A NOTE ON DISCRETE MARKOVIAN DECISION PROCESS

Ogawara, Masami
Tokyo Woman's Christian College

https://doi.org/10.5109/13012

出版情報：統計数理研究. 11 (1/2), pp.35-42, 1964-03. Research Association of Statistical Sciences

A NOTE ON DISCRETE MARKOVIAN DECISION PROCESSES

By

Masami OGAWARA

(*Received* November 15, 1963)

Introduction. In the current method of policy improvement for Markovian decision processes, it seems to be assumed that the optimal policy is composed of successions of the same decision. (Howard [2], Blackwell [3]) When the discount factor β is less than one, this fact is easily proved by use of Blackwell's theorems (§2). In the case of $\beta=1$, as we shall show in §3, similar theorems hold but the situation is not affirmative. In the last paragraph (§4) we shall do some generalizations of discrete Markovian decision processes.

§ **1. Definitions and Notations.** In the following, symbol $\equiv$ means definition or identity. We consider a system such that the states of which are labeled by the integers and form a finite set $S\equiv\{1, 2, \cdots, N\}$. We assume that the set A of our possible actions with the system is also finite; $A\equiv\{1, 2, \cdots, M\}$. In this note however these assumptions of finiteness are not essential.

Decision function $d=d(i)$ is a mapping from $i \in S$ onto $d \in A$, and the set D of decision functions is finite; the number of its elements is M^N.

Now, we observe the system at intervals of unit time and make a decision on our action according to the state. In that case we assume that the system moves with the transition probability matrix

$$P(d)\equiv(p_{ij}(d))$$

which depends on the decision function d and that we receive an immediate reward $r_{ij}(a)$ for one-step transition $i\rightarrow j$ and an action $a \in A$, where $r_{ij}(a)$ is a random variable. The expected immediate reward is then

$$r(i,\ d(i))\equiv\sum_j p_{ij}(d(i))r_{ij}(d(i))$$

and

$$r(d)\equiv\begin{pmatrix} r(1,\ d(1)) \\ \vdots \\ r(N,\ d(N)) \end{pmatrix} \quad (d \in D)$$

is said to be an expected immediate reward vector.

By a policy π, we mean a sequence of decision functions;

$$\pi \equiv (d_1 d_2 d_3 \cdots) \quad (d_n \in D,\ n=1, 2, \cdots)$$

According to Blackwell, we shall use the following notations for various special policies; for a policy $\pi \equiv (d_1\, d_2 \cdots)$

$$\begin{aligned}
&(d\pi) \equiv (d\, d_1\, d_2 \cdots) \\
&(g_1 \cdots g_n \pi) \equiv (g_1 \cdots g_n\, d_1\, d_2 \cdots) \\
&(d^{(n)}\pi) \equiv (\underbrace{d\, d \cdots d}_{n}\ d_1\, d_2 \cdots) \\
&d^{(\infty)} \equiv (dd \cdots) \\
&(T\pi) \equiv (d_2\, d_3 \cdots)
\end{aligned}$$

and so forth. We denote the set of all policies by Π.

The n-step transition probability matrix for a policy $\pi \equiv (d_1\, d_2 \cdots)$ is denoted by

$$P_n(\pi) \equiv P(d_1) P(d_2) \cdots P(d_n) \quad (n=0, 1, 2, \cdots),$$

where $\quad P_0(\pi) \equiv I.$

Finally, if the present value of unit income n-steps in future is β^n, the β is called a discount factor; if the unit time interest rate is ρ, $\beta=(1+\rho)^{-1}$ and $0<\beta \leqq 1$.

§ 2. **Optimal Policies for $\beta<1$.** If $\beta<1$ and $r(d)$ has finite components on D, the expected total rewards vector corresponding to a policy $\pi \equiv (d_1\, d_2 \cdots)$ is given by the convergent infinite series

$$\begin{aligned}
V(\pi) &= \sum_{n=0}^{\infty} \beta^n P_n(\pi) r(d_{n+1}) \\
&= r(d_1) + \beta\, P(d_1) V(T\pi)\,;
\end{aligned} \tag{2.1}$$

the ith component of $N \times 1$ column vector $V(\pi)$ is the expectation of total rewards when the process started from a state i.

Now we define a semi-order in the N dimensional vector space $\mathfrak{v}$. For two $N \times 1$ column vectors

$$v_1 = \begin{pmatrix} v_{11} \\ \vdots \\ v_{1N} \end{pmatrix}, \quad v_2 = \begin{pmatrix} v_{21} \\ \vdots \\ v_{2N} \end{pmatrix}$$

if $v_{1i} \geqq v_{2i}$ $(i=1, 2, \cdots, N)$ we denote $v_1 \geqq v_2$ or $v_2 \leqq v_1$

and if $v_1 \geqq v_2$ and $v_1 \neq v_2$ we denote $v_1 > v_2$ or $v_2 < v_1$.

The join of two vectors v_1 and v_2, $v_1 \vee v_2$, is defined by the vector whose ith component is max (v_{1i}, v_{2i}) and the meet $v_1 \wedge v_2$ is the vector whose ith component is min (v_{1i}, v_{2i}). Thus $\mathfrak{v}$ is a vector lattice and the absolute value of $v (v \in \mathfrak{v})$ is defined as a vector whose ith component is given by $|v_i|$. If we define, morerover, the norm of v by $\|v\| = \max_{1 \leq i \leq N} |v_i|$, $\mathfrak{v}$ is a Banach lattice. The subspace of $\mathfrak{v}$, $\mathfrak{v}(\Pi) = \{V(\pi) ; \pi \in \Pi\}$, which is induced by (2.1) from Π, is also a Banach lattice.

Corresponding to space $\mathfrak{v}$, we introduce the semi-order into the policy space Π ; for two policies π_1 and π_2, we write

$$\pi_1 \geqq \pi_2 \quad \text{if } V(\pi_1) \geqq V(\pi_2)$$
$$\pi_1 = \pi_2 \quad \text{if } V(\pi_1) = V(\pi_2)$$
$$\text{and} \quad \pi_1 > \pi_2 \quad \text{if } V(\pi_1) > V(\pi_2).$$

The policy π^* is said to be optimal if $\pi^* \geqq \pi$ for all $\pi \in \Pi$. If the subspace $\mathfrak{v}(\Pi)$ is complete, then such optimal policy π^* exists.

If there exist policies π', π'' such that

$$V(\pi') = V(\pi_1) \vee V(\pi_2), \quad V(\pi'') = V(\pi_1) \wedge V(\pi_2)$$

for any two policies π_1, π_2, then we may define

$$\pi' = \pi_1 \vee \pi_2, \quad \pi'' = \pi_1 \wedge \pi_2 ;$$

thus Π would be a lattice and $V(\Pi)$ a homomorphic mappimg from Π into $\mathfrak{v}(\Pi)$.

In the following, the convergence of vector sequence may be underst-ood in the sense of norm defined above.

Lemma 2.1. *For any two policies* $\pi \equiv (d_1 d_2 \cdots)$ and $\pi' \equiv (d'_1 d'_2 \cdots)$

$$\lim_{n \to \infty} V(d'_1 \cdots d'_n \pi) = V(\pi')$$

Proof. We have

$$V(d'_1 \cdots d'_n \pi) = \sum_{\nu=0}^{n-1} \beta^\nu P_\nu(\pi') r(d'_{\nu+1}) + \beta^n P_n(\pi') \sum_{\nu=0}^{\infty} \beta^\nu P_\nu(\pi) r(d_{\nu+1})$$

and there exists a constant vector R such that $\|r(d)\| < R < \infty$ $(d \in D)$. Therefore, the norm of the second term on the right hand side is less than $\beta^n R/(1-\beta)$ and $\beta^n \to 0 (n \to \infty)$. Thus the second term $\to 0 (n \to \infty)$.

Following to Blackwell we define 'monotone' operator $L(d)$ by

$$L(d)v \equiv r(d) + \beta P(d)v \quad (v \in \mathfrak{v},\ \mathrm{d} \in D).$$

Then, if $v_1 \geqq v_2$ or $v_1 > v_2$, $L(d)v_1 \geqq L(d)v_2$. The following two theorems are easily proved by lemma 2.1 and the monotoneity of operator $L(d)$. (Blackwell [3])

Theorem 2.1. *If* $\pi^* = (d\pi^*)$ *for all* $d \in D$, *then* π^* *is optimal.*

Theorem 2.2. *If* $(d\pi)>\pi$, *then* $d^{(\infty)}>\pi$.

Lemma 2.2. *If* $\pi_1\leqq\pi_2$, *then* $(d\pi_1)\leqq(d\pi_2)$ *for all* $d\in D$.

Proof. By the assumption $V(\pi_1)\leqq V(\pi_2)$. Hence $L(d)V(\pi_1)\leqq L(d)V(\pi_2)$ or $V(d\pi_1)\leqq V(d\pi_2)$. Thus we get $(d\pi_1)\leqq(d\pi_2)$.

Theorem 2.3. *If* $\pi\equiv(d_1\, d_2\cdots)$ *is optimal, then* $\pi=d_1^{(\infty)}$.

Proof. Since $\pi\geqq(d_2\, d_3\cdots)$, we get by theorem 2.2

$$d_1^{(\infty)}\geqq(d_2\, d_3\cdots)$$

and by lemma 2.2

$$(d_1\, d_1^{(\infty)})\geqq(d_1\, d_2\cdots) \text{ or } d_1^{(\infty)}\geqq\pi,$$

whereas π is optimal, $\pi=d_1^{(\infty)}$.

Owing to this theorem we may search an optimal policy in the confined set of policies of the type $d^{(\infty)}(d\in D)$ and the theorems 2.1 and 2.2 give the basis of policy improvement routine for $\beta<1$, by seting $\pi\equiv d_1^{(\infty)}(d_1\in D)$.

§ 3. Optimal Policies for $\beta=1$.

The expected total reward for n-1 transitions under a policy $\pi\equiv(d_1\, d_2\cdots)$ is given by

$$V_n(\pi)\equiv\sum_{\nu=0}^{n-1}P_\nu(\pi)r(d_{\nu+1})$$

and the expected mean reward per step for n-1 transitions is

$$\bar{V}_n(\pi)\equiv V_n(\pi)/n. \tag{3.1}$$

In general, (3.1) does not converge as $n\to\infty$, and so refering to the min-max doctrine let us define the value of policy π by

$$\bar{V}(\pi)\equiv\liminf_{n\to\infty}\bar{V}_n(\pi),$$

where the right hand side means the vector each component of which is the inferior limit of the corresponding component of $\bar{V}_n(\pi)$.

Similarly to the foregoing paragraph, we set the correspondence between the semi-order in $\mathfrak{v}(\Pi)$ (space of $\bar{V}(\pi)$) and that in Π as follows,

$$\bar{V}(\pi_1)\leqq\bar{V}(\pi_2) \rightleftarrows \pi_1\leqq\pi_2$$
$$\bar{V}(\pi_1)<\bar{V}(\pi_2) \rightleftarrows \pi_1<\pi_2$$
$$\bar{V}(\pi_1)=\bar{V}(\pi_2) \rightleftarrows \pi_1=\pi_2$$

and π^* is an optimal policy, if $\pi^*\geqq\pi$ for all $\pi\in\Pi$. We may point out here only that the lattice theoretical interpretation to the present case can be given quite similarly to the foregoing case of $\beta<1$.

Now, for two policies

$$\pi_1=(d_1^1 d_2^1\cdots) \text{ and } \pi_2=(d_1^2 d_2^2\cdots)$$

we set

$$V_{m,n}(\pi_1, \pi_2)\equiv\sum_{\nu=0}^{m-1} P_\nu(\pi_1)r(d_{\nu+1}^1)+P_m(\pi_1)\sum_{\nu=0}^{n-1} P_\nu(\pi_2)r(d_{\nu+1}^2)$$
$$=m\overline{V}_m(\pi_1)+nP_m(\pi_1)\overline{V}_n(\pi_2)$$
$$\overline{V}_{m,n}(\pi_1, \pi_2)\equiv V_{m,n}(\pi_1, \pi_2)/(m+n).$$

Then we get the following

Lemma 3.1. $\liminf_{m\to\infty} \overline{V}_{m,n}(\pi_1, \pi_2)=\overline{V}(\pi_1)$

$$\liminf_{n\to\infty} \overline{V}_{m,n}(\pi_1, \pi_2)=P_m(\pi_1)\overline{V}(\pi_2)\equiv\overline{V}(d_1^1\cdots d_m^1\pi)$$

Theorem 3.1. *If there exists an integer n_1 such that*

$$\overline{V}_{1,n}(d, \pi)\geqq\overline{V}_{n+1}(\pi) \text{ for all } n\geqq n_1$$

then $d^{(\infty)}\geqq\pi$.

Proof. We define a monotone operator $L(d)$ associated with each $d\in D$ by

$$L(d)v=r(d)+P(d)v \quad (v\in\mathfrak{v})$$

Then, from our assumption, we get

$$L(d)V_n(\pi)\geqq V_{n+1}(\pi) \qquad n\geqq n_1$$

Consequently $\quad L^m(d)V_n(\pi)\geqq L^{m-1}(d)V_{n+1}(\pi)\geqq\cdots\geqq V_{n+m}(\pi).$

That is $\quad \overline{V}_{m,n}(\overset{(\infty)}{d}, \pi)\geqq\overline{V}_{m+n}(\pi)$

Hence $\quad \liminf_{m\to\infty} \overline{V}_{m,n}(d^{(\infty)}, \pi)\geqq\liminf_{m\to\infty} \overline{V}_{m+n}(\pi)$

Therefore $\quad \overline{V}(d^{(\infty)})\geqq\overline{V}(\pi)$

Thus we get $\quad d_{(\infty)}\geqq\pi.$

Theorem 3.2. *Assume that $\lim_{n\to\infty} \overline{V}_n(\pi)=\overline{V}(\pi)$ exists for a policy π. If $(d\pi)>\pi$, then $d^{(\infty)}>\pi$, and if $(d\pi)<\pi$, then $d^{(\infty)}<\pi$.*

Proof. By the assumption there exist two vectors α, β such that

$$L(d)V_n(\pi)\geqq(n+1)\alpha>(n+1)\beta\geqq V_{n+1}(\pi) \text{ for } n \text{ large.}$$

Since the operator $L(d)$ is monotone,

$$L^2(d)V_n(\pi)\geqq L(d)(n+1)\alpha\geqq L(d)(n+1)\beta\geqq L(d)V_{n+1}(\pi)$$
$$\geqq(n+2)\alpha>(n+2)\beta\geqq V_{n+2}(\pi)$$

By repeating the operation, we get

$$L^m(d)V_n(\pi) \geqq (n+m)\alpha > (n+m)\beta \geqq V_{n+m}(\pi).$$

Deviding each term by $(n+m)$ and letting $m\to\infty$, we have

$$\bar{V}(d^{(\infty)}) \geqq \alpha > \beta \geqq \bar{V}(\pi)$$

Hence $$d^{(\infty)} > \pi.$$

The second part of the theorem is proved in the same way.

Now, the following fact is well known in the theory of Markov chain.

Lemma 3.2. *For any $d \in D$*

$$\lim_{n\to\infty}(1+P(d)+P(d)^2+\cdots+P(d)^n)/(n+1) = P_d$$

exists and $P_dP(d) = P(d)P_d = P_d^2 = P_d$.

From this lemma we get

Theorem 3.3. *If $\pi = d^{(\infty)}$, then $(d^{(n)}\pi) = \pi$ for $n=1, 2, \cdots$.*

Proof. By assumption $\bar{V}(\pi) = \bar{V}(d^{(\infty)})$. On the other hand, by lemma 3.2, $\bar{V}(d^{(\infty)}) = P_d r(d)$ and $P(d)^n \bar{V}(d^{(\infty)}) = P_d r(d) = \bar{V}(d^{(\infty)})$. Therefore, $P(d)^n \bar{V}(\pi) = P(d)^n \bar{V}(d^{(\infty)}) = \bar{V}(d^{(\infty)}) = \bar{V}(\pi)$. Consequently, by lemma 3.1, we get $(d^{(n)}\pi) = \pi$.

Finally we observe

Theorem 3.4. *If $\pi \equiv (d_1 d_2 \cdots)$ is optimal, then*

i) $\pi = (d_1^{(n)}\pi)$ $n=1, 2, \cdots$

ii) *π and $d_1^{(\infty)}$ belong to the same class $\Pi^* \equiv \{\pi;\ P(d_1)\bar{V}(\pi) = \bar{V}(\pi)\}$.*

Proof. First, we see that lemma 2.2 holds in case of $\beta=1$ too; for any two policies π_1 and π_2

(3.2) if $\pi_1 \leqq \pi_2$, then $(d\pi_1) \leqq (d\pi_2)$ for all $d \in D$.

In fact, since $P(d)$ is also a monotone operator, from $\bar{V}(\pi_1) \leqq \bar{V}(\pi_2)$ we get $P(d)V(\pi_1) \leqq P(d)V(\pi_2)$, that is $\bar{V}(d\pi_1) \leqq \bar{V}(d\pi_2)$ which leads to (3.2).

Now, because of optimality of π, we have $(d_1 d_2 \cdots) \geqq (d_2 d_3 \cdots)$ and by (3.2) $(d_1 d_1 d_2 \cdots) \geqq (d_1 d_2 \cdots)$ or $(d_1\pi) \geqq \pi$, while $(d_1\pi) \leqq \pi$. Hence $(d_1\pi) = \pi$, that is $P(d_1)\bar{V}(\pi) = \bar{V}(\pi)$. Operating $P(d_1)$ $n-1$ times on both sides, we get $P(d_1)^n\bar{V}(\pi) = \bar{V}(\pi)$ which proves (i). On the other hand, from $\bar{V}(d_1^{(\infty)}) = P_d r(d_1)$ we get $P(d_1)\bar{V}(d_1^{(\infty)}) = P_{d_1} r(d_1) = \bar{V}(d_1^{(\infty)})$.

Thus, together with the equation $P(d_1)\bar{V}(\pi) = \bar{V}(\pi)$ obtained above, part (ii) of the theorem was proved.

Corollary. *If $\pi \equiv (d_1 d_2 \cdots)$ is optimal,*

$$\max_{d\in D} P(d)\bar{V}(\pi) = P(d_1)\bar{V}(\pi) = \bar{V}(\pi).$$

Even if $\pi\equiv(d_1\, d_2\cdots)$ is optimal, the equivalence $\pi=d_1^{(\infty)}$ may not be concluded (except the special case of $r(d_1)=V(\pi)$) which was necessary consequence in case of $\beta<1$. However we may think of that the policy $d_1^{(\infty)}$ is 'nearly optimal' in the sense of theorem 3.4 and corollary to it.

By the way, in relation to our equation $P(d_1)V(\pi)=V(\pi)$, there is Bellman's theorem ([1], p. 329). However, since $(a_{ij}(q))$ in his theorem is a Markov matrix as well as our $P(d_1)$, the solution has the form $\alpha y_i+\beta$ $(i=1,2,\cdots, N)$ where α and β are arbitrary real constants.

§ 4. A general formulation.

We may generalize the state space to k-dimensional vector space R^k and the atcion space to m-dimensional vector space R^m; decision function $d(s)$ $(s\in R^k)$ is a mapping from R^k into R^m, and D is the set of all decision functions, while immediate reward $r(s_1,\ s_2)$ corresponding to a transition of state $s_1\rightarrow s_2$ is a real valued random variable.

Suppose that states $s=s(t)$ are observed at equally spaced discrete time points $t=0,1,2,\cdots$ and an action $d(s)$ is decided each time according to the state and let the transition probability distribution function associated with the decision function d be

$$F(s_0,\ s\,;\ d)=Pr(s(n)\leqq s\mid s(n-1)=s_0\,;\ d(s_0)) \qquad (n=1,2,\cdots)$$

that is independent of time n, where s_0 and s are k-dimensional vectors and $\leqq$ means the semi-order defined similarly to that in §2.

Policy is a sequence of decision functions; $\pi\equiv(d_1\, d_2\cdots)$, a point of policy space Π. The n step transition probability distribution function for a policy $\pi\equiv(d_0\, d_1\, d_2\cdots)$ is given by

$$F_n(s_0, s\,;\pi)=\int dF(s_0, s_1\,; d_0)dF(s_1, s_2\,; d_1)\cdots dF(s_{n-2}, s_{n-1}\,;\ d_{n-2})F(s_{n-1}, s\,; d_{n-1}) \qquad n=1,2,\cdots$$

where $F_1(s_0,\ s\,;\ \pi)=F(s_0,\ s\,;\ d_0)$.

The expected immediate reward started with state s_1 and decision $d(s_1)$ is

$$r(s_1,\ d(s_1))=\int r(s_1,\ s_2)dF(s_1,\ s_2\,;\ d)$$

and the expected total reward for a policy $\pi\equiv(d_0\, d_1\, d_2\cdots)$ with initial state s_0 is given by

$$V(\pi\,;\ s_0)=\sum_{n=0}^{\infty}\beta^n\int r(s\,;\ d_n(s))dF_n(s_0,\ s\,;\ \pi)$$

where β is a discount factor and $0\leqq\beta<1$.

Now, into the space $\mathfrak{v}$ of real valued functions of k real variables we can introduce semi-order and norm in a quite similar way as before. If the subspace $\mathfrak{v}(\Pi)=\{V(\pi)\ ;\ \pi\in\Pi\}$ of $\mathfrak{v}$ is a complete vector lattice and if

$$\sup_{\pi\in\Pi} V(\pi)=V(\pi^*),\quad \pi^*\in\Pi$$

then π^* is an optimal policy, and the theorems 2.1～2.3 hold in the same form.

In case of $\beta=1$, we may consider

$$\overline{V}_n(\pi,\ s_0)=(n+1)^{-1}\sum_{\nu=0}^{\infty}\int r(s,\ d_\nu(s))dF_\nu(s_0,\ s\ ;\ \pi)$$

$$\overline{V}(\pi,\ s_0)=\liminf_{n\to\infty}\ \overline{V}_n(\pi\ ;\ s_0),\ (s_0\in R^k)$$

and we can proceed in almost same way as preceding paragraph. The dynamic form of maximum expected reward $V_n(\pi,\ s_0)$ for an optimal policy $\pi\equiv(d_1\,d_2\cdots)$ in n steps started in state s_0 may be given by

$$V_n(\pi\ ;\ s_0)=\max_{d_1\in D}\left[r(s_0,\ d_1(s_0))+\beta\int V_{n-1}(T\pi\ ;\ s)dF(s_0,\ s\ ;d_1(s_0))\right],\ 0\leqq\beta\leqq1.$$

Acknowledgement. This note is one of the studies developed in the DP Developing Group sponsored by the Japanese Union of Scientists and Engineers. The author would like to express his thanks to Prof. T. Kitagawa, chief of the group, and the other members for the discussions given by them.

TOKYO WOMAN'S CHRISTIAN COLLEGE

References

[1] R. BELLMAN: *Dynamic Programming*, Princeton (1957).
[2] R.. A. HOWARD: *Dynamic Programming and Markov Processes*, MIT and Wiley (1960).
[3] D. BLACKWELL: *Discrete Dynamic Programming*, Ann. Math. Statist., Vol. **33** (1962), 719-726.