POLICY IMPROVEMENT IN MARKOV DECISION PROCESSES AND MARKOV POTENTIAL THEORY

Yasuda, Masami
Statistical Laboratory, College of General Education, Chiba University

https://doi.org/10.5109/13123

出版情報：統計数理研究. 18 (1/2), pp.55-67, 1978-03. Research Association of Statistical Sciences
バージョン：
権利関係：

POLICY IMPROVEMENT IN MARKOV DECISION PROCESSES AND MARKOV POTENTIAL THEORY

By

Masami YASUDA*

(Received October 5, 1977)

1. Introduction

A connection between Markov Decision Process (MDP) and Markov potential theory has two sides. One is the potential theoritic development of MDP and the other is the alternative proof of the results in MDP owing to Markov potential theory. Shaufele [12] belongs to the later, but it seems interesting from the standpoint of the mathematical programming to establish the development of MDP by using certain potential notion. Several approaches have been tried. Watanabe [16] interpreted the monotonicity of Howard's iteration [8] in the relation to the a dual problem of Linear Programming. By the property of a potential kernel, Furukawa [6] and Aso and Kimura [1] proved a policy improvement. A formulation of MDP by potential theoretic notion has been tried by Hordijk [7].

In many cases it is restricted to a transient potential theory because its analysis is simpler. In this paper we shall define a new potential in order to serve a general policy improvement. Our aim is to expose theorems which are available to several cases of MDP.

By the potential theoretic terms, we can interpret policy improvements of MDP as follows; The increase of rewards in MDP consists of the potential with a charge of an increment of the policy improvement and a regular function. If it is transient, then the potential is reduced to the ordinary one and the regular function equals zero. Hence this consists with that of Watanabe [16]. The merit of the potential is that it connects the policy improvement with the increment of rewards.

We shall consider the following cost criteria of MDP; (1) discounted case, (2) average case, (3) nearly optimal case and (4) sensitive discounted case. Case (1) and (2) are representitive and discussed by many authors. Especially we list up Howard [7] and Blackwell [2], [3] for (1) and Howard [8] and Derman [4], [5] for (2). Case (3) is due to Blackwell [2]. Extending case (3), case (4) is studied by Miller and Veinott [11] and Veinott [14], [15].

* Statistical Laboratory, College of General Education, Chiba University, Chiba 280, Japan.

2. Potential theory

Let S be a denumerable set and $P=\{P_{ij}; i, j\in S\}$ be a Markov matrix. A sequence of powers of P has a Cesaro sum;

$$P^*=\lim_{k\to\infty}\frac{1}{k}\{I+P+\cdots+P^{k-1}\} \tag{2.1}$$

where I is a unit matirix and lim means a pointwise convergence. A function f on S is a colum vector. If

$$\lim_k k^{-1}\sum_{l=0}^{k-1}P^l f$$

exist and finite, then we say $\{P^k f\}$ is Cesaro summable to P^*f. It is known that

$$0\leqq P^*=P^*P=PP^*=P^*P^*,\quad P^*\mathbf{1}\leqq\mathbf{1} \tag{2.2}$$

where $\mathbf{1}$ is a column vector whose elements equal 1 and the inequality between matrices or vectors means that it holds for each element. A function f on S such that $Pf=f$ is called *regular* for P. See Kemeny, Snell and Knapp [9] for terminology of Markov chains.

DEFINITION 1. Let

$$H_n=\sum_{k=0}^{n-1}(P^k-P^*),\quad n\geqq 1 \tag{2.3}$$

and a function f on S for which H_nf are well defined and finite. If the sequence $\{H_nf\}$ is Cesaro summable to a function g, then f is called a *charge* with respect to the Markov matrix P. We denote the Cesaro sum by Hf. It is called a called a *potential* with a charge f. That is,

$$g=Hf=\lim_n\frac{1}{n}\sum_{k=1}^{n}H_k f. \tag{2.4}$$

We note that a pointwise limit H does not necessarily exist in denumerable states. If P is a non-cyclic strong ergodic chain, $\lim H_n$ exists and finite (Kemeny, Snell and Knapp [9]). If S is a finite state, then $I-P+P^*$ is non-singular and

$$H=[I-P+P^*]^{-1}-P^*. \tag{2.5}$$

If S is a single transient chain with a denumerable state, then P^* is zero matrix and

$$H=[I-P]^{-1}. \tag{2.6}$$

Hence this kernel consists with a ordinary one and equals

$$\lim_n\sum_{k=0}^{n-1}P^k.$$

THEOREM 1. *If f and g are functions for which*

$$f=(I-P)g \tag{2.7}$$

and P^*g is finite-valued, then Hf is finite-valued and

$$g=Hf+h \tag{2.8}$$

where h is regular and $h=P^*g$.

PROOF. From the definition (2.3) of H_n, (2.7) implies

$$H_nf=g-P^{n+1}g, \quad n\geqq 1. \tag{2.9}$$

Since P^*g is finite-valued, the sequence of $\{P^ng\}$ is Cesaro summable. Hence $\{H_nf\}$ is Cesaro summable and Hf is finite valued. Thus

$$Hf=g-P^*g. \tag{2.10}$$

Set $h=P^*g$. Then the function h is regular by (2.2). This proves the theorem.

From this theorem, the *Riesz decomposition*—a super regular function is uniquely decomposed into a non-negative charge and a regular function—is obtained but it is not refered later.

THEOREM 2. (a) *If a function f is non negative charge with $P^*f=0$, then the potential is non negative;*

$$Hf\geqq 0. \tag{2.11}$$

(b) *If f is regular, then*

$$Hf=0. \tag{2.12}$$

PROOF. (a) Since $P^*f=0$,

$$H_nf=\sum_{k=0}^{n-1}P^kf.$$

Hence $H_nf\geqq 0$ for all $n\geqq 1$.

(b) $H_nf=0$, $n\geqq 1$ are obtained by $P^nf=P^*f=f$. So (2.12) is immediate.

A potential of a non negative charge is called a *pure* potential. So, by (2.11), a pure potential with $P^*f=0$ is non negative. In a transient chain, it holds $P^*f=0$. Hence the pure potential is always non negative. But in our case the pure potential is not always non negative.

3. Theorems

In this section we will expose fundamental theorems which show how are policy improvements of MDP connected with the increment of rewards. The potential which is defined at the previous section links the two into closer relation with each other.

DEFINITION 2. For each $n\geqq 1$, $r_n(P)$ is a given function on S and is defined for a Markov matrix P.

(a) Let sequences of functions $\{w_n=w_n(P)\,;\,n\geqq 0\}$ and $\{u_n=u_n(\tilde{P})\,;\,n\geqq 0\}$, for Markov matrices P and $\tilde{P}$ respectively, satisfy the following iteration:

$$w_0=u_0=0 \quad \text{and for} \quad n\geqq 0$$

$$(I-P)w_{n+1}+Pw_n=r_{n+1}(P), \tag{3.1}$$

$$(I-\tilde{P})u_{n+1}+\tilde{P}u_n=r_{n+1}(\tilde{P}). \tag{3.2}$$

(b) Let functions f_n and g_n for $n\geqq 0$ be

$$f_n=r_n(P)-(I-P)u_n-Pu_{n-1}, \tag{3.3}$$

$$g_n=w_n-u_n \tag{3.4}$$

for which $\{u_n\}$ and $\{w_n\}$ satisfy (3.1) and (3.2) respectively.

It will become clear the meaning of these functions in the next section, so we take a slight look here. For a sequence $\{w_n\}$ or $\{u_n\}$, it represents reward of MDP corresponding a stationary policy. One of the functions among the sequence is our objective reward and others are complemental. The index n is not a time parameter but denotes like an order of the reward underlying MDP. The function g_n is an increment of rewards. It is comparing two stationary policies. The function f_n is an increment of a policy improvement for fixed rewards $\{u_n\}$. The form of f_n is obtained from a perturbation of a policy. Or we associate it with the recursive property of Dynamic Programming and also with Linear Programming.

Our aim is to maximize the reward in MDP and select a policy which reward is greater than others, that is, an optimal policy. Thus the policy improvement among stationary policies is to select a policy so that g_n in (3.4) is positive for a fixed policy P. The policy improvement contemplates seeking a routine which implies the positiveness of g_n. Routins for cases of MDP (Howard [8], Derman [4], [5], Blackwell [2], Miller and Veinott [11], Veinott [14], [15], Furukawa [6] and Aso and Kimura [1]) are summarized certain positivity of $\{f_n\}$. We postpone to argue the detail until the next section.

LEMMA 1. (a) *Let* $\{w_n\}$ *satisfy* (3.1). *For any sequence* $\{u_n\}$ (*not necessary to satisfy* (3.2)), $\{f_n\}$ *and* $\{g_n\}$ *defined by* (3.3) *and* (3.4) *have the following relation*;

$$f_n=(I-P)g_n+Pg_{n-1}, \quad n\geqq 1, \quad g_0=0. \tag{3.5}$$

(b) *If* P^*g_n *is finite-valued for* $n(\geqq 1)$, *then so is* $H(f_n-Pg_{n-1})$ *and*

$$g_n=H(f_n-Pg_{n-1})+P^*g_n. \tag{3.6}$$

Further if P^*g_{n+1} *is finite-valued, then so is* P^*f_{n+1} *and*

$$g_n=H(f_n-Pg_{n-1})+P^*f_{n+1}, \tag{3.7}$$

$$P^*g_{n-1}=P^*f_n. \tag{3.8}$$

PROOF. For (a), it is sufficient to eliminate $\{u_n\}$ from Definition 2(a). Since $\{w_n\}$ satisfies (3.1), we have (3.5). It is trivial for $n=0$. For (b), let $f=f_n-Pg_{n-1}$ and $g=g_n$. Applying Theorem 1, we can prove (3.6). The later parts follows from (3.5). This completes the proof.

Suppose that g_{n-1} is regular with respect to P for $n\geqq 1$. Since Pg_{n-1} is also regular, Theorem 2 (b) implies $H(Pg_{n-1})=0$. Hence we obtain

$$g_n=Hf_n+P^*f_{n+1} \tag{3.9}$$

from (3.7). Thus we insist on (3.9) that the increment of reward g_n consists of the potential Hf_n with the charge f_n, the increment of a policy improvement, and the regular function P^*f_{n+1}. If P is transient, then $P^*=0$ and so

$$g_n=Hf_n \tag{3.10}$$

where $H=[I-P]^{-1}$. Therefore, g_n is the potential with the charge of the increment in a policy improvement (Watanabe [14]). In this case since the pure potential is non-negative so $f_n\geqq 0$ implies $g_n\geqq 0$. This is the policy improvement of Howard [8] in the discounted case (1). For a general chain, the positivity of f_{n-1} and f_n among $\{f_n\}$ determines that of g_{n-2} and g_{n-1}. This is a principle for average case (2) and sensitive discounted case (4), which will be discussed later in detail. For nearly optimal case (3), it require further conditions so as to deduce the positivity of g_n.

ASSUMPTION 1. For a Markov matrix P, the state space S consists of

$$S=\bigcup_{\nu=1}^{\infty} R_\nu \cup T \tag{3.11}$$

such that R_ν is a finite recurrent class for each ν and T is a transient one.

This assumption holds for a finite state space. It may be replaced that S contains only strong ergodic classes and a transient one. Therefore, a point-wise limit of $\{H_n\}$ exists in the Ceraso summability, so H_{ij}, i, $j\in S$ are defined and

$$[Hf]_i=\sum_{j\in S} H_{ij}f(j)$$

for a bounded function f.

For functions $f, g, \cdots, h$ are column vectors, so $(f, g, \cdots, h)$ is a matrix. We say that it is *lexicographically* non-negative if the first nonvanishing element of each row of the matrix is positive, written by $(f, g, \cdots, h)\geqq_l 0$. From this definition, if $(f, g, \cdots, h)\geqq_l 0$, then $f\geqq 0$, that is, $f(i)\geqq 0$ for each $i\in S$.

Let $\{f_n\}$ and $\{g_n\}$ are defined by (3.3) for given Markov matrix P with Assumption 1 and sequence $\{r_n\}$.

THEOREM 3. *Suppose that* $\{w_n\}$ *satisfies* (3.1) *and* P^*g_k, $k=1, 2, \cdots, n-1$ *are finite-valued for* $n\geqq 4$. *If*

$$(f_1, f_2, \cdots, f_{n-2})=0, \tag{3.12}$$

$$(f_{n-1}, f_n)\geqq_l 0, \tag{3.13}$$

then

$$(g_1, g_2, \cdots, g_{n-3}) = 0, \tag{3.14}$$

$$(g_{n-2}, g_{n-1}) \geqq_l 0. \tag{3.15}$$

PROOF. Firstly we prove (3.14) from (3.12) by induction on n. For $n=4$, (3.12) implies $(f_1, f_2)=0$ and so $f_1=0$ and $f_2=0$. By (3.7), $g_1=Hf_1+P^*f_2$. Hence $g_1=0$. For $n\geqq 4$, it is sufficient to prove $g_{n-3}=0$ under (3.12). By (3.7),

$$g_{n-3}=H(f_{n-3}-Pg_{n-4})+P^*f_{n-2}.$$

Since $g_{n-4}=0$ by the assumption of induction and $f_{n-3}=f_{n-2}=0$ by (3.12), thus we obtain $g_{n-3}=0$. This proves (3.14). Nextly we determine the forms of g_{n-2} and g_{n-1} in (3.15). By (3.12) and (3.14),

$$g_{n-2}=P^*f_{n-1} \tag{3.16}$$

and so it is regular function by (2.2). Hence we have

$$g_{n-1}=Hf_{n-1}+P^*f_n \tag{3.17}$$

from Theorem 2 (b). In order to prove (3.15), we must show that

(a) $g_{n-2}\geqq 0$ and

(b) if $g_{n-2}(i)=0$ for some i, then $g_{n-1}(i)\geqq 0$.

The former (a) is immediate from (3.16) because $f_{n-1}\geqq 0$ by (3.13). To prove (b), we need the assumption (3.11). Suppose the state i belong to some recurrent class R_ν (abbr. by R). For $j\in R$, any P^*_{ij}, which is (i, j) element of P^*, are strict positive and so

$$g_{n-2}(i)=\sum_{j\in R}P^*_{ij}f_{n-1}(j)=0$$

implies $f_{n-1}(j)=0$, $j\in R$. Hence $f_n(j)\geqq 0$ for $j\in R$ by (3.13). Thus

$$g_{n-1}(i)=\sum_{j\in R}H_{ij}f_{n-1}(j)+\sum_{j\in R}P^*_{ij}f_n(j)$$

$$=\sum_{j\in R}P^*_{ij}f_n(j)\geqq 0.$$

Suppose that i belongs to a transient class T. From $f_{n-1}\geqq 0$ and

$$g_{n-2}(i)=\sum_{j\in S}P^*_{ij}f_{n-1}(j)=0$$

by (3.16), we have

$$\sum_{j\in S}H_{ij}f_{n-1}(j)\geqq 0 \tag{3.18}$$

applying Theorem 2 (a). Since $i\in T$, so $P^*_{ij}=0$ for $j\in T$ and $P^*_{ij}\geqq 0$ for $j\in R$. Therefore if a state j is such that $P^*_{ij}>0$, then $f_{n-1}(j)=0$ and so $f_n(j)\geqq 0$ by (3.13). Hence $P^*_{ij}f_n(j)\geqq 0$. If a state j is such that $P^*_{ij}=0$, cleary $P^*_{ij}f_n(j)=0$. Thus we have

$$\sum_{j\in S}P^*_{ij}f_n(j)\geqq 0. \tag{3.19}$$

For $i\in T$,

$$g_{n-1}(i)=\sum_{j\in S}H_{ij}f_{n-1}(j)+\sum_{j\in S}P^*_{ij}f_n(j)$$

Thus (3.18) and (3.19) imply $g_{n-1}(i)\geqq 0$ and this completes the proof.

COROLLARY 1. *If* $(f_1, f_2)\geqq_l 0$, *then* $g_1\geqq 0$.

PROOF. Since $P^*f_1=0$, $g_1=Hf_1+P^*f_2$ by (3.7) and (3.8). Hence we can prove it similarly.

COROLLARY 2. *Without Assumption* 1 *on a Markov matrix* P *but assume* P^*g_1 *and* P^*f_2 *are finite-valued. Then* $f_1\geqq 0$ *and* $f_2\geqq 0$ *imply* $g_1\geqq 0$.

PROOF. It is immediate from the relation which is in the proof of Corollary 1 and Theorem 2 (a).

THEOREM 4. *Added in Theorem* 3, *suppose that a state* i *such as* $f_{n-1}(i)=f_n(i)=0$ *satisfies* $P_{ij}=\tilde{P}_{ij}$ *for all* $j\in S$. *Then* $g_{n-2}=0$ *implies* $(g_{n-1}, g_n)\geqq_l 0$.

PROOF. Since $g_{n-2}=0$, we have

$$P^*f_{n-1}=0, \tag{3.20}$$

$$g_{n-1}=Hf_{n-1}+P^*f_n, \tag{3.21}$$

$$g_n=H[f_n-Pg_{n-1}]+P^*g_n \tag{3.22}$$

by (3.7) and (3.8). Further, $g_{n-1}\geqq 0$ by the result (3.15) in Theorem 3. Hence we will prove that $g_{n-1}(i)=0$ for a state i implies $g_n(i)\geqq 0$. Previously note that

$$f_{n-1}(j)=0,\quad j\in\bigcup_{\nu}R_\nu \tag{3.23}$$

by (3.30).

(a) Suppose that a state $i\in R_\nu$ for some ν (abbr. by R in the proof). For any $k\in R$,

$$g_{n-1}(k)=\sum_{j\in R}H_{kj}f_{n-1}(j)+\sum_{j\in R}P^*_{ij}f_n(j)$$

$$=\sum_{j\in R}P^*_{kj}f_n(j)$$

by (3.23). The assumption that $g_{n-1}(i)=0$ implies

$$f_n(j)=0,\quad j\in R \tag{3.24}$$

because $P^*_{ij}>0$, $j\in R$. Also

$$[Pg_{n-1}]_j=\sum_{k\in R}P_{jk}g_{n-1}(k)=\sum_{k\in R}P^*_{jk}f_n(k)$$

is regular with respect to $\{P_{ij}; i, j\in R\}$. Thus, by Theorem 2 (b),

$$g_n(i)=\sum_{j\in R}H_{ij}\{f_n(j)-[Pg_{n-1}]_j\}+\sum_{j\in R}P^*_{ij}g_n(j)$$

$$=\sum_{j\in R}P^*_{ij}g_n(j).$$

From (3.23) and (3.24), the assumption implies $P_{jk}=\tilde{P}_{jk}$ for all j, $k\in R$. Hence $u_n(j)=w_n(j)$ for all $j\in R$ because corresponding Markov matrix are same. Thus we have $g_n(j)=u_n(j)-w_n(j)=0$ for all $j\in R$.

(b) Suppose $i\in T$. By (3.23), (3.21) implies

$$g_{n-1}(i)=\sum_{j\in T}H_{ij}f_{n-1}(j)+\sum_{j\in\bar{R}}P^*_{ij}f_n(j)$$

where $\bar{R}=\cup_\nu R_\nu$. Let T_0, R_0 be subsets of T and $\bar{R}$ respectively such that $H_{ij}=0$, $j\in T_0$ and $P^*_{ij}=0$, $j\in R_0$. Noting $H_{ij}>0$ for i, $j\in T$, $g_{n-1}(i)=0$ implies

$$f_{n-1}(j)=0, \qquad j\in T\setminus T_0, \tag{3.25}$$

$$f_n(j)=0, \qquad j\in\bar{R}\setminus R_0. \tag{3.26}$$

From (3.23) and (3.25), we have $f_{n-1}(j)=0$ for $j\in\bar{R}\cup(T\setminus T_0)$. Hence $H_{kj}f_{n-1}(j)=0$ for each $k\in T$ and $j\in S$ because one of factors equals zero. Similarly $P^*_{kj}f_n(j)=0$ for each $k\in T$ and $j\in S$. Thus we obtained that

$$g_{n-1}=0. \tag{3.27}$$

From (3.27) and the result of (a), $g_n(j)=0$ for $j\in\bar{R}$. Hence, for $i\in T$,

$$[H[f_n-Pg_{n-1}]]_i=[Hf_n]_i=\sum_{j\in S}H_{ij}f_n(j)$$

$$=\sum_{j\in R_0}H_{ij}f_n(j)+\sum_{j\in T\setminus T_0}H_{ij}f_n(j)$$

and

$$[P^*g_n]_i=\sum_{j\in\bar{R}}P^*_{ij}g_n(j)=0$$

by (3.27) and (3.26). Thus

$$g_n(i)=\sum_{j\in R_0}H_{ij}f_n(j)+\sum_{j\in T\setminus T_0}H_{ij}f_n(j).$$

Applying Theorem 2 (a), the assertions

$$\sum_{j\in R_0}P^*_{ij}f_n(j)=0$$

and $f_n(j)\geqq 0$, $j\in R_0\cup(T\setminus T_0)$, which derived by (3.25) and (3.13), implies $g_n(i)\geqq 0$. This completes the proof.

THEOREM 5. *Suppose that a Markov matrix P is transient. If $(f_1, f_2, \cdots, f_{n-1})=0$ and $f_n\geqq 0$ for $n\geqq 2$, then $g_{n-1}=0$ and $g_n\geqq 0$.*

PROOF. This is clear from (3.10).

4. Policy improvements of MDP

In this section we will formulate MDP of (1)—(4) in section 1 and then expose a principle of policy improvements.

Let S a denumerable state space. A policy $\boldsymbol{P}$ is a sequence of Markov matrices, that is,

$$\boldsymbol{P}=(P_1, P_2, \cdots). \tag{4.1}$$

Π denotes the set of all policies. Let r_P is a column vector defined by Markov matrix P and

$$\boldsymbol{P}^{(n)}r=P_1P_2\cdots P_nr_{P_{n+1}}, \quad n\geqq 1, \tag{4.2}$$

$$\boldsymbol{P}^{(0)}r=r_{P_1}$$

where $\boldsymbol{P}=(P_1, P_2, \cdots)\in\Pi$. $\boldsymbol{P}^{(n)}r$ is the n-th expected reward. MDP is to maximize the total sum of the sequence of the n-th expected reward and several criteria are considered as follows.

4. 1. Discounted case

Let $\boldsymbol{P}\in\Pi$ and

$$v_D(\boldsymbol{P})=\lim_{n\to\infty}\sum_{k=0}^{n}\boldsymbol{P}^{(k)}r. \tag{4.3}$$

A policy $\boldsymbol{P}$ is *preferable* to a policy $\tilde{\boldsymbol{P}}$ if

$$v_D(\boldsymbol{P})\geqq v_D(\tilde{\boldsymbol{P}}). \tag{4.4}$$

It is called *optimal* if (4.4) holds for all $\tilde{\boldsymbol{P}}\in\Pi$.

THEOREM 6. *Suppose* $v_D(\tilde{\boldsymbol{P}})<\infty$. *If a Markov matrix* P *is*

$$r_P+Pv_D(\tilde{\boldsymbol{P}})-v_D(\tilde{\boldsymbol{P}})\geqq 0, \tag{4.5}$$

then the stationary policy $\boldsymbol{P}=(P, P, \cdots)$ *is preferable to* $\tilde{\boldsymbol{P}}$.

PROOF. Let $r_n(P)=r_P$ for $n=2$ and zero othewise. Consider only two terms in (3.1) and (3.2) respectively such as $u_1=w_1=0$, $u_2=v_D(\tilde{\boldsymbol{P}})$, $w_2=v_D(\boldsymbol{P})$. By (3.3), we have $f_1=0$ and $f_2=r_P+Pv_D(\tilde{\boldsymbol{P}})-v_D(\tilde{\boldsymbol{P}})$. The condition (4.5) means $f_2\geqq 0$. So Theorem 5 implies $g_1=0$ and $g_2=v_D(\boldsymbol{P})-v_D(\tilde{\boldsymbol{P}})\geqq 0$. Hence (4.4) holds.

4. 2. Average case

Let

$$v_A(\boldsymbol{P})=\liminf_n(n+1)^{-1}\sum_{k=0}^{n}\boldsymbol{P}^{(k)}r. \tag{4.6}$$

$$v_A^{(2)}(\boldsymbol{P})=\liminf_n(n+1)^{-1}\sum_{m=0}^{n}\sum_{k=0}^{m}(\boldsymbol{P}^{(k)}r-v_A(\boldsymbol{P})). \tag{4.7}$$

A policy $\boldsymbol{P}$ is *preferable* to $\tilde{\boldsymbol{P}}$ if

$$v_A(\boldsymbol{P}) \geqq v_A(\tilde{\boldsymbol{P}}). \tag{4.8}$$

If (4.8) holds for all $\tilde{\boldsymbol{P}} \in \Pi$, then $\boldsymbol{P}$ is *average optimal.*

THEOREM 7. *Suppose* $v_A(\tilde{\boldsymbol{P}})$, $v_A^{(2)}(\tilde{\boldsymbol{P}}) < \infty$. *If*

$$Pv_A(\tilde{\boldsymbol{P}}) - v_A(\tilde{\boldsymbol{P}}) \geqq 0 \tag{4.9}$$

or (4.9) *holds with equality and*

$$r_P + Pv_A^{(2)}(\tilde{\boldsymbol{P}}) - v_A(\tilde{\boldsymbol{P}}) - v_A^{(2)}(\tilde{\boldsymbol{P}}) \geqq 0. \tag{4.10}$$

then a stationary policy $\boldsymbol{P}=(P, P, \cdots)$ *is preferable to* $\tilde{\boldsymbol{P}}$.

PROOF. Set $r_n(P)=r_P$ for $n=2$ and zero otherwise. Consider $u_1=v_A(\tilde{\boldsymbol{P}})$, $u_2=v_A^{(2)}(\tilde{\boldsymbol{P}})$ and $w_2=v_A^{(2)}(\boldsymbol{P})$. They satisfy (3.1) and (3.2). By (3.3),

$$g_1 = v_A(\boldsymbol{P}) - v_A(\tilde{\boldsymbol{P}}), \tag{4.11}$$

$$f_1 = Pv_A(\tilde{\boldsymbol{P}}) - v_A(\tilde{\boldsymbol{P}}), \tag{4.12}$$

$$f_2 = r_D + Pv_A^{(2)}(\tilde{\boldsymbol{P}}) - Pv_A(\tilde{\boldsymbol{P}}) - v_A^{(2)}(\tilde{\boldsymbol{P}}). \tag{4.13}$$

If (4.9) holds with equality, then the left hand side of (4.10) equals the right of (4.13). Hence Theorem 3 or Corollary 1 imply $g_1 \geqq 0$ and so we have (4.8).

Conditions (4.9) and (4.10) are a denumerable version of Derman [5]. It is also possible to extend that $(f_1, f_2) \geqq_l 0$ as Corollary 1. From Corollary 2, if both (4.9) and (4.10) hold then $g_1 \geqq 0$. This result dues to Aso and Kimura [1].

4. 3. Nearly optimal case

Let

$$v_D(\beta, \boldsymbol{P}) = \lim_n \sum_{k=0}^{n} \beta^k \boldsymbol{P}^{(k)} r \tag{4.14}$$

for a scaler β $(0<\beta<1)$ which is called a *discount factor* and a policy $\boldsymbol{P} \in \Pi$. A policy $\boldsymbol{P}$ is *preferable* to a policy $\tilde{\boldsymbol{P}}$ if

$$\lim_{\beta \to 1} \{v_D(\beta, \boldsymbol{P}) - v_D(\beta, \tilde{\boldsymbol{P}})\} \geqq 0. \tag{4.15}$$

A policy $\boldsymbol{P}$ is *nearly optimal* if

$$\lim_{\beta \to 1} \{v_D(\beta, \boldsymbol{P}) - U(\beta)\} = 0 \tag{4.16}$$

where

$$U(\beta) = \sup \{v_D(\beta, \tilde{\boldsymbol{P}}) ; \tilde{\boldsymbol{P}} \in \Pi\}.$$

Suppose that a Markov matrix $\tilde{P}$ satisfies Assumption 1. Hence $\tilde{P}^* r_{\tilde{P}}$ and $\tilde{H} r_{\tilde{P}}$ are finite-valued where $\tilde{P}^*$, $\tilde{H}$ are defined by the matrix $\tilde{P}$. Let

$$u_1 = \tilde{P}^* r_{\tilde{P}}, \quad u_2 = \tilde{H} r_{\tilde{P}}. \tag{4.17}$$

They satisfy (3.2). We consider the following inequalities;

$$Pu_1 \geqq u_1, \tag{4.18}$$

$$r_P + Pu_2 \geqq Pu_1 + u_2. \tag{4.19}$$

For a fixed $\tilde{P}$, let $G(\tilde{P})$ be the set of Markov matrices P such that (4.18) holds; (4.18) holds with equality and (4.19) holds; or for each i for which (4.18) and (4.19) holds with equality, $P_{ij}=\tilde{P}_{ij}$ for all $j \in S$.

THEOREM 8. *If a Markov matrix $P \in G(\tilde{P})$, then the stationary policy $\boldsymbol{P}=(P, P, \cdots)$ is preferable to $\tilde{\boldsymbol{P}}=(\tilde{P}, \tilde{P}, \cdots)$. $\boldsymbol{P}$ is nearly optimal if $G(P)=\phi$.*

PROOF. Let w_1 and w_2 be vectors determined for $\boldsymbol{P}$ similarly as (4.17). Since

$$v_D(\beta, P) = (1-\beta)^{-1} w_1 + w_2 + o(1-\beta),$$

the former assertion is equivalent to $(g_1, g_2) \geqq_l 0$. Hence the results are immediate from Theorem 4.

This extends Veinott [14] to a denumerable state with Assumption 1.

4. 4. Sensitive discounted case

As a beginning of the case (3) due to Blackwell [2], Miller, Veinott [11] considered the behaviour of $v_D(\beta, \boldsymbol{P})$ as β tends to 1 and obtained a Laurant expansion about $\beta=1$. We show here that an approximate form of the expansion and discuss the policy improvement. As additional results, an average version are obtained.

A stationary policy $\boldsymbol{P}$ is *preferable* to $\tilde{\boldsymbol{P}}$ under n-th ($n \geqq 1$) sensitive discount case if

$$\lim_{\beta \to 1} (1-\beta)^{-n} \{v_D(\beta, \boldsymbol{P}) - v_D(\beta, \tilde{\boldsymbol{P}})\} \geqq 0. \tag{4.20}$$

If (4.20) holds for all $\tilde{\boldsymbol{P}} \in \Pi$, then $\boldsymbol{P}$ is *optimal*. We consider Markov matrices P and $\tilde{P}$ under Assumption 1. Suppose that $r_n(P)=r_P$ for $n=2$ and zero otherwise. Let $\{w_n\}$ be

$$w_1 = P^* r_P, \tag{4.21}$$

$$w_n = (-1)^n H^{n-1} P^{n-2} r_P, \quad n \geqq 2 \tag{4.22}$$

and $\{u_n\}$ are defined in (4.21) and (4.22) which replaced P by $\tilde{P}$. Then we see that $\{u_n\}$, $\{w_n\}$ satisfy (3.1) and (3.2). Because $(I-P)H=I-P^*$ and $HP=PH$. If we define $\{f_n\}$ and $\{g_n\}$ by (3.3) and (3.4) respectively, the policy improvement of this case is as follows.

THEOREM 9. *If $(f_1, f_2, \cdots f_{n-2})=0$, $(f_{n-1}, f_n) \geqq_l 0$, then a stationary policy $\boldsymbol{P}$ is preferable to $\tilde{\boldsymbol{P}}$ under n-th sensitive discount case.*

PROOF. It is sufficient show that

$$v_D(\beta, \boldsymbol{P}) = (1-\beta)^{-1} w_1 + w_2 + (1-\beta) w_3 + \cdots + (1-\beta)^{N-2} w_N + o(1-\beta). \tag{4.23}$$

This is an approximate form of a Laurant expansion. From the itration (3.1),

$$r_{n+1}(P) = (I-P) w_{n+1} + P w_n = (I-P)(w_{n+1} - w_n) + w_n.$$

Multiplying $(1-\beta)^{n-1}$, then we are summing up and hence

$$r_P = [I-\beta P] \sum_{n=1}^{N} (1-\beta)^{n-2} w_n + (1-\beta)^{N-1} (I-P) w_{N+1}.$$

Since $0<\beta<1$,

$$v_D(\beta, \boldsymbol{P}) = [I-\beta P]^{-1} r_P.$$

Thus we have

$$v_D(\beta, \boldsymbol{P}) = \sum_{n=1}^{N} (1-\beta)^{n-2} w_n + (1-\beta)^{N-1} [I-\beta P]^{-1} (I-P) w_{N+1}.$$

That is, (4.23) holds and so this completes the proof.

Additionally we introduce an average overtaking MDP. Let for

$$\boldsymbol{P} = (P_1, P_2, \cdots) \in \Pi, \quad v_n^{(0)}(\boldsymbol{P}) = \boldsymbol{P}^{(n)} r$$

for $n \geqq 0$,

$$v_n^{(1)}(\boldsymbol{P}) = \sum_{m=0}^{n-1} v_m^{(0)}(\boldsymbol{P})$$

for $n \geqq 1$ and

$$v_n^{(k)}(\boldsymbol{P}) = \sum_{m=1}^{n} v_m^{(k-1)}(\boldsymbol{P})$$

for $n \geqq 1$, $k \geqq 2$. If, for fixed $k \geqq 2$,

$$\liminf_{n\to\infty} n^{-1} \{v_n^{(k)}(\boldsymbol{P}) - v_n^{(k)}(\tilde{\boldsymbol{P}})\} \geqq 0, \tag{4.24}$$

then $\boldsymbol{P}$ is preferable to $\tilde{\boldsymbol{P}}$ under k-fold average overtaking case. The special case of $k=2$ is given by Veinott [14]. We can prove that, extending Lippman [10],

THEOREM 10. *For $k \geqq 1$, k-th sensitive discount case is equivalent to k-fold average overtaking case.*

Hence the same policy improvement as Theorem 9 holds.

References

[1] ASO, H. and KIMURA, M., *An application of Markov potential theory to Markov decision processes*, Int. J. System Sci., **4** (1973), 907-932.

[2] BLACKWELL, D., *Discrete dynamic programming*, Ann. Math. Statist., **33** (1962), 719-726.

[3] BLACKWELL, D., *Discounted dynamic programming*, Ann. Math. Statist., **36** (1965), 226-235.

[4] DERMAN, C., *Markov sequential control processes—denumerable state space*, J. Math. Anal. Appl., **10** (1965), 295-302.

[5] DERMAN, C., Finite State Markovian Decision Processes, Academic Press (1970).

[6] FURUKAWA, N., *Markov decision processes with compact action spaces*, Ann. Math. Satist., **43** (1972), 1612-1622.

[7] HORDIJK, O., Dynamic Programming and Markov Potential Theory, Mathematical Center Tracts (1974).

[8] HOWARD, R.A., Dynamic Programming and Markov Processes, John Wiley (1960).

[9] KEMENY, J.G., SNELL, J.L. and KNAPP, A.M., Denumerable Markov Chains, von Nostrand (1966).

[10] LIPPMAN, S.A., *Criterion equivalence in discrete dynamic programming*, OR, **17** (1969), 920-923.

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

[12] SCHAUFELE, R.A., *A potential theoritic proof of a theorem of Derman and Veinott*, Ann. Math. Statist., **38** (1967), 585-587.

[13] ROTHBLUM, U.G., *Normalised Markov decision chain* I; *sensitive discount optimality*, OR, **23** (1975), 785-795.

[14] VEINOTT, A.F., *On finding optimal policies in discounted dynamic programming with no discounting*, Ann. Math. Statist., **37** (1966), 1284-1294.

[15] VEINOTT, A.F., *Discrete dynamic programming with sensitive discount optimality criteria*, Ann. Math. Statist., **40** (1969), 1635-1660.

[16] WATANABE, H., *Markov programming and potential* (in Japanese), Symposium at Union of Japanese Scientist and Engineers, (1967).