

可算状態マルコフ決定過程の sensitive discount 最適系について

和歌山大 教育 門田良信

§1. 序

有限状態空間を持つ定常マルコフ決定過程の sensitive discount 最適系については, Blackwell [1], Miller and Veinott [8] および Veinott [10] 等により, よく研究されている. 彼等の結果の幾つかを可算状態空間へ拡張する事が, ここでの目的である.

§2 では定義および表記法を与える.

§3 では, 各定常政策に対応したマルコフ連鎖が平均時間エルゴード定理を満たす事を条件として, ρ -割引総期待利得の Laurent 級数展開を導く.

§4 では §3 と同じ条件の下で, Veinott [10] による政策改良法を拡張する.

§5 では, §3 の条件に定常政策に関する一種の一様性の条件を追加して, ∞ -割引最適定常政策が存在する事を示す.

これらの問題は最近よく研究されているようである。一般に $n \geq 1$ なる場合の n -割引最適系は、Laurent 級数展開を使って考察されるが、この級数は無条件では得る事ができない。以下の記述における条件は、Hordijk and Sladky [5], Hordijk [4], Taylor [9], Wijngaard [11] 等と比べると、推移確率に関する条件がゆるい点に特徴がある。従って [8], [10] の完全な拡張となっている。また [9], [11] は一般の状態空間においてそれぞれ、Laurent 級数展開、0-割引最適定常政策の存在を導いている。

§2. 準備.

可算状態空間を S で表わす。各状態 $i \in S$ において取り得る決定の有限集合を A_i とする。各 $i \in S$ および $a \in A_i$ に対しては、 S 上の推移確率 ($p_a(i, j); j \in S$) と利得 $r(i, a)$ (実数値) が定まっているものとする。ここで、 $\sum_{j \in S} p_a(i, j) = 1$ であり、 $r(i, a)$ は i, a に関して一様有界であると仮定しておく。 $F = \prod_{i \in S} A_i$ (直積) とする。 F の元を f , その第 i 要素を $f(i)$ と表わすと、各 f に対して、確率行列 $P(f) = (p_{f(i)}(i, j); i, j \in S)$ と利得の列ベクトル $r(f) = (r(i, f(i)); i \in S)$ が定まる。このように定義された組 $\mathcal{M} = (S, F, \{P(f)\}_{f \in F}, \{r(f)\}_{f \in F})$ を定常マルコフ決定過程と呼ぶ事にする。

F の元の無限列 $\pi = (f_1, f_2, \dots)$ を政策と呼ぶ。特にすべての n について、 $f_n = f$ ならば、 π を定常政策と呼ぶ。従って定常政策 π は F の元 f と同一視されるので、 π を f で表す事にする。任意の政策 $\pi = (f_1, f_2, \dots)$ に対して S 上のベクトル $V_p(\pi)$, $\chi(\pi)$ を、それぞれ

$$V_p(\pi) = \sum_{n=0}^{\infty} \beta^{n+1} P(f_1)P(f_2)\cdots P(f_n)r(f_{n+1}),$$

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

と定義する。但し、 $\beta = \frac{1}{1+\rho}$, $\rho > 0$ とし、 $n=0$ のとき、 $P(f_1)P(f_2)\cdots P(f_n) = I$ (単位行列) とする。 $V_p(\pi)$ の各要素は、システムが状態 i から出発して政策 π に従った時に得られる利得を、各時間毎に β だけ割引いたものの総和を表わしている。従って、 $V_p(\pi)$ を ρ 割引 (または β -割引) 総期待利得と呼ぶ。同様の意味で、 $\chi(\pi)$ を平均期待利得と呼ぶ。

ある政策 π^* が他のどのような政策 π に対しても、 $V_p(\pi^*) \geq V_p(\pi)$ を満たしていれば、 π^* を ρ -割引最適と呼び、 $\chi(\pi^*) \geq \chi(\pi)$ を満たしているならば、平均最適であると呼ぶ。また任意の $n = -1, 0, \dots$ に対して、

$$(1) \quad \liminf_{\rho \rightarrow 0^+} \rho^{-n} \{ V_p(\pi^*) - V_p(\pi) \} \geq 0$$

を満たしているならば、 n -割引最適であると呼ぶ。すべての n について π^* が n -割引最適ならば、 ∞ -割引最適と呼ぶ。定常政策の集合 F の中だけで考えるならば、 ∞ -割引最適であるた

めの必要十分条件は、任意の $f \in F$ に対して $p_0 = p_0(f)$ が存在して、 $0 < p < p_0$ ならば $V_p(f^*) \geq V_p(f)$ が成立する事である。

上記3つの最適系については、任意の p に対して p -割引最適定常政策が存在する事が、Blackwell [2] によって知られている。平均最適および n -割引最適政策は、状態空間が可算以上になった場合には無条件には存在せず、また存在関係は入り組んだものとなる (Flynn [3] 参照)。以下においては主に n -割引最適系を考察するが、そのために全般を通じて用いられる条件も次に記す。

確率行列 $P(f)$ によって定義されるマルコフ連鎖に関して、次の表記も約束する。任意の $k=1, 2, \dots$ について、 $p_f^{(1)}(i, j) = p_f(i, j)$, $p_f^{(k)}(i, j) = \sum_{l \in S} p_f^{(1)}(i, l) p_f^{(k-1)}(l, j)$ と表わす。また、 $p_f^{(k)}(i, E) = \sum_{j \in E} p_f^{(k)}(i, j)$ とする。 $p_f^{(k)}$ の Cesàro 極限 $p_f^*(i, j) = \lim_{n \rightarrow \infty} \frac{1}{n+1} \sum_{k=0}^n p_f^{(k)}(i, j)$ が存在する事は、よく知られている。 $p_f^{(k)}(i, E)$ と同様にして $p_f^*(i, E)$ が定義される。また、 $p_f^*(i, j)$ も (i, j) 要素とする行列を $P(f)^*$ で表わす。

条件 I. 各 $f \in F$ に対して次の式を満たす定数 B_f が存在する。すべての自然数 n , $i \in S$, $E \subset S$ に対して、

$$(2) \quad \left| \sum_{k=0}^n (p_f^{(k)}(i, E) - p_f^*(i, E)) \right| \leq B_f$$

が成立する。

(2) 式の両辺を n で割ってみると、いわゆるマルコフ連

鎖の平均時間エルゴード定理である。従って、例えば $P(f)$ が Doeblin 条件を満たしていれば、条件 I は成立する。(Yosida and Kakutani [12] 参照) 条件 I によれば、 $P(f)^*$ が確率行列となる事も明らかである。従って、 $\chi(f) = P(f)^* r(f)$ が成立する。

§3. $V_p(f)$ の Laurent 級数展開.

この § では任意の $f \in F$ を固定して考える。従って f の表記も省略して、 $P(f)$, P_f , $r(f)$, $\chi(f)$, $V_p(f)$ 等を P , p , r , χ , V_p 等と表わす。

$S \times S$ の行列 H_p と S 上のベクトル h_p を、

$$H_p = \sum_{n=0}^{\infty} p^{n+1} (P^n - P^*), \quad h_p = H_p r$$

と定義する。 V_p の Laurent 級数展開を導くために、まず $\lim_{p \rightarrow 0+} H_p$, $\lim_{p \rightarrow 0+} h_p$ の存在とその性質について調べる。

補助定理 1. 条件 I のもとで、

(a) $\|h_p\| \leq C$ を満たす定数 C が存在する。但し、ベクトルのノルム $\|\cdot\|$ は、各要素の絶対値の $\sup$ で定義する。

(b) $\lim_{p \rightarrow 0+} \|h_p - h\| = 0$ を満たすベクトル h が存在する。

(c) $H = \lim_{p \rightarrow 0+} H_p$ が存在する。任意の $i \in S$, $E \subset S$ に対して、 $H_p(i, E) = \sum_{j \in E} (H_p)_{ij}$ と表わせば、 $H(i, E) = \lim_{p \rightarrow 0+} H_p(i, E)$

$i, E)$ が存在して、 $H(i, E) = \sum_{j \in E} (H)_{ij}$ が成立する。

略証. (a) は [7] によって示されている. 集合 E の特性関数を r とおけば, (c) は (b) から直ちに得られる. 従って (b) を示せば十分である. h_p で $p \rightarrow 0+$ とした時の集積点の 1 つを h_0 とする. $0 \leq p < \infty$ なる任意の p に対して, h_p は

$$(3) \quad (I - \beta P) h_p = \beta (I - P^*) r, \quad P^* h_p = 0$$

を満たす唯一の有界ベクトルである. この事より $\lim_{p \rightarrow 0+} h_p = h_0$ が示される. また (3) を n 回反復して用いれば,

$$h_p = \frac{1}{n} \sum_{m=0}^{n-1} \sum_{k=0}^m \beta^{k+1} (P^k - P^*) r + \frac{1}{n} \sum_{k=1}^n \beta^k (P^k - P^*) h_p$$

が成立する. (a) と条件 I により, $\lim_{p \rightarrow 0+} \|h_p - h_0\| = 0$ が容易に示される. $\square$

次の補助定理 2 および定理 1 (a) は Miller and Veinott [8] の拡張であり, 定理 1 (b) は Veinott [10] の拡張である. その証明も補助定理 1 を使って, 彼等の方法と同じ方針でなされる.

補助定理 2. $0 < p < \infty$ とする. 条件 I のもとで,

(a) $M_p = \sum_{n=0}^{\infty} \beta^{n+1} P^n$ とおくと, $P^* M_p = \frac{1}{p} P^*$ か $M_p = P^* M_p + H_p$ が成立する.

(b) $L_p = \sum_{n=0}^{\infty} p^n (-1)^n H^n$ とおくと, $(I + pH) L_p = L_p (I + pH) = I$ および $H_p = L_p H = H L_p$ が成立する.

定理 1. 条件 I が成立するとする.

(a) $0 < \rho < \frac{1}{2 \|H\|}$ なる ρ もとると.

$$(4) \quad V_\rho = \frac{1}{\rho} y_{-1} + \sum_{n=0}^{\infty} \rho^n y_n$$

が成立する. 但し, $y_{-1} = x$, $n=0, 1, \dots$ なる n に対しては,

$y_n = (-1)^n H^{n+1} r$ とする. また行列 H のノルムは, $\|H\| =$

$\sup_{i \in S} \sup_{E \in S} |H(i, E)|$ で定義する.

(b) 任意の $n \geq -1$ に対して, $(y_{-1}, y_0, \dots, y_n)$ は方程式

$$(5) \quad y_{-1} - \rho y_{-1} = 0, \quad y_{-1} + y_0 - \rho y_0 = r,$$

$$y_{k-1} + y_k - \rho y_k = 0, \quad k=1, 2, \dots, n$$

を満たす. 逆に $n \geq 0$ に対して, 有界なベクトルの列 $(x_{-1},$

$x_0, \dots, x_n)$ が上記方程式の解になっていければ,

$$x_k = y_k, \quad k=-1, 0, \dots, n-1$$

が成立する.

(4) 式は [8] によって Laurent 級数展開と呼ばれたものである. 定常政策の中だけで考えるならば, -1 -割引最適である事と, 平均最適である事が, 同値となっている事が確かめられる.

§4. 政策改良法.

行列 A, B が同じ型である時, $A \succ B$ とは, 行列 $A-B$ の各行の 0 でない最初の項が正である事を示す. $A \succ B$ かつ $A \neq B$ ならば, $A \succ B$ と表わす.

任意の $f \in F$ に対して、 $Y_n(f) = (y_{-1}(f), y_0(f), \dots, y_n(f))$ とする。 $\mathcal{D}_{-2} = F$, $n = -1, 0, 1, \dots$ については $\mathcal{D}_n$ を、すべての $g \in F$ に対して $Y_n(f) \succeq Y_n(g)$ が成立するような $f \in F$ の全体として定義する。明らかに $\{\mathcal{D}_n\}$ は単調減少な集合の列となっている。また、 $Y_n(f) \succeq Y_n(g)$ ならば、 $\lim_{p \rightarrow 0+} \frac{1}{p^n} (V_p(f) - V_p(g)) \geq 0$ が成立している。

任意の $f, g \in F$ に対して、

$$\psi_n(gf) = \begin{cases} P(g)y_{-1}(f) - y_{-1}(f) & n = -1 \text{ の時,} \\ r(g) + P(g)y_0(f) - y_{-1}(f) - y_0(f) & n = 0 \text{ の時,} \\ P(g)y_n(f) - y_n(f) - y_{n-1}(f) & n = 1, 2, \dots \text{の時} \end{cases}$$

と定義する。 $n = -1, 0, \dots$ について、 $\Psi_n(gf) = (\psi_{-1}(gf), \psi_0(gf), \dots, \psi_n(gf))$, $G_n(f) = \{g \in F; \Psi_n(gf) \succ 0\}$ と定義する。

次の定理 3 は、 n -割引最適系における Howard [6] のタイプの政策改良法である。政策改良法の目的は $\mathcal{D}_n$ の元を求める事にあるが、定理 2 はその判定基準を与えている。定理 2, 3 は Veinott [10] の拡張である。その証明は [10] と同じ方針で行われ、2, 3 の補助定理が必要となるので省略する事にする。

定理 2. 条件 I のもとで、任意の $f \in F$, $n = -1, 0, 1, \dots$ に対して、

(a) $G_{n+1}(f) = \emptyset$ ならば $f \in \mathcal{D}_n$ である.

(b) $f \in \mathcal{D}_n$ ならば $G_n(f) = \emptyset$ である.

定理 3. 条件 I のもとで、任意の $f \in \mathcal{D}_{n-1}$ を取る. $G_n(f)$ の i 成分を $G_n(f)_i$ ($\subset A_i$) と表わし.

$$g(i) = \begin{cases} a \in G_n(f)_i \text{ なる任意の } a, & G_n(f)_i \neq \emptyset \text{ の時,} \\ f(i) & G_n(f)_i = \emptyset \text{ の時,} \end{cases}$$

として $g \in F$ を定義する. $g \neq f$ ならば、 $Y_n(g) \succ Y_n(f)$ が成立する.

§ 5. 存在定理.

F 上に A_i の直積位相を導入する. 即ち、 $\{f_n\} \subset F$ がある $f \in F$ に収束するとは、任意の $i \in S$ に対して自然数 N が存在して、 $n \geq N$ ならば $f_n(i) = f(i)$ が成立する事である. この事を $\lim_{n \rightarrow \infty} f_n = f$ と記す.

$Y_n(f)$ の F 上での連続性を得るために次の条件を設ける.

条件 II. 条件 I が成立していて、更に (2) における B_f に関して $\sup_{f \in F} \{B_f\} < \infty$ が成立する.

状態空間 S が有限集合ならば、条件 II は満たされる. 条件 II の意味に関しては、[7] に若干の記述がある.

補助定理 3. 条件 II のもとで、各 $n = -1, 0, \dots$ に対して、

$\lim_{k \rightarrow \infty} f_k = f$ ならば $\lim_{k \rightarrow \infty} Y_n(f_k) = Y_n(f)$ が成立する.

証明. 補助定理1 (a) とノルム $\|\cdot\|$ の性質により、任意の $g \in F$ について $\|y_\ell(g)\| = \|H(g)^{\ell+1} r(g)\| \leq 2^{\ell+2} \|H(g)\|^{\ell+1} \|r(g)\|$ が成立する. 従って条件IIにより $\{\|y_\ell(f_k)\|; \ell = -1, 0, \dots, n+1, k=1, 2, \dots\}$ は有界である. $\{y_{-1}(f_k), y_0(f_k), \dots, y_{n+1}(f_k)\}$ の $k \rightarrow \infty$ とした時の集積点の1つを $\{x_{-1}, x_0, \dots, x_{n+1}\}$ とする. 各 f_k に関して方程式(5)を考え、 $\{x_\ell\}$ を与えた部分列に関する極限をとると、 $\{x_\ell\}$ は f に対する(5)の解となる. 解の一意性により、 $x_\ell = y_\ell(f)$, $\ell = -1, 0, \dots, n$ である. $\square$

Blackwell [2] の存在定理により、 $\{p_k\} \searrow 0$ に対して p_k -割引最適定常政策の列 $\{f_k\}$ が定まる. F はコンパクトだから、 $\{f_k\}$ が収束するように部分列がとれる. それを改めて $\{f_k\}$ と表わす. 即ち、 f_k は p_k -割引最適で $\lim_{k \rightarrow \infty} f_k = f_*$ となる $\{f_k\}$, $f_* \in F$ が存在する. f_* を p -割引最適政策の極限と呼ぶ事にする.

補助定理4. 条件IIを満たすマルコフ決定過程 $\mathcal{M}$ が与えられているとする. ある n について $Y_n(f) = Y_n$ (一定) がすべての $f \in F$ に対して成立していれば、 p -割引最適政策の極限 f_* は $\mathcal{D}_{n+2}$ の元となる.

証明は $n = -2$ の場合には [7] にある. その他の n についても同様の方針で示されるので、省略する.

補助定理5. 条件IIのもとで $f \in \mathcal{D}_n$ ならば、 f は n -割

引最適定常政策である。

略証. 帰納法による. $n=-1, 0$ の時には補助定理4により明らかである. $f \in \mathcal{D}_{n-1}$ の時も仮定して, $f \in \mathcal{D}_n$ の時を示す.
 (1) で $\liminf$ を与え P -割引最適政策の極限 f_* に収束する $\{(p_k, f_k)\}$ がとれる. $f_k \in \mathcal{D}_n$ となる k が有限個しかない時にはそれらの k も除いて, 可算個ある時にはそのような k だけを採用して, 改めて $\{(p_k, f_k)\}$ を作る. $V_{p_k}(f) - V_{p_k}(\pi)$ を $V_{p_k}(f) - V_{p_k}(f_k)$ と $V_{p_k}(f_k) - V_{p_k}(\pi)$ の和で表わし, p_k で割って k に関する $\liminf$ をとる. 各々の場合に第1項が非負となる事は, それぞれ帰納法の仮定, Laurent展開により明らかである. よって $\pi^* = f$ として (1) が成立する. $\square$

与えられたマルコフ決定過程を $\mathcal{M}_2$ と表わす. 任意の $f_0 \in \mathcal{D}_n$ を1つ固定して, $F_n = \{g \in F; \psi_n(gf_0) = 0\}$ と定義する. $f_0 \in F_n$ だから $F_n \neq \emptyset$ である. 任意の $n=-1, 0, \dots$ に対してマルコフ決定過程 $\mathcal{M}_n = (S, F_n, \{P(f)\}_{f \in F_n}, \{r(f)\}_{f \in F_n})$ を, $\mathcal{M}_2$ の F_n 上への制限として定義する. $\mathcal{D}_n \neq \emptyset$ である限り $\mathcal{M}_n$ がうまく定義される事は, $\psi_n(gf)$ の定義により明らかである.

定理4. 条件IIが成立していて, $\mathcal{D}_n \neq \emptyset$ と仮定する. $\mathcal{M}_2$ から $\mathcal{M}_n$ を構成し, $\mathcal{M}_n$ において P -割引最適政策の極限 f_* をとると, もとの $\mathcal{M}_2$ に関して $f_* \in \mathcal{D}_{n+1}$ が成立する.

証明. $\psi_n(gf_0)$ の性質により, $D_{n+1} \supset F_n \supset D_n$ が成立する. 従って $\mathcal{M}_n$ に補助定理4が適用できる. 即ち, 任意の $f \in F$ が $f \in F_n$ を満たすならば, $Y_{n+1}(f_*) \geq Y_{n+1}(f)$ である. $f \notin F_n$ ならば $f \notin D_n$ だから, $Y_{n+1}(f_*) \geq Y_{n+1}(f_0) \geq Y_{n+1}(f)$ である. ゆえに $f_* \in D_{n+1}$. $\square$

定理4は $D_n \neq \emptyset$ ならば $D_{n+1} \neq \emptyset$ を示している. 各 D_n はコンパクトだから次の系1は明らかである. 補助定理5により系1は, ∞ -割引最適定常政策の存在を示している.

系1. 条件IIのもとで $\bigcap_{n=2}^{\infty} D_n \neq \emptyset$ が成立する.

References

- [1] Blackwell, D. Discrete dynamic programming. Ann. Math. Statist., 33, (1962), 719-726.
- [2] Blackwell, D. Discounted dynamic programming. Ann. Math. Statist., 36, (1965), 226-235.
- [3] Flynn, J. Conditions for the equivalence of optimality criteria in dynamic programming. Mathematics of operations Research, 2, (1976), 1-14.
- [4] Hordijk, A. Regenerative Markov decision models. Mathematical Programming Study, 6, (1976), 49-72.
- [5] Hordijk, A. and Sladky, K. Sensitive optimality criteria in countable state dynamic programming. Mathematics of Operations Research, 2, (1977), 1-14.
- [6] Howard, R. A. Dynamic Programming and Markov Processes.

Wiley, New York, (1960).

- [7] Kadota, Y. Countable state Markovian decision processes under the Doeblin conditions. To appear in Bull. Math. Statist., (1979).
- [8] Miller, B. L. and Veinott, A. F., Jr. Discrete dynamic programming with a small interest rate. Ann. Math. Statist., 40, (1969), 366-370.
- [9] Taylor, H. M. A Laurent series for the resolvent of a strongly continuous stochastic semi-group. Mathematical Programming Study, 6, (1976), 258-263.
- [10] Veinott, A. F., Jr. Discrete dynamic programming with sensitive discount optimality criteria. Ann. Math. Statist., 40, (1969), 1635-1660.
- [11] Wijngaard, J. Sensitive optimality in stationary Markovian decision problems on a general state space. Mathematical Centre Tracts, 93, (1977), 85-93.
- [12] Yosida, K. and Kakutani, S. Operator-theoretical treatment of Markoff's process and mean ergodic theorem. Ann. Math. 42, (1941), 188-228.