数理解析研究所講究録
第557巻 1985年 1-13

有限マルコフ決定過程における Overtaking最適基準について

和歌山大教育　門田 良信（Yoshinobu Kadota）

§0. 序

定常有限マルコフ決定過程（以下FMDPと略す）の最適化基準は有限段階に関するものと無限段階に関するものとの2通りがある。無限段階に関する基準には多数あって、歴史的に中心課題となってきたβ-割引最適基準、平均最適基準、あるいはそれらの欠点を補うものとしてのsensitive最適基準等の他にも様々なものが案出されているようである。元々最適化基準は、FMDP理論を現実に適用する場合に目的に合ったものを選び出して活用するものであるから、その種類の豊富さはこの理論が対象とする現実問題の広さを、一面において示していると言えよう。

それらの多様な最適化基準の中にあってOvertaking最適性が特にその興味を引く点は、次の3つにあると考える。第1に数理経済学の方面からGale[7]等により、その有用性が強調されている事。第2に、この基準が我々の直観的な「良さ」

に合致している事。次にR. Bellmanの最適性の原理との関係である。この原理が時刻の最終段階から逆登って行動(actions)を規定する事により全体に渡る最も良い政策を求めようとしているのに反して，この基準では時刻の流れと共に行動を規定しながら前者に劣らない政策を得ようとしている点である。しかしこの3点を裏面から眺めるならば，Overtaking最適基準はFMDPの分野では未だ十分な成果が得られていない事となり，表面的に明快であればある程，解析的には難しくなり，また欲ばりすぎの基準であるために厳しい条件が必要となってくる事になる。事実，Overtaking最適基準を満たす最適政策の存在を扱った論文は少なく，FMDPにおいては筆者の知る限り *Denardo and Rothblum* [5] くらいである。

この報告の目的は，Overtaking最適基準と他の幾つかの最適化基準を比較する事によって，[5]の得た定理の意味を検討する事にある。§1では，以下の議論のモデルとなるFMDPと政策，およびOvertaking，平均最適基準等の定義を与える。§2では次の§3, 4で使われるFMDPの基礎理論を簡単に紹介する。その多くは *Blackwell* [1], *Veinott* [12] による。§3では[5]によるOvertaking最適基準に関する結果を紹介する。§4では*Veinott*[12]のsensitive最適基準，Blackwell最適基準(仮称)，Overtaking最適基準との相互関係を見る。

これより[5]の条件が相当に厳しいものである事が解るであろう。§4では3つの反例を調べて，[5]の条件は決して厳しすぎるものではない事，およびOvertaking最適とFlynn[6]による強Overtaking最適との相異を見る。なお，使われる記号は[1]，[12]等に従うものとする。

§1. FMDP

空でない有限集合 $S=\{1,2,\cdots,S\}$ を状態空間と呼び，各 $i\in S$ に対して行動空間（action space）と呼ぶ空でない有限集合 A_i が定まっているとする。いま時刻 $n=1,2,3,\cdots$ で状態 $i\in S$ を観察し行動 $a\in A_i$ を取った時，利得 $r(i,a)$ が得られ，時刻 $n+1$ ではシステムは状態 $j\in S$ へ確率 $P(j|i,a)$ で移るものとする。ここで任意の i,a に対して $r(i,a)$ は実数値をとり，$P(j|i,a)$については，任意の $i,j\in S$，$a\in A_i$ に対して $P(j|i,a)\geq 0$ かつ $\sum_{j\in S} P(j|i,a)\leq 1$ を満たすとする。$F=\underset{i\in S}{\times} A_i$（直積）とする。任意の $f\in F$ の第 i 要素を $f_i\in A_i$ で表わせば，$P(j|i,a)$の定義により f に対して S 次元部分確率行列 $P(f)\equiv(P(j|i,f_i);\ i,j\in S)$ と S 次元実ベクトル $r(f)=(r(i,f_i);\ i\in S)$ が定まる。以上のように定義された4つの組 $\mathcal{M}=(S,F,P,r)$ をFMDP（Finite state markov Decision Process）と呼ぶ。

Fの元の列 $\pi=(f_1,f_2,\cdots)$ を政策(policy)と呼ぶ。特にすべてのnについて$f_n=f$ならば、πを定常政策と呼びfで表わす。従ってFは定常政策全体の集合をも示す事になる。任意の政策$\pi=(f_1,f_2,\cdots)$と$n=1,2,\cdots$, $0\leq\beta<1$に対してS次元ベクトル$V_n(\pi)$, $V_\beta(\pi)$を

$$V_n(\pi)=\sum_{k=1}^{n} P(f_1)P(f_2)\cdots P(f_{k-1})\,r(f_k),$$

$$V_\beta(\pi)=\sum_{k=1}^{\infty} P(f_1)P(f_2)\cdots P(f_{k-1})\,r(f_k)\,\beta^{k}$$

と定義する。ただし、$k=1$のとき行列積$P(f_1)P(f_2)\cdots P(f_{k-1})$$=I$(単位行列)とする。$V_n(\pi)$, $V_\beta(\pi)$の第i要素はそれぞれ、システムが時刻1で$i\in S$から出発した時に政策πを使って得られる、時刻nまでの期待利得の和および$n\to\infty$とした時のβ-割引総期待利得を表わしている。

$n=1,2,3,\cdots$に対して $V_n^*=\sup\{V_n(\pi)\,;\,\pi$はすべての政策$\}$とおく。ここで$\sup$は、後にでてくる$\liminf$, $\limsup$, $\lim$, $\max$, 等号$=$, 不等号$\geq$も全く同様に、ベクトルの各要素毎にとられるものとする。いま、ある政策π^*が強Overtaking最適であるとは、

$$\lim_{n\to\infty}[\,V_n(\pi^*)-V_n^*\,]=0 \qquad (1)$$

が成立する事を言い、Overtaking最適であるとは、任意の政策πに対して

$$\liminf_{n\to\infty}[\,V_n(\pi^*)-V_n(\pi)\,]\geq 0$$

を満たす事を言う。また π^* が平均最適であるとは，任意の π に対して

$$\liminf_{n\to\infty} \frac{1}{n}[\, V_n(\pi^*) - V_n(\pi)\,] \geq 0 \qquad (2)$$

が成立する事である。

$V_n(\pi^*) \leq V_n^*$ であるから，(1) は $\liminf_{n\to\infty}[\, V_n(\pi^*) - V_n^*\,] \geq 0$ と同じ意味である。(2) は Flynn (1976) の定義であり，通常使われる平均最適の定義 $\liminf_{n\to\infty} V_n(\pi^*)/n \geq \liminf_{n\to\infty} V_n(\pi)/n$ よりも若干厳しいものである。

明らかに，π^* が強 Overtaking 最適ならば Overtaking 最適であり，Overtaking 最適ならば平均最適である。

§2. 関数方程式.

n-割引最適基準における関数方程式を構成し，この方程式の解を与える定常政策の部分集合を定義する。それらは §3, §4 の定理に用いられる。また以下においてベクトル X の第 i 要素は，X_i, $[X]_i$ 等と表わす。

任意の $f \in F$ に対して $P(f)^* = \lim_{n\to\infty} \frac{1}{n+1}\sum_{k=0}^{n} P(f)^k$ が存在して

$$x(f) \equiv P(f)^* r(f) = \lim_{n\to\infty} V_n(f)/n$$

が成立する。更に $(I - P(f) + P(f)^*)$ は正則で

$$H(f) = (I - P(f) + P(f)^*)^{-1} - P(f)^*,$$

$$g_{-1}(f) = x(f), \quad g_n(f) = (-1)^n H(f)^{n+1} r(f), \quad n = 0, 1, 2, \cdots$$

とおく。$\beta=\frac{1}{1+\rho}$, $\rho>0$ とし, $V_\beta(f)$ を ρ の関数とみて $V_\rho(f)$ で表わせば,

$$V_\rho(f)=\sum_{n=-1}^{\infty} g_n(f)\rho^n \qquad (3)$$

が成立する。(3) を $V_\rho(f)$ の Laurent 展開と呼ぶ。

任意の $i\in S$, $a\in A_i$, $f\in F$ に対して, $r_0(i,a)=r(i,a)$, $n\neq 0$ ならば $r_n(i,a)=0$, $g_{-2}(f)=0$ (零ベクトル) として

$$\gamma_n(i,a,f)=r_n(i,a)+\sum_{j\in S}p(j|i,a)g_n(f)_j-g_n(f)_i-g_{n-1}(f)_i \qquad (4)$$

$n=-1,0,1,\cdots$ とおく。また, 時刻1では $g\in F$ を使いその後は定常政策 $f\in F$ を使って得られる β-割引総期待利得を $V_\rho(gf)$ で表わすと, (3) と $V_\rho(gf)=\beta(r(g)+\beta P(g)V_\rho(f)$ より, $[V_\rho(gf)-V_\rho(f)]_i=\sum_{n=-1}^{\infty}\gamma_n(i,g_i,f)$ が成立する。$\gamma_n(i,g_i,f)$ を第 i 要素とするベクトルを $\gamma_n(gf)$ で表わす。

S次元ベクトルの列 $\{g_n;\ n=-2,-1,0,1\cdots\}$ を帰納的に

$$g_{-2}=0,\quad g_{-1}=\max\{g_{-1}(f);\ f\in F\}$$

$$g_n=\max\{g_n(f);\ g_{-1}=g_{-1}(f),\ g_0=g_0(f),\cdots,g_{n-1}=g_{n-1}(f),f\in F\}$$

と定義する。[12] により すべての n について $g_n=g_n(f_0)$ となる $f_0\in F$ が一般に存在するから, この定義は可能である。(4) の右辺で $g_n(f)_j$, $g_{n-1}(f)_i$, $g_n(f)_i$ をそれぞれ $g_{n,j}$, $g_{n-1,i}$, $g_{n,i}$ でおきかえてできる式を, $\gamma_n(i,a)$ で表わし, $\gamma_n(i,f_i)$ を第 i 要素とするベクトルを $\gamma_n(f)$ で表わす事にする。

$$A_i(n)\equiv\{a\in A_i;\ \gamma_k(i,a)=0,\ k=-1,0,\cdots,n\}$$

かつ $F_n = \underset{i\in S}{X} A_i(n)$（直積）とすれば，$F_n \subset F$ は n に関する単調非増加列となり、前出の f_0 は $f_0 \in \bigcap_{n=-1}^{\infty} F_n$ を満たす。$\bigcap_{n=-1}^{\infty} F_n$ は後に現れる Blackwell 最適定常政策の集合となる。また n を固定するとき、任意の $i\in S$ と任意の $k=-1,0,\cdots,n+1$ に対して $\gamma_k(i,\alpha)=0$ としたものは n-割引最適基準における関数方程式を示す。特に $n=-1$ のときには平均最適基準のそれになる。

§3. Overtaking 最適基準に関する [5] の結果.

前 § で得られた $F_0 = \underset{i\in S}{X} A_i(0)$ を使って Overtaking 最適政策が存在するための 1 つの十分条件を記す。

条件 I. 任意の $f, g \in F_0$ に対して $P(f)^*_{ii} > 0$（$P(f)^*_{ii}$ は $P(f)^*$ の第 (i,i) 要素）ならば，$f_i = g_i$ である。

マルコフ連鎖 $P(f)$ の状態空間 S は，過渡的状態の集合 T と有限個の再帰的状態のクラス $R_1, R_2, \cdots, R_k$ に分割される。各 R_i の周期を d_i，$\{d_1, d_2, \cdots, d_k\}$ の最小公倍数を d とする。条件 I のもとでは、$i \in R_1 \cup R_2 \cup \cdots \cup R_k$ ならば $A_i(0)$ は唯一の行動を含む事になる。従って上記の $T, R_1, R_2, \cdots, R_k, d_1, d_2, \cdots, d_k$ および d は $f \in F_0$ に無関係に一意に定まる。

条件 II. $d = 1$（非周期的）.

定理 1. [5].

条件 I のもとで任意の定常政策 $f \in F_0$ と政策 π をとる。任意

の $i \in S$ について次の（1）～（3）は同値.

（1） $\limsup_{n\to\infty} \{ \sum_{\ell=0}^{d-1} [V^{n+\ell}(\pi) - V^{n+\ell}(f)]_i \} \geq 0$

（2） $n = 1, 2, \cdots$ に対して

$$0 = [P(f_1)P(f_2)\cdots P(f_{n-1}) \gamma_1(f_n)]_i = [P(f_1)P(f_2)\cdots P(f_{n-1}) \gamma_0(f_n)]_i .$$

（3） $\lim_{n\to\infty} \{ \sum_{\ell=0}^{d-1} [V^{n+\ell}(\pi) - V^{n+\ell}(f)]_i = 0$

系 2. [5].

条件 I, II のもとで、任意の定常政策 $f \in F_0$ は Overtaking 最適政策となる。政策 $\pi = (f_1, f_2, \cdots)$ が Overtaking 最適となるための必要十分条件は，定理 1 の（2）がすべての $i \in S$ について成立する事である。

§4. sensitive 最適基準との関係

ある政策 π^* について，π^* が Blackwell 最適であるとは，$\beta_0 < \beta < 1$ なる任意の β と任意の政策 π に対して $V_\beta(\pi^*) \geq V_\beta(\pi)$ が成立するような β_0 ($0 \leq \beta_0 < 1$) が存在する事を言う。

$n = -1, 0, 1, 2, \cdots$ について π^* が n-割引最適であるとは，任意の政策 π に対して $\liminf_{\rho\to 0} \rho^{-n} [V_\rho(\pi^*) - V_\rho(\pi)] \geq 0$ が成立する事を言う。

任意の政策 π と $k = 1, 2, 3, \cdots$ に対して $V_k^n(\pi)$ を

$$V_k^1(\pi) = V_k(\pi), \quad V_k^n(\pi) = \sum_{\ell=1}^{k} V_\ell^{n-1}(\pi)$$

として定義する。$n = -1, 0, 1, \cdots$ に対して政策 π^* が n-平均最適

であるとは、任意の政策 π に対して

$$\liminf_{k\to\infty} \frac{1}{k}\left[V_k^{n+2}(\pi^*) - V_k^{n+2}(\pi)\right] \geq 0$$

が成立する事を言う。特に-1-平均最適と§1の平均最適とは一致する。また0-平均最適の事をaverage overtaking最適とも言う。

[9]によれば $\pi^*=(f_1, f_2, \cdots)$ が n-平均最適となるための必要十分条件は、$k=1,2,3,\cdots$ について

$$P(f_1)P(f_2)\cdots P(f_{k-1})\gamma_m(f_k) = 0, \quad m=-1,0,1,\cdots,n \qquad (4)$$

かつ $$\lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} P(f_1)P(f_2)\cdots P(f_{k-1})\gamma_{n+1}(f_k) = 0 \qquad (5)$$

が成立する事である。特に $n=S$ のときには(5)は不必要となり、$n \geq S$ のときの(4),(5)は $n=S$ のときの(4)に一致する。

今までに紹介した幾つかの最適化基準を満たす政策の集合を、次の記号で表わす。

SOT ; 強Overtaking最適政策の集合。

OT ; Overtaking最適政策の集合。

B ; Blackwell最適政策の集合。

$\mathcal{D}_n$; n-割引最適政策の集合。$n=-1,0,1,\cdots$。

$\mathcal{O}_n$; n-平均最適政策の集合。$n=-1,0,1,\cdots$。

このとき[4, 5, 6, 8, 9, 10, 11, 12]の結果を総合すると、一般に次の包含関係を得る。

$$SOT \subset OT \subset B = \mathcal{D}_{s+m} = \cdots = \mathcal{D}_s \subset \mathcal{D}_{s-1} \subset \cdots \subset \mathcal{D}_1 \subset \mathcal{D}_0 \subset \mathcal{D}_{-1}$$
$$\| \qquad\qquad \| \quad \| \qquad\qquad \| \quad \| \quad \|$$
$$\mathcal{A}_{s+m} = \cdots = \mathcal{A}_s \subset \mathcal{A}_{s-1} \subset \cdots \subset \mathcal{A}_1 \subset \mathcal{A}_0 \subset \mathcal{A}_{-1}$$

ただし，$m=1,2,3,\cdots$ とする。

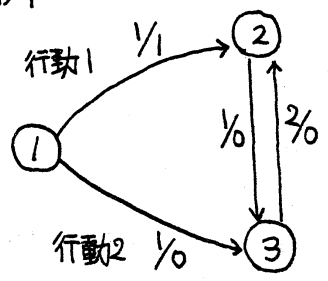
$\mathcal{D}_n \cap F \subset F_n \subset \mathcal{D}_{n-1} \cap F$ が成立する事から，条件Ⅰ，Ⅱのもとでは $OT \cap F = B \cap F = \mathcal{D}_0 \cap F = \mathcal{A}_0 \cap F = F_0 \subset \mathcal{D}_{-1}$ が成立する。また系2と(4)，(5)より，条件Ⅰ，Ⅱのもとで $\mathcal{A}_0 \subset OT \subset \mathcal{A}_{-1}$ が成立する。これらは条件Ⅰ，Ⅱの厳しさを物語っている。

一般に $B \cap F \neq \phi$ なる事が知られている。また $f \in \mathcal{D}_{n-1} \cap F$ を使って $f \in \mathcal{D}_n \cap F$ を見つける政策改良法は[12]によって得られている。線型計画による解法も[3]によって一部得られている。

§5．反例

例1 [4]．$d>1$ のとき $OT = \phi$ となる例

$S=\{1,2,3\}$，$A_1=\{1,2\}$，$A_2=A_3=\{1\}$，$d=2$．



状態1で行動1，行動2をとる定常政策を，それぞれ f，g とする。

確率/利得

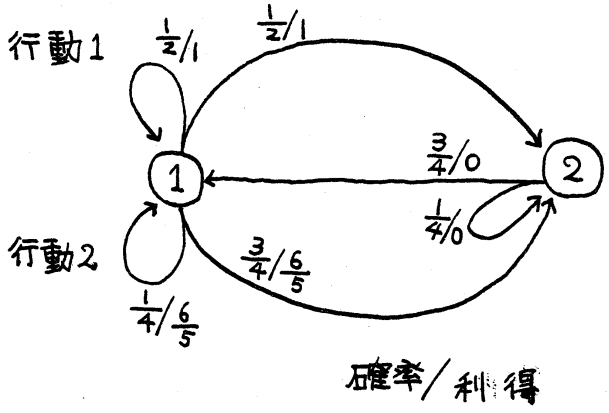
$$P(f)=\begin{pmatrix}0&1&0\\0&0&1\\0&1&0\end{pmatrix},\quad r(f)=\begin{pmatrix}1\\0\\2\end{pmatrix},\quad P(g)=\begin{pmatrix}0&0&1\\0&0&1\\0&1&0\end{pmatrix},\quad r(g)=\begin{pmatrix}0\\0\\2\end{pmatrix}.$$

n が奇数のとき $V_n(f)_1 = n$, $V_n(g)_1 = n-1$ であり n が偶数のとき $V_n(f)_1 = n-1$, $V_n(g)_1 = n$ である。従って、$OT = \phi$ である。実際には $f, g \in D_0$, $\{f\} = D_1 \cap F = B \cap F$ となっている。

例2 [2]. $d=1$ だけでは $OT = \phi$ となる例.

$S = \{1, 2\}$, $A_1 = \{1, 2\}$, $A_2 = \{1\}$.

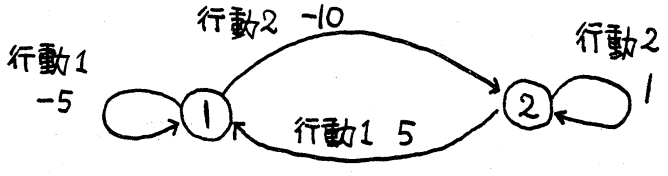
状態1で行動1, 行動2をとる決定を, それぞれ f, g とする。



$$P(f) = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{3}{4} & \frac{1}{4} \end{pmatrix}, \quad r(f) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad P(g) = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{3}{4} & \frac{1}{4} \end{pmatrix}, \quad r(g) = \begin{pmatrix} \frac{6}{5} \\ 0 \end{pmatrix}.$$

最適性の原理により, $V_n^* = V_n(\pi)$ となる F の元の列は (…f, g, f, g) (g で終り f と g が交互に現われる) となる。従って $\pi_1 = (f, g, f, g, \cdots)$, $\pi_2 = (g, f, g, f, \cdots)$ とすれば, n が奇数のとき $V_n(\pi_2) = V_n^* > V_n(\pi_1)$, n が偶数のとき $V_n(\pi_1) = V_n^* > V_n(\pi_2)$ となるから、$OT = \phi$ である。

例3 [6]. 条件Ⅰ, Ⅱを満たし $OT \neq \phi$ であるが $SOT = \phi$ となる例.



図の $-5, -10, 5, 1$ は利得を示し、矢印の推移確率はすべて1とする。

$S=\{1,2\}$, $A_1=A_2=\{1,2\}$. 状態1でも2でも常に行動2をとる定常政策をfとする。

$$V_1^*=\begin{pmatrix}-5\\5\end{pmatrix} \qquad n=2,3,4,\cdots \text{に対して} \quad V_n^*=\begin{pmatrix}-7+n\\4+n\end{pmatrix}$$

$$V_n(f)=\begin{pmatrix}-11+n\\n\end{pmatrix} \text{より}, \quad n=2,3,4,\cdots \text{に対して} \quad V_n(f)-V_n^*=\begin{pmatrix}-4\\-4\end{pmatrix}$$

となる。従ってfもSOTである。一方，平均最適定常政策はfだけであり，状態2の行動1は1度以上使うべきでないから，$SOT=\phi$，$OT\cap F=\{f\}$となる事が解る。

参考文献

[1] Blackwell, D. (1962). Discrete dynamic programming. Ann. Math. Statist. 33, 719-726.

[2] Brown, B. W. (1965). On the iterative method of dynamic programming on a finite space discrete time Markov process. Ann. Math. Statist. 36, 1279-1285.

[3] Denardo, E. V. (1970). Computing a bias-optimal policy in a discrete-time Markov decision problem. Operations Res. 18, 279-289.

[4] Denardo, E. V. and Miller, B. L. (1968). An optimality condition for discrete dynamic programming with no

discounting. Ann. Math. Statist. 39, 1220-1227.

[5] Denardo, E.V. and Rothblum, U.G. (1979). Overtaking optimality for markov decision chains. Math. Operations Res. 4, 144-152.

[6] Flynn, J. (1980). On optimality criteria for dynamic programs with long finite horizons. Jour. Math. Ann. Applications 76, 202-208.

[7] Gale, D. (1967). On optimal developement in a multi--sector economy. Review of Economic Studies 34, 1-18.

[8] Lippman, S.A. (1969). Criterion equivalence in discrete dynamic programming. Operations Res. 17, 920-923.

[9] Sladký, K. (1974). On the set of optimal controls for markov chains with rewards. Kybernetika 10, 350-367.

[10] Veinott, A.F. Jr. (1966). On finding optimal policies in discrete dynamic programming with no discounting. Ann. Math. Statist. 1284-1294.

[11] ———. (1968). Discrete dynamic programming with sensitive optimality criteria (preliminary report). Ann. math. Statist. 39, 1372.

[12] ———. (1969). Discrete dynamic programming with sensitive optimality criteria. Ann. Math. Statist. 40, 1635-1660.