数理解析研究所講究録
第299巻 1977年 93-113



# Policy improvement in Markov Decision Processes and Markov potential theory

千葉大 教養部 安田正実

## 1. Introduction

Markov Decision Process (MDP) と Markov potential theory との関係は，直接的なもの，すなわち MDP における potential theory と間接的なもの，potential theory の結果による MDP のそれの別証明とに分類される。後者に属するものとして Shaufele [11] の論文があるが，前者の方が我々計画法を研究するものにとって興味深い。これについては，まだ理論体系が形成されていない。しかし，渡辺浩[14] は LP の dual problem と関連して，Howard [7] の iteration における単調性の意味を potential の用語で説明したし，N. Furukawa [6] や H. Aso and M. Kimura [1] は potential kernel の性質を用い，policy improvement を証明した。また A. Hordijk [8] も MDP をその定式化から potential の議論として構成している。これらはその理論形成を意図しているように思われる。

しかし，多くの場合，解析的なとり扱いが容易なことから transient chain に対応する case に制限されている。そこで，この報告では，ＭＤＰの policy improvement に役立つよう，一般の chain に対して新しい potential を定義する。そしてこの potential を用いてＭＤＰの policy improvement の解釈とその証明を与えるのが目的である。

まず解釈は「ＭＤＰの reward の増加は improvement における increment を charge とする potential と regular function の和で表わされる」という形である。Transient chain のとき，potential は従来のそれと一致し，regular function は０であって渡辺浩[14]の与えたものに帰着される。また各種の improvement の証明も denumerable state space であっても可能で，その方法の本質はすべてに通じて同一であることがわかる。ここで考えるＭＤＰの policy improvement は次の場合；

(1) discounted case　(2) average case

(3) nearly optimal case

(4) sensitive discounted case

このうち (1)，(2)はＭＤＰの代表的なもので数多くの文献がある。とくに (1)について Howard [7]，Blackwell [2]，[3]，(2)について Howard [7]，Derman

[4], [5] を挙げておく。 Blackwell [2] は (1) と同時に (3) もその存在性を議論している。 その後これを糸口として発展した (4) は Miller and Veinott [10], Veinott [12]; [13] に依るものである。

## 2. Potential theory

$S$ を finite または denumerable な自然数の部分集合、 $P$ を $S \times S$ 上の Markov 行列, $P^*$ を $P$ のベキ $\{P^n\}$ の cesaro 和

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

$I$ は単位行列で lim は pointwise の収束。 この $P^*$ の存在は知られていて 次の性質をもつ。

$$0 \leqq P^* = P^*P = PP^* = P^*P^*, \quad P^*\underline{1} \leqq \underline{1} \qquad (2.2)$$

ここで $\underline{1}$ は恒等的に 1 をとる関数またはベクトルとする。 $\leqq$ は成分毎で, §3 にて用いる "$\overset{l}{\leqq}$" の定義は各行ベクトルの差について最初に 0 でない要素があれば, それが正のときと定める。 関数 $f$ が $P$ について regular とは $Pf = f$ のとき。 Markov chain についての定義などは Kemeny, Snell and Knapp[9]を参照されたい。

<u>定義 1</u> $S$ 上の関数 $f$ が Markov 行列 $P$ に関する charge であるとは

$$H_n = \sum_{k=0}^{n-1} (P^k - P^*) \qquad (n \geq 1) \qquad (2.3)$$

としたとき，$H_n f$ が well defined で $\{H_n f\}$ がある関数 $g$ に cesaro 総和可能であるとき。これを $g = Hf$ と表わし，$g$ を charge $f$ の $P$ に関する potential という。すなわち

$$g = Hf = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} H_k f \qquad (2.4)$$

とくに $S$ が finite ならば $I - P + P^*$ が正則（Blackwell [2]）であるから

$$H = [I - P + P^*]^{-1} - P^* \qquad (2.5)$$

となる。また $P$ が single chain で transient ならば

$$H = [I - P]^{-1} \qquad (2.6)$$

で，従来の potential と一致している。この (2.3) で定めた potential kernel $H$ のもつ次の性質が，後の議論では有効な働きをする。

<u>定理1</u>　関数 $f, g$ において $P^* g < \infty$,

$$f = (I - P) g \qquad (2.7)$$

とする。そのとき $Hf < \infty$ で

$$g = Hf + h \qquad (2.8)$$

ここで $h$ は regular で $h = P^* g$

(証明)　(2.7) と $H_n$ の定義 (2.3) から

$$H_n f = g - P^{n+1} g \qquad (n \geq 1) \qquad (2.9)$$

$P^*g < \infty$ であるから $\{P^n g\}$ は cesaro 総和可能。

したがって (2.9) より $\{H_n f\}$ もそうで $Hf < \infty$

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

を得る。 ここで $h = P^*g$ とおくと (2.10) は (2.8) と同じ式である。 ∎

この定理から Riesz 分解—super regular function は non negative charge と regular function に一意に分解される—も成り立つことがわかる。

定理2 (1) 関数 $f$ が non negative charge でかつ $P^*f = 0$ ならば potential

$$Hf \geq 0 \tag{2.11}$$

(2) $f$ が regular ならば

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

(証明) (1) $H_n f = \sum_{k=0}^{n-1} P^k f$, $f \geq 0$ より $H_n f \geq 0$ で $n$ について単調増加で $Hf < \infty$ から (2.11) を得る。

(2) すべての $n$ に対して $f = P^*f = P^n f$ であるから $H_n f = 0$ (2.12) が成り立つことは明らか。 ∎

とくに $P$ が transient であれば $P^* = 0$ であって (1) は potential $H$ が pure であることを示している。

3. Theorems　この§では policy improvement に用いる一般的な定理を述べる。

定義2　$P$, $\hat{P}$ を2つの Markov 行列とし, $\{r_n(P)\}_1^\infty$ を $P$ によって定まるあらかじめ与えられた $S$ 上の有界関数列とする。

(1)　$\{w_n = w_n(P)\}$, $\{u_n = u_n(\hat{P})\}$ は次を満たすとする。

$$(I-P)\,w_{n+1} + P\,w_n = r_{n+1}(P) \qquad n \geq 0 \quad (3.1)$$

$$(I-\hat{P})\,u_{n+1} + \hat{P}\,u_n = r_{n+1}(\hat{P}) \qquad n \geq 0 \quad (3.2)$$

ただし　$w_0 = u_0 = 0$ とする。

(2)　上の(1)で定めた $\{u_n\}$, $\{w_n\}$ に対して　$(n \geq 1)$

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

$$g_n = w_n - u_n \qquad (3.4)$$

次の§で, これらの具体的な意味ははっきりするから, いまの段階では簡単にふれておく。　この $\{w_n\}$, $\{u_n\}$ は §1 での(1)~(4)などの case に対する MDP における reward で, それぞれ stationary policy に対応しているものである。　しかし 列の中で, 1つだけが目的の reward であって, それ以外は補助的な reward の役割をしている。このパラメータ $n$ は時間ではなく1つの MDP における reward の order のようなものを表わしている。　(3.4)はしたがって reward の increment となる。　また (3.3) は

reward $\{u_n\}$ に対する policy improvement の increment である。 この $\{f_n\}$ の形は policy を perturb すれば得られるが, 通常の Dynamic Programming や Linear programming にも関連している。 MDP の policy improvement すべてにわたる原則は $g_n$ すなわち reward の increment を正にすることであるが, それは $f_n$ の符号に定まるということが本質である。 §4で述べることにしておき, そのための準備としていくつかの定理を用意する。

<u>補題1</u> (1) $\{w_n\}$ は (3.1) の再帰式を満たすとする。任意の関数列 $\{u_n\}$ ( (3.2)を満たさなくてもよい ) に対して (3.3), (3.4) で $\{f_n\}$, $\{g_n\}$ を定めれば

$$f_n = (I-P)g_n + P g_{n-1} \quad (n \geq 1) \tag{3.5}$$

$$f_0 = 0 \quad g_0 = 0$$

(2) もし $P^* g_n < \infty$ ならば $H(f_n - P g_{n-1}) < \infty$ で

$$g_n = H(f_n - P g_{n-1}) + P^* g_n \quad (n \geq 1) \tag{3.6}$$

さらに $P^* g_{n+1} < \infty$ ならば $P^* f_{n+1} < \infty$ で

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

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

(証明) (1) 定義2 (2)において $u_n$ を消去すればよい。$w_n$ が (3.1) をみたすことを用いれば (3.5) が成りたつ。

(2) $f = f_n - P g_{n-1}$, $g = g_n$ とおいて 定理1を適用。∎

この補題で $P$ に関して $g_{n-1}$ $(n \geq 1)$ が regular であるとすれば，$Pg_{n-1}$ も regular となるから，定理2(2)によって (3.7) は

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

となる。　したがって (3.9) を解釈すれば

「reward increment $g_n$ は policy improvement の increment $f_n$ を charge とする potential $Hf_n$ と regular function $P^*f_{n+1}$ との和に書き表わされる。」

もし $P$ が transient ならば $P^*=0$ で (3.9) は

$$g_n = H f_n \tag{3.10}$$

となり，「policy improvement の increment を charge とする potential」(渡辺浩[4]) で表わされる。　$f_n \geq 0$ ならば (2.11) より $g_n \geq 0$ で，これは Howard の policy improvement の単調性を示している。　transient でないときは，$f_n \geq 0$ であっても $g_n \geq 0$ とはならない。　しかし，定理3で述べるように，$\{f_n\}$ のうちで $f_{n-1}$, $f_n$ の2つの正負が $g_{n-2}$, $g_{n-1}$ の正負を決定する。これは §1 の (2), (4) に対する原則で，さきの transient のときが (1) に対応している。　(3) については特別で，$g_n$ の正負も結論とするので，さらに仮定が必要である。

いま次の仮定を設ける。　以下の定理はことわらない限

りこれを仮定するものとする。

仮定1　Markov 行列 P について state space S は

$$S = \bigcup_{k=1}^{\infty} R_k \cup T \tag{3.11}$$

で表わされ，各 $R_k$ は finite recurrent class，T は transient class とする。

S が有限ならば これはつねに成りたっている。

与えられた P と $\{r_n\}$ によって $\{f_n\}$, $\{g_n\}$ を (3.3) (3.4) で定めたとする。

定理3　$\{w_n\}$ が (3.1) を満たし $n \geq 4$ で $P^* g_i < \infty$ $i = 1, \cdots, n-1$ とする。

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

$$(f_{n-1}, f_n) \overset{\ell}{\geqq} 0 \tag{3.13}$$

ならば 次が成り立つ

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

$$(g_{n-2}, g_{n-1}) \overset{\ell}{\geqq} 0 \tag{3.15}$$

(証明)　(3.12) から (3.14) を帰納法で導びく。 (3.8) より $g_1 = H f_1 + P^* f_2$ だから $(f_1, f_2) = 0$ ならば $g_1 = 0$ なることは明らか。 (3.12) のとき $g_{n-3} = 0$ を示せばよい。

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

で，帰納法の仮定によって $g_{n-4} = 0$ と $f_{n-3} = f_{n-2} = 0$ から $g_{n-3} = 0$ を得，(3.14) が示される。

次に (3.15) を証明する。 (3.12), (3.14) より (3.7) は

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

の形。 これは (2.2) より regular であるから (2.12) を用いると

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

を得る。 (3.13) より $f_{n-1} \geq 0$ だから (3.16) によって $g_{n-2} \geq 0$ は明らか。 したがって (3.15) を示すには各 $i \in S$ に対し $g_{n-2}(i) = 0$ ならば $g_{n-1}(i) \geq 0$ をいえばよい。 ここで (3.11) の仮定を用いる。

1) $i \in R_k$ のとき

$P^*$ の $i, j \in R_k$ 要素は positive であって, いま

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

だから $f_{n-1}(j) = 0$ $j \in R_k$ よって (3.13) から $f_n(j) \geq 0$ $j \in R_k$。 (3.17) から

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

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

2) $i \in T$ のとき

(3.16) によって $\sum_{j \in S} P^*_{ij} f_{n-1}(j) = 0$。 (3.13) から $f_{n-1} \geq 0$ だから 定理2 (1) より

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

また $p^*_{ij} > 0$ $j \in \bigcup_k R_k$ より

$$f_{n-1}(j) = 0 \qquad j \in \bigcup_k R_k$$

(3.13) より

$$f_n(j) \geqq 0 \qquad j \in \bigcup_k R_k$$

これから

$$\sum_{j \in \bigcup_k R_k} p^*_{ij} f_n(j) \geqq 0 \tag{3.19}$$

$i \in T$ に対して (3.17) は

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

であるから, (3.18), (3.19) より $g_{n-1}(i) \geqq 0$ ▨

系1 $(f_1, f_2) \overset{\ell}{\geqq} 0$ ならば $g_1 \geqq 0$

(証明) $p^* f_1 = 0$, $g_1 = H f_1 + p^* f_2$ より同様に証明できる。 ▨

系2 $p$ について 仮定1 なしで $p^* g_1 < \infty$, $p^* f_2 < \infty$ とする。

$f_1 \geqq 0$ かつ $f_2 \geqq 0$ ならば $g_1 \geqq 0$

(証明) 系1 の証明中の関係式と定理2(1) より明らか。 ▨

定理4 定理3につけ加え, $f_{n-1}(i) = f_n(i) = 0$ なる $i \in S$ に対しては $p_{ij} = \tilde{p}_{ij}$ $\forall j \in S$ とする。 もし $g_{n-2} = 0$ ならば $(g_{n-1}, g_n) \overset{\ell}{\geqq} 0$

(証明) $g_{n-2} = 0$ とすれば (3.7), (3.8) より

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

$$g_{n-1} = H f_{n-1} + p^* f_n \tag{3.21}$$

$$g_n = H[f_n - P g_{n-1}] + P^* g_n \qquad (3.22)$$

を得る。 (3.15) において $g_{n-1} \geq 0$ であるから $g_{n-1}(i) = 0$ のとき $g_n(i) \geq 0$ を示せば十分である。 $f_{n-1} \geq 0$ で (3.20) より

$$f_{n-1}(i) = 0 \qquad i \in R = \bigcup_{k=1}^{\infty} R_k \qquad (3.23)$$

も成りたっている。 もし

1) $i \in R$ ならば

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

$$= \sum_R P^*_{ij} f_n(j)$$

$g_{n-1}(i) = 0$ としているから

$$f_n(j) = 0 \qquad j \in R \qquad (3.24)$$

(3.22) において (3.24) を用いれば

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

$$= \sum_R P^*_{ij} g_n(j)$$

(3.23), (3.24) であるから 仮定(定理4)をみたす。

したがって この $i \in S$ に対しては $P_{ij} = \tilde{P}_{ij}$ $\forall j \in S$

このときには 明らかに

$$u_n(j) = w_n(j) \qquad j \in R$$

したがって $g_n(j) = u_n(j) - w_n(j) = 0 \qquad j \in R$

2) $i \in T$ ならば (3.21) において (3.23) を用いると

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

ここで $H_{ij} > 0 \quad j \in T \cap T_0^c$, $H_{ij} = 0 \quad j \in T \cap T_0$

$P^*_{ij} > 0 \quad j \in R \cap R_0^c$, $P^*_{ij} = 0 \quad j \in R \cap R_0$ とすれば

$$f_{n-1}(j) = 0 \qquad j \in T \cap T_0^c \tag{3.25}$$

$$f_n(j) = 0 \qquad j \in R \cap R_0^c \tag{3.26}$$

(3.23), (3.25) より $f_{n-1}(j) = 0 \quad j \in R \cup T_0^c$ で (3.25) (3.26) を用い, $R_0$, $T_0$ の定め方から

$$g_{n-1} = H f_{n-1} + P^* f_n = 0$$

1) の結果を用いれば $g_{n-1}(j) = 0 \quad j \in R$ より

$$g_n(j) = 0 \qquad j \in R \tag{3.27}$$

を得る。 これから $i \in T$ に対する (3.22) は

$$g_n(i) = \sum_{R \cap R_0} H_{ij} f_n(j) + \sum_{T \cap T_0^c} H_{ij} f_n(j)$$

となる。

$\sum_{R \cap R_0} P^*_{ij} f_n(j) = 0$, $f_n(j) \geq 0 \quad j \in R_0 \cup T_0^c$ であるから, 定理2 (1) より $g_n(i) \geq 0$ となり, これが求める結果である。 ∎

P に制限を加えれば

<u>定理5</u> Markov 行列 P は transient とする。

$(f_1, \cdots, f_{n-1}) = 0$, $f_n \geq 0 \quad (n \geq 2)$ ならば

$$g_{n-1} = 0, \quad g_n \geq 0$$

(証明) (3.10) より明らか。 ∎

4. Policy improvement of MDP　一般的な記号を定め，各論に分けて §1 の (1) ~ (4) を論ずる。

$S$ を state space で finite または denumerable。

Action space を表に出さず transition probability でおきかえる。つまり policy の定義を

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

ただし $P_i$, $i=1,2,\cdots$ は $S \times S$ 上の Markov 行列とする。

$\underline{\underline{P}}$ の全体を $\Pi$ と表わす。Markov 行列 $P$ によって定まる $S$ 上の有界関数を $r_P$ とするとき

$\underline{\underline{P}} = (P_1, P_2, \cdots) \in \Pi$ に対して

$$\underline{\underline{P}}^{(n)} r = P_1 \cdots P_n r_{P_{n+1}} \qquad n \geq 1 \tag{4.2}$$

と書き，これを $n$ 期の expected reward という。ただし $\underline{\underline{P}}^{(0)} r = r$。この (4.2) の関数列をいかに total するかによって §1 の (1) ~ (4) の case が考えられている。

4.1. Discounted case　$\underline{\underline{P}}, \underline{\underline{\tilde{P}}} \in \Pi$ とする。

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

などとするとき，$\underline{\underline{P}}$ の方が $\underline{\underline{\tilde{P}}}$ より better とは

$$v_D(\underline{\underline{P}}) \geq v_D(\underline{\underline{\tilde{P}}}) \tag{4.4}$$

(4.4) がすべての $\underline{\underline{\tilde{P}}} \in \Pi$ でなりたつとき $\underline{\underline{P}}$ を discount optimal という。

定理 6　$v_D(\underline{\underline{\tilde{P}}}) < \infty$ とする。Markov 行列 $P$ に対し

$$r_P + P v_D(\tilde{\underline{P}}) - v_D(\hat{\underline{P}}) \geqq 0 \qquad (4.5)$$

ならば stationary policy $\underline{P} = (P, P, \cdots)$ の方が $\hat{\underline{P}}$ より better である。

(証明) 定義 2で $r_n(P) = r_P \quad n=2$, $n \neq 2$ では $r_n(P) = 0$ とおく。 また $u_1 = w_1 = 0$, $u_2 = v_D(\hat{\underline{P}})$ $w_2 = v_D(\underline{P})$ とする。 こうすれば (3.3) より

$$f_1 = 0, \quad f_2 = r_P + P v_D(\tilde{\underline{P}}) - v_D(\hat{\underline{P}})$$

(4.5) から $f_2 \geqq 0$ であるから 定理5によって

$$g_1 = 0, \quad g_2 = v_D(\underline{P}) - v_D(\hat{\underline{P}}) \geqq 0$$

よって (4.4) がなりたつ。 ∎

4.2. Average case $\underline{P} \in \Pi$ に対し

$$v_A(\underline{P}) = \lim_{n\to\infty} \inf \frac{1}{n+1} \sum_{k=0}^{n} \underline{P}^{(k)} r \qquad (4.6)$$

$$v_A^{(2)}(\underline{P}) = \lim_{n} \inf \frac{1}{n+1} \sum_{m=0}^{n} \sum_{k=0}^{m} (\underline{P}^{(k)} r - v_A(\underline{P})) \qquad (4.7)$$

とする。 $\underline{P}, \tilde{\underline{P}} \in \Pi$ に対し

$$v_A(\underline{P}) \geqq v_A(\tilde{\underline{P}}) \qquad (4.8)$$

のとき $\underline{P}$ の方が better といい, すべての $\hat{\underline{P}}$ に対してなりたつとき $\underline{P}$ を average optimal という。

定理7 $v_A(\hat{\underline{P}}), v_A^{(2)}(\hat{\underline{P}}) < \infty$ とする。

$$P v_A(\hat{\underline{P}}) - v_A(\hat{\underline{P}}) \geqq 0 \qquad (4.9)$$

または $P v_A(\tilde{\underline{P}}) - v_A(\hat{\underline{P}}) = 0$, $r_P + P v_A^{(2)}(\tilde{\underline{P}}) - v_A(\hat{\underline{P}}) - v_A^{(2)}(\hat{\underline{P}}) \geqq 0$ (4.10)

ならば $\underline{P}=(P,P,\cdots)$ のほうが $\underline{\tilde{P}}$ より better つまり (4.8) がなりたつ。

(証明) $r_n(P)=r_P$, $n=2$, $r_n(P)=0$, $n\neq 2$ として $u_1=v_A(\underline{\tilde{P}})$, $w_1=v_A(\underline{P})$, $u_2=v_A^{(2)}(\underline{\tilde{P}})$, $w_2=v_A^{(2)}(\underline{P})$ とおけば (3.3) より

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

$$f_1 = P v_A(\underline{\tilde{P}}) - v_A(\underline{\tilde{P}}) \tag{4.12}$$

$$f_2 = r_P + P v_A^{(2)}(\underline{\tilde{P}}) - v_A(\underline{\tilde{P}}) - v_A^{(2)}(\underline{\tilde{P}}) \tag{4.13}$$

となる。 ここで 定理3または系1によって $g_1 \geq 0$ つまり (4.8) を得る。 ∎

[注意] 定理7の (4.9), (4.10) の条件は $(f_1,f_2)\overset{l}{\geq}0$ より強い。 これは Derman [5] を denumerable にいいかえたからである。 (4.9) と (4.10) の後半が同時になりたてば系2より $g_1\geq 0$ となる (Aso and Kimura [1]) が $P^*g_1$, $P^*f_2<\infty$ または $v_A(\underline{\tilde{P}})$, $v_A^{(2)}(\underline{\tilde{P}})$ の有界性が必要となる。

4.3. Nearly optimal case 以下では stationary policy $\underline{P}=(P,P,\cdots)$ $\underline{\tilde{P}}=(\tilde{P},\tilde{P},\cdots)$ を考える。 $0<\beta<1$ で

$$v_D(\beta,\underline{P}) = \lim \sum_{k=0}^{n} \beta^k P^k r_P \tag{4.14}$$

とする。 $\underline{P}$ のほうが $\underline{\tilde{P}}$ より better であるとは

$$\lim_{\beta\to 1} \{v_D(\beta,\underline{P}) - v_D(\beta,\underline{\tilde{P}})\} \geq 0 \tag{4.15}$$

$\underline{P}$ が nearly optimal とは

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

ただし $U(\beta)=\sup\{v_D(\beta,\underline{P});\ \underline{P}\in\Pi\}$

Markov 行列 $\tilde{P}$ に対して cesaro 和を $\tilde{P}^*$, potential kernel を $\tilde{H}$ とするとき, 仮定1のもとでは $\tilde{P}^* r_{\tilde{P}}$, $\tilde{H} r_{\tilde{P}} < \infty$ となるから

$$u_1 = \tilde{P}^* r_{\tilde{P}},\quad u_2 = \tilde{H} r_{\tilde{P}} \qquad (4.17)$$

とおく。 この $u_1, u_2$ は (3.2) をみたす。 いま

$$P u_1 \geqq u_1 \qquad (4.18)$$

$$r(P)+P u_2 \geqq P u_1 + u_2 \qquad (4.19)$$

を考える。

$$G(\tilde{P})=\{P;\ (4.18)\text{が成立, (4.18)が等号で(4.19)が成立}$$

$$\text{または (4.18)(4.19) が等号でそのi}\in S\text{に対しては}$$

$$P_{i,j}=\tilde{P}_{i,j}\quad \forall j\in S\}$$

とすれば

<u>定理8</u> $P\in G(\tilde{P})$ ならば $\underline{P}=(P,P,\cdots)$ は $\underline{\tilde{P}}=(\tilde{P},\cdots)$ よりも better 。 $G(P)=\phi$ ならば $\underline{P}$ は nearly optimal。

(証明) $w_1, w_2$ を $P$ に対して (4.17) と同様に定める。

$v_D(\beta,\underline{P})=\frac{1}{1-\beta}w_1+w_2+o(1-\beta)$ と展開できるから,

$(g_1,g_2)\overset{\ell}{\geqq}0$ と $\lim\{v_D(\beta,\underline{P})-v_D(\beta,\underline{\tilde{P}})\}\geq 0$ とは同値。 したがって 定理4から得られる。 ∎

この定理は $S$ が finite のとき Veinott [12] が得たものを denumerable へと拡張している。 ただし denumerable であるので，収束の条件が必要となって，仮定1が必要である。 もちろん この仮定1は finite のとき 自動的に満たされ，収束などの議論は気にしなくてよい。

4.4. Sensitive discounted case Discount case での Blackwell [2] による (1), (3) の結果を糸口として Miller, Veinott 等は (4.14) を $1-\beta$ で Laurant 展開することを考えた。 これが sensitive discounted case であって 次のように定義される。

Stationary policy $\underline{\underline{P}}$ が $\underline{\underline{\tilde{P}}}$ より $n$ th order で better とは

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

また $n$-th order で optimal とは (4.20) が すべての $\underline{\underline{\tilde{P}}}$ に対して成り立つときをいう。

$$w_1 = P^* r_P \quad (4.21)$$

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

と定義する。 $u_n$ は $P$ のかわりに $\tilde{P}$ に対して定める。

$$g_n = w_n - u_n \quad (n \geq 1) \quad (4.23)$$

$$f_n = r_n(P) - (I-P)u_n - P u_{n-1} \quad (n \geq 1) \quad (4.24)$$

ただし　$r_n(P) = r_P$, $n=2$,　$r_n(P) = 0$, $n \neq 2$,　$u_0 = 0$

定理9　$(f_1, \cdots, f_{n-2}) = 0$,　$(f_{n-1}, f_n) \lneqq 0$

ならば　stationary policy $\underline{\underline{P}} = (P, P, \cdots)$ は $\underline{\underline{\tilde{P}}} = (\tilde{P}, \hat{P}, \cdots)$ より $(n-1)$th order で better, つまり

$$\lim_{\beta \to 1} (1-\beta)^{-n+1} \{ v_D(\beta, \underline{\underline{P}}) - v_D(\beta, \underline{\underline{\tilde{P}}}) \} \geqq 0 \qquad (4.25)$$

が成立。

(証明)　(4.21), (4.22) などで定めた $\{u_n\}$, $\{w_n\}$ は (3.1), (3.2) を満たすことがわかる。　したがって (4.23) (4.24) で定めた $\{f_n\}$, $\{g_n\}$ は (3.3), (3.4) と同じものである。　よって　定理9を示すには

$$(g_1, \cdots, g_{n-1}) \lneqq 0 \qquad (4.26)$$

と (4.25) が同値であれば　定理3より明らか。　その同値であることは

$$v_D(\beta, \underline{\underline{P}}) = (1-\beta)^{-1} u_1 + w_2 + (1-\beta) w_3 + \cdots$$
$$\cdots + (1-\beta)^N w_{n+2} + o(1-\beta)^{N+1} \quad (N \gg n) \qquad (4.27)$$

となるから。　(4.27) が成り立つことは (3.1) の再帰式を使って証明する。　∎

この定理と同様な結果は "average overtaking optimal" な case についても 成り立つが, ここでは省略する。

# 参考文献


[1] Aso, H. and M. Kimura ; An application of Markov potential theory to Markov decision processes, INT. J. System sci. (1973)

[2] Blackwell, D. ; Discrete dynamic programming, AMS (1962)

[3] ——— ; Discounted dynamic programming, AMS (1965)

[4] Derman, C ; Markov sequential control processes — denumerable state space, J. Math. Anal. Appl. (1965)

[5] ——— ; Finite state Markovian decision processes, Academic Press (1970)

[6] Furukawa, N. ; Markov decision processes with compact action spaces, AMS (1972)

[7] Howard, R.A. ; Dynamic programming and Markov proceses, Wiley (1960)

[8] Hordijk, A. ; Dynamic programming and Markov potential theory, Mathematical Center Tracts (1974)

[9] Kemmeney, J.G., J.L. Snell and A.W. Knapp ; Denumerable Markov chains, Van Nostrand (1966)

[10] Miller, B.L. and A.F. Veinott, Jr. ; Discrete dynamic

programming with a small interest rate, AMS (1969).

[11] Schaufele, R.A. ; A potential theoritic proof of a theorem of Derman and Veinott, AMS (1967)

[12] Veinott, A.F. Jr. ; On finding optimal policies in discounted dynamic programming with no discounting AMS (1966)

[13] ——— ; Discrete dynamic programming with sensitive discount optimality criteria, AMS (1969)

[14] 渡辺浩 ; マルコフ計画法とポテンシャル, 日科技連計画シンポジウム (1967)