alpha_k P(g_n g_{n-1} \dots g_k) r(g_{k-1}) \right] \\ &\quad - \left[\sum_{k=2}^N \alpha_n \alpha_{n-1} \dots \alpha_k P(g_n g_{n-1} \dots g_k) x^* + (x^* + \sum_{k=N+1}^n \alpha_n \alpha_{n-1} \dots \alpha_k P(g_n g_{n-1} \dots g_k) x^*) \right]. \end{aligned}$$

As for the undefined g_0 in the above formula, we interpret as $r(g_0) = \hat{r}(g_0) = y_0$.

It is clear that the quantity between the first two brackets of the last side is never greater than y_n . Let $M = \min \{x^*(i); i \in S\}$ where $x^*(i)$ is the i -th element of x^* , and let e be the S -dimensional column vector whose elements are all equal to 1. Then we have

$$\begin{aligned} \sum_{k=2}^N \alpha_n \alpha_{n-1} \cdots \alpha_k P(g_n g_{n-1} \cdots g_k) x^* &\geq \sum_{k=2}^N \alpha_n \alpha_{n-1} \cdots \alpha_k M e \\ &= \alpha_n \alpha_{n-1} \cdots \alpha_N \gamma_{N-1} M e. \end{aligned}$$

On the other hand from Lemma 4.1 we have $P(g_k)x^* = x^*$ for all $k \geq N$. Hence we have

$$\begin{aligned} x^* + \sum_{k=N+1}^n \alpha_n \alpha_{n-1} \cdots \alpha_k P(g_n g_{n-1} \cdots g_k) x^* &= x^* + \sum_{k=N+1}^n \alpha_n \alpha_{n-1} \cdots \alpha_k x^* \\ &= (\gamma_n - \alpha_n \alpha_{n-1} \cdots \alpha_N \gamma_{N-1}) x^*. \end{aligned}$$

From the results stated above we get

$$\hat{y}_n \leq y_n - \gamma_n x^* + \alpha_n \alpha_{n-1} \cdots \alpha_N \gamma_{N-1} (x^* - M e).$$

From the condition (iii) for $\{\alpha_n\}$ and Lemma 3.4 it follows that

$$\hat{V}_\infty = \lim_{n \rightarrow \infty} \hat{y}_n \leq \varliminf_{n \rightarrow \infty} \{y_n - \gamma_n x^*\}.$$

On the other hand from Lemma 3.1 we get

$$\lim_{n \rightarrow \infty} \hat{y}_n \geq \overline{\lim}_{n \rightarrow \infty} \{y_n - \gamma_n x^*\}.$$

Combining these two inequalities, we get (4.6). This completes the proof.

Another version of Theorem 4.4 in terms of $\mathcal{M}$ is as follows.

COROLLARY 4.1. *Let the stationary strategy f be α -discount optimal (in $\mathcal{M}$) for all α close enough to 1. Then it holds that*

$$\lim_{n \rightarrow \infty} \{y_n - \gamma_n x^*\} = H(f)r(f).$$

PROOF. Take such f as stated in the corollary. From the equalities $H(f)P^*(f) = 0$ and $x^* = x(f) = P^*(f)r(f)$, and from the Laurent series expansion (3.5) of $\hat{V}_\alpha(f^\infty)$, we get

$$(4.7) \quad \hat{V}_\alpha(f^\infty) = (1 + \rho) \sum_{k=0}^{\infty} (-\rho)^k H(f)^{k+1} r(f)$$

for α close enough to 1. The results in [7] show that f is also α -discount optimal in $\hat{\mathcal{M}}$ for all α close enough to 1. Since ρ tends to 0 as α increases to 1, the corollary is obvious from Theorem 4.4 and (4.7).

COROLLARY 4.2. *It holds that*

$$\lim_{n \rightarrow \infty} \{y_n - \alpha_n y_{n-1}\} = x^*.$$

PROOF. From $\gamma_n = 1 + \alpha_n \gamma_{n-1}$, it follows that

$$y_n - \alpha_n y_{n-1} = (y_n - \gamma_n x^*) - \alpha_n (y_{n-1} - \gamma_{n-1} x^*) + x^*.$$

The right-hand side tends to x^* by virtue of Theorem 4.4 and the condition (ii).

COROLLARY 4.3. *There exists a positive integer N such that for any $n \geq N$ every $f \in \Gamma_n$ is optimal in $\mathcal{M}$.*

PROOF. The proof can be obtained along the line similar to Theorem 3 (b) in [5] by the use of Theorems 4.3 and 4.4, so that is omitted.

Acknowledgement

The author would like to express his sincere gratitude to Professor N. Furukawa whose guidance and encouragements helped make this paper possible. The author also would like to express his deep appreciation to Dr. S. Iwamoto for his useful advices in preparing this paper.

References

- [1] D. BLACKWELL, *Discrete dynamic programming*, Ann. Math. Statist. **33** (1962), 719–726.
- [2] B. W. BROWN, *On the iterative method of dynamic programming on a finite space discrete time Markov process*, Ann. Math. Statist. **36** (1965), 1279–1285.
- [3] K. L. CHUNG, *Markov Chains with Stationary Transition Probabilities*, Berlin, Springer-Verlag (1960).
- [4] G. H. HARDY, *Divergent Series*, Oxford Univ. Press (1949).
- [5] A. HORDIJK and H. TIJMS, *A modified form of the iterative method of dynamic programming*, Ann. Statist. **3** (1975), 203–208.
- [6] R. A. HOWARD, *Dynamic Programming and Markov Processes*, Wiley, New York (1960).

- [7] B. L. MILLER and A. F. VEINOTT, JR., *Discrete dynamic programming with a small interest rate*, Ann. Math. Statist. **40** (1969), 366–370.