数理解析研究所講究録
第358巻 1979年 1-13

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．$F = \mathop{\times}_{i \in S} A_i$（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$\pi=(f_1, f_2, \cdots)$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$\pi=(f_1, f_2, \cdots)$に対してS上のベクトル$V_\rho(\pi)$, $x(\pi)$を，それぞれ

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

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

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．$V_\rho(\pi)$の第$i$要素は，システムが状態$i$から出発して政策$\pi$に従った時に得られる利得を，各時間毎に$\beta$だけ割引いたものの総和を表わしている．従って，$V_\rho(\pi)$を$\rho$-割引（または$\beta$-割引）総期待利得と呼ぶ．同様の意味で，$x(\pi)$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，$V_\rho(\pi^*)\geq V_\rho(\pi)$を満たしていれば，$\pi^*$を$\rho$-割引最適と呼び，$x(\pi^*)\geq x(\pi)$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$n=-1, 0, \cdots$に対して，

(1) $$\liminf_{\rho\to 0+}\rho^{-n}\{V_\rho(\pi^*)-V_\rho(\pi)\}\geq 0$$

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 $\rho_0 = \rho_0(f)$ が存在して、$0 < \rho < \rho_0$ ならば $V_\rho(f^*) \geq V_\rho(f)$ 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 ρ に対して ρ-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を次に記す.

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、次の表記を約束する. 任意の $k = 1, 2, \cdots$ について、$p_f^{(1)}(i, j) = p_f(i, j)$, $p_f^{(k)}(i, j) = \sum_{\ell \in S} p_f^{(1)}(i, \ell)\, p_f^{(k-1)}(\ell, j)$ 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 $p_f^*(i, j) = \lim_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} p_f^{(k)}(i, j)$ 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、$p_f^*(i, j)$ を (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、

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

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，$x(f) = P(f)^* r(f)$ が成立する．

§3. $V_\rho(f)$ 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 の表記を省略して，$P(f)$, p_f, $r(f)$, $x(f)$, $V_\rho(f)$ 等を P, p, r, x, V_ρ 等と表わす．

$S \times S$ の行列 H_ρ と S 上のベクトル h_ρ を，

$$H_\rho = \sum_{n=0}^{\infty} \beta^{n+1}(P^n - P^*), \qquad h_\rho = H_\rho r$$

と定義する．V_ρ の Laurent 級数展開を導くために，まず $\lim_{\rho\to 0+} H_\rho$，$\lim_{\rho\to 0+} h_\rho$ の存在とその性質について調べる．

<u>補助定理 1</u>．条件 I のもとで，

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

(b) $\lim_{\rho\to 0+} \| h_\rho - h \| = 0$ を満たすベクトル h が存在する．

(c) $H = \lim_{\rho\to 0+} H_\rho$ が存在する．任意の $i \in S$，$E \subset S$ に対して，$H_\rho(i, E) = \sum_{j\in E} (H_\rho)_{ij}$ と表わせば，$H(i, E) = \lim_{\rho\to 0+} H_\rho($

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_ρ で $\rho\to 0+$ とした時の集積点の1つを h_0 とする．$0\leqq\rho<\infty$ なる任意の ρ に対して，h_ρ は

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

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

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

が成立する．(a) と条件Iにより、$\lim_{\rho\to 0+}\|h_\rho-h_0\|=0$ が容易に示される． □

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．

<u>補助定理2</u>．$0<\rho<\infty$ とする．条件Iのもとで，

(a) $M_\rho=\sum_{n=0}^{\infty}\beta^{n+1}P^n$ とおくと，$P^*M_\rho=\frac{1}{\rho}P^*$ かつ $M_\rho=P^*M_\rho+H_\rho$ が成立する．

(b) $L_\rho=\sum_{n=0}^{\infty}\rho^n(-1)^nH^n$ とおくと，$(I+\rho H)L_\rho=L_\rho(I+\rho H)=I$ および $H_\rho=L_\rho H=HL_\rho$ が成立する．

<u>定理1.</u> 条件Iが成立するとする．

(a) $0<\rho<\dfrac{1}{2\|H\|}$ なるρをとると、

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

が成立する．但し、$y_{-1}=x$，$n=0,1,\cdots$ 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、$\|H\|=\sup_{i\in S}\sup_{E\subset S}|H(i,E)|$ で定義する．

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

(5) $$y_{-1}-Py_{-1}=0,\quad y_{-1}+y_0-Py_0=r,$$
$$y_{k-1}+y_k-Py_k=0,\quad k=1,2,\cdots,n$$

を満たす．逆に $n\geqq 0$ に対して、有界なベクトルの列 $(z_{-1},z_0,\cdots,z_n)$ が上記方程式の解になっていれば、

$$z_k=y_k,\quad k=-1,0,\cdots,n-1$$

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，$A\succsim B$ 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．$A\succsim B$ 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、$Y_n(f)=(y_{-1}(f), y_0(f), \cdots, y_n(f))$ とする. $\mathcal{D}_{-2}=F$, $n=-1,0,1,\cdots$ については $\mathcal{D}_n$ を、すべての $g\in F$ に対して $Y_n(f)\succsim Y_n(g)$ 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、$Y_n(f)\succsim Y_n(g)$ ならば、$\lim_{\rho\to 0+}\frac{1}{\rho^n}(V_\rho(f)-V_\rho(g))\geqq 0$ が成立している.

任意の $f, g\in F$ に対して,

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

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

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

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

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

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

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

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 $\lim_{n\to\infty} f_n=f$ 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.

<u>条件II.</u> 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.

<u>補助定理3.</u> 条件IIのもとで、各 $n=-1,0,\cdots$ に対して、$\lim_{k\to\infty} f_k=f$ ならば $\lim_{k\to\infty} y_n(f_k)=y_n(f)$ が成立する.

証明. 補助定理1（a）とノルム‖ ‖の性質により、任意の$g \in F$について $\| y_\ell(g) \| = \| H(g)^{\ell+1} r(g) \| \leqq 2^{\ell+2} \| H(g) \|^{\ell+1} \| r(g) \|$ が成立する. 従って条件IIにより$\{ \| y_\ell(f_k) \| ; \ell=-1,0,\cdots,n+1, k=1,2,\cdots \}$は有界である. $\{ y_{-1}(f_k), y_0(f_k), \cdots, y_{n+1}(f_k) \}$の$k \to \infty$とした時の集積点の1つを$\{ z_{-1}, z_0, \cdots, z_{n+1} \}$とする. 各$f_k$に関して方程式（5）を考え、$\{ z_\ell \}$を与えた部分列に関する極限をとると、$\{ z_\ell \}$は$f$に対する（5）の解となる. 解の一意性により、$z_\ell = y_\ell(f)$, $\ell=-1,0,\cdots,n$ である. □

Blackwell [2] の存在定理により、$\{ \rho_k \} \searrow 0$ に対してρ_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ρ_k-割引最適で $\lim_{k \to \infty} f_k = f_*$ となる$\{ f_k \}$, $f_* \in F$ が存在する. f_*をρ-割引最適政策の極限と呼ぶ事にする.

<u>補助定理4.</u> 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、ρ-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.

<u>補助定理5.</u> 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. $f\in\mathcal{D}_{n-1}$ の時を仮定して, $f\in\mathcal{D}_n$ の時を示す. (1)で liminf を与え ρ-割引最適政策の極限 f_* に収束する $\{(\rho_k, f_k)\}$ 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 を除いて, 可算個ある時にはそのような k だけを採用して, 改めて $\{(\rho_k, f_k)\}$ を作る. $V_{\rho_k}(f)-V_{\rho_k}(\pi)$ を $V_{\rho_k}(f)-V_{\rho_k}(f_k)$ と $V_{\rho_k}(f_k)-V_{\rho_k}(\pi)$ の和で表わし, ρ_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. □

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 $F_n\neq\phi$ である. 任意の $n=-1,0,\cdots$ 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. $\mathcal{D}_n\neq\phi$ 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.

<u>定理4.</u> 条件IIが成立していて, $\mathcal{D}_n\neq\phi$ と仮定する. $\mathcal{M}_{-2}$ から $\mathcal{M}_n$ を構成し, $\mathcal{M}_n$ において ρ-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、$\mathcal{D}_{n-1} \supset F_n \supset \mathcal{D}_n$ 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、$Y_{n+1}(f_*) \succsim Y_{n+1}(f)$ である. $f \notin F_n$ ならば $f \notin \mathcal{D}_n$ だから、$Y_{n+1}(f_*) \succsim Y_{n+1}(f_0) \succsim Y_{n+1}(f)$ である. ゆえに $f_* \in \mathcal{D}_{n+1}$. □

定理4は $\mathcal{D}_n \neq \phi$ ならば $\mathcal{D}_{n+1} \neq \phi$ を示している. 各 $\mathcal{D}_n$ 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.

<u>系1.</u> 条件Ⅱのもとで $\bigcap_{n=-2}^{\infty} \mathcal{D}_n \neq \phi$ 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 dynammic 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 stocastic 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] Wijngaad, 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.