数理解析研究所講究録
864巻 1994年 9-16

ベクトル値マルコフ決定過程と線形不等式系

長岡高専　涌田和芳（Kazuyoshi Wakuta）

1．序

割引利得基準をもつ有限状態，ベクトル値マルコフ決定過程を考える．1980年代の初め，ベクトル値マルコフ決定過程に対する関心が高まった（Furukawa(1980a), Henig(1983)）．そして，最適政策を求める有効なアルゴリズムがいくつか提案された(Viswanathan et al.(1977), Furukawa(1980b), White & Kim(1980), White(1982))．本報告では，最適な（確定的）定常政策の特徴付けを与え，それを利用した政策改良法を考える．ベクトル値マルコフ決定過程の政策改良法は，Furukawa(1980b)が最初に議論した．そこでは最適政策を，利得がベクトル値版の最適方程式の極大解であるものとして特徴付け，まず政策改良によって最適方程式の解を求め，その中の極大解を探した．一方，Thomas(1983)は，平均利得基準をもつベクトル値マルコフ決定過程の政策改良法を与えた．そこでは，(Furukawa(1980b)とは異なる)政策改良法によって最適政策の候補を集め，その中から最適政策を探した．Thomas(1983)の方法は割引利得基準にも応用できる．Thomas(1983)の政策改良は 2つのベクトルを比較するという単純なものであるが，最適政策の特徴付けはしていない．また，これらの論文では定常政策しか考慮されていない．定常政策の中では最適であるが，(確率的で，システムの履歴を記憶する政策を含む）すべての政策の中では最適でないものがある．したがって，すべての政策を考慮した方が良いが，上述の方法は適用できない．

最近著者は，完全エルゴード性が満たされる場合の平均利得基準をもつベクトル値マルコフ決定過程について，最適な定常政策の特徴付けとそれを利用した政策改良法を考えた(Wakuta(1993a,b))．本報告では，これを割引利得基準の場合にに応用する．上述の政策改良法では，政策反復が停止したときの政策が最適かどうかわからなくて良いが，すべての政策を考慮する場合，最適性の判定が必要となる．本報告ではすべての政策を考慮し，最適政策を線形不等式系で特徴付ける．それを解いて，政策反復が停止したときの政策の最適性を判定する．政策改良はThomas(1983)に従う．

2．ベクトル値マルコフ決定過程

$a=(a_1,\ldots,a_m),(b_1,\ldots,b_m)\in R^m$　に対して

$a\geqq b \Leftrightarrow a_k\geqq b_k, k=1,\ldots,m$

$a\geq b \Leftrightarrow a\geqq b,\ a\neq b$

$a>b \Leftrightarrow a_k>b_k, k=1,\ldots,m$

$U \subset R^m$ に対して
$e(U) = \{x \in U \mid x \leqq y \text{ for some } y \in U \text{ implies } x = y\}$

ベクトル値マルコフ決定過程(VMDP)

$S = \{1,2,\ldots,N\}$ ：状態空間
A ＝有限集合：行動空間，$A(i)$：$i \in S$ で実行可能な行動集合
$GrA = \{(i,a) \mid i \in S, a \in A(i)\}$
$p(j \mid i,a)$, $i,j \in S$, $a \in A(i)$ ：推移確率
$r(i,a) = (r^1(i,a),\ldots,r^m(i,a)) \in R^m$ ：利得関数
β $(0 \leqq \beta < 1)$ ：割引因子
Π：すべての政策の集合
F：$f(i) \in A(i)$, $i \in A(i)$ である S から A への写像 f の集合

$$I_\pi(i_1) = E_\pi[\sum_{n=1}^{\infty} \beta^{n-1} r(i_n, a_n) \mid i_1]$$

$$V(i_1) = \bigcup_{\pi \in \Pi} I_\pi(i_1)$$

$$F(i_1) = \bigcup_{f \in F} I_{f^\infty}(i_1)$$

すべての $i_1 \in S$ に対して，$I_{\pi^*}(i_1) \in e(V(i_1))$ であるとき，π^* はVMDPに対して最適であるという

3．最適政策の特徴付け
各 $i_1 \in S$ に対して $c(i_1) \in R^m$ を選び，

$$r^c(i_1, i_n, a_n) = \langle c(i_1), r(i_n, a_n) \rangle$$

を利得関数にもつ非定常動的計画NDP(c)を考える（利得関数以外はVMDPと同じ）．
NDP(c)における政策 π の利得は

$$J_\pi(i_1) = E_\pi[\sum_{n=1}^{\infty} \beta^{n-1} r^c(i_1, i_n, a_n) \mid i_1], \quad i_1 \in S$$

である．明らかに $J_\pi(i_1) = \langle c(i_1), I_\pi(i_1) \rangle$, $i_1 \in S$. すべての $\pi \in \Pi$ と $i_1 \in S$ に対して $J_{\pi^*}(i_1) \geqq J_\pi(i_1)$ であるとき，π^* はNDP(c)に対して最適であるという．特にc が一定ならば，NDP(c)は通常のマルコフ決定過程(MDP(c))となる．

命題3.1. 各 $i_1 \in S$ に対して，$V(i_1)$ は定常政策によって生ずる有限個の頂点により生成されるコンパクト凸集合である，i.e., $V(i_1) = co F(i_1)$, $i_1 \in S$.

命題3.2.(Yu(1985)) π^* があるc(i_1)>0, $i_1 \in S$ をもつNDP(c)に対して最適ならば，VMDPに対しても最適であり，逆も成り立つ．

命題3.3.VMDPに最適定常政策が存在する．

Bをすでに最適でないとわかっている定常政策の集合とし，F'=F−B とする．更に，$F'(i_1) = \bigcup_{f \in F'} \{I_{f^\infty}(i_1)\}$, $i_1 \in S$ とおく．

命題3.4.定常政策 $(f^*)^\infty$ がVMDPに対して最適ならば

$$I_{(f^*)^\infty}(i_1) \in e(\mathrm{co}F'(i_1)), i_1 \in S$$

であり，逆もなりたつ．

半空間 $H_c = \{x \in R \mid \langle c, x \rangle \leqq 0\}$, $c \in R^m$ を考える．

定理3.1.定常政策 $(f^*)^\infty$ がVMDPに対して最適ならば $f^* \in F'$ であり，各 $i_1 \in S$ に対して

$$I_{f^\infty}(i_1) - I_{(f^*)^\infty}(i_1) \in H_{c(i_1)}, f \in F'$$

なるc(i_1)>0 が存在する．また逆も成り立つ．

$$R_f(i) = E_{f^\infty}[\sum_{n=1}^{\infty} \beta^{n-1} r(i_n, a_n) \mid i_1 = i], i \in S$$

$$L_f(i,a) = r(i,a) + \beta \sum_{j \in S} p(j \mid i,a) R_f(j), (i,a) \in GrA$$

とおく．

定理3.2.(Wakuta(1992))(i) 定常政策 $(f^*)^\infty$ がVMDPに対して最適ならば，各 $i_1 \in S$ に対して，$p_{(f^*)^\infty}(i_1)\{i_n\} > 0$ なるすべての $(i_n, a_n) \in GrA$ に対して

$$L_{f^*}(i_n, a_n) - R_{f^*}(i_n) \in H_{c(i_1)}$$

が成り立つようなc(i_1)>0 が存在する．

(ii) 各 $i_1 \in S$ に対して，$p_\pi(i_1)\{i_n\} > 0$, $\pi \in \Pi$ なるすべての $(i_n, a_n) \in GrA$ に対して

$$L_{f^*}(i_n, a_n) - R_{f^*}(i_n) \in H_{c(i_1)}$$

が成り立つような $c(i_1) > 0$ が存在すれば，定常政策 $(f^*)^\infty$ はVMDPに対して最適である．

任意の $f \in F$ と $i_1 \in S$ に対して，$1 \times m$ 行ベクトル

$$d_f^g(i_1) = I_{g^\infty}(i_1) - I_{f^\infty}(i_1), \quad g \in F'$$

を定義する．各 $i_1 \in S$ に対して $d_f^g(i_1)$, $g \in F'$ を行ベクトルにもつ行列を $D_f(i_1)$ で表す．このとき，定理3.1は次のようになる．

定理3.3. 定常政策 $(f^*)^\infty$ がVMDPに対して最適ならば，$f^* \in F'$ で，各線形不等式系

$$(S_1): \begin{cases} x > 0 \\ D_{f^*}(1)x \leqq 0 \end{cases}, \ldots, (S_N): \begin{cases} x > 0 \\ D_{f^*}(N)x \leqq 0 \end{cases}$$

は解をもつ．また逆も成り立つ．

任意の $f \in F$ に対して，$1 \times m$ 行ベクトル

$$q_f(i,a) = r(i,a) + \beta \sum_{n=1}^{\infty} p(j \mid i,a) I_{f^\infty}(j) - I_{f^\infty}(i), \quad i \in S, a \in A(i)$$

を定義し，次のようにおく．

$Q_f(i)$: $q_f(i,a)$, $(i,a) \in GrA$ を行ベクトルにもつ行列

$\bar{Q}_f(i_1)$: $q_f(i_n, a_n)$, $(i_n, a_n) \in GrA$, $p_{f^\infty}(i_1)\{i_n\} > 0$ を行ベクトルにもつ行列

$\bar{\bar{Q}}_f(i_1)$: $q_f(i_n, a_n)$, $(i_n, a_n) \in GrA$, $p_\pi(i_1)\{i_n\} > 0$, $\pi \in \Pi$ を行ベクトルにもつ行列

このとき定理3.2 (i)(ii)は次のようになる.

定理3.4 (i) 定常政策 $(f^*)^\infty$ がVMDPに対して最適ならば, 各線形不等式系

$$(T_1): \begin{cases} x>0 \\ \overline{Q}_{f^*}(1)\, x \leqq 0 \end{cases}, \cdots, (T_N): \begin{cases} x>0 \\ \overline{Q}_{f^*}(N)\, x \leqq 0 \end{cases}$$

は解をもつ.

(ii) 各線形不等式系

$$(U_1): \begin{cases} x>0 \\ \overline{Q}_{f^*}(1) x \leqq 0 \end{cases}, \cdots, (U_N): \begin{cases} x>0 \\ \overline{Q}_{f^*}(N)\, x \leqq 0 \end{cases}$$

が解をもてば, 定常政策 $(f^*)^\infty$ はVMDPに対して最適である.

注意3.1 上述の線形不等式系は, フーリエ消去法により解をもつかどうか判定できる(cf. Stoer & Witzgall(1970)).

4.政策改良法

S上のすべてのm－値関数の集合を $B^m(S)$ で表す. $u, v \in B^m(S)$ に対して

$$u \geqq v \Leftrightarrow u(i) \geqq v(i),\ i \in S$$
$$u \geq v \Leftrightarrow u \geqq v,\ u \neq v$$

任意の $f \in F$ に対して

$$T_f u(i) = r(i, f(i)) + \beta \sum_{j \in S} p(j \mid i, f(i))\, u(j),\ i \in S$$

として $B^m(S)$ 上のオペレーターを定義する.

補題4.1(Furukawa(1980b)). T_f は単調である, i.e., $u \geqq v$ ならば $T_f u \geqq T_f v$ である.

$$I_f^g(i)=r(i,g(i))+\beta\sum_{j\in S}p(j\mid i,g(i))\,I_{f^\infty}(j),\ i\in S$$

とおく．特に $I_f^f(i)=I_{f^\infty}(i),\ i\in S$.

補題4.2.

(i) $I_f^g\geq I_f^f \Rightarrow I_{g^\infty}\geq I_{f^\infty}$

(ii) $I_f^g\leq I_f^f \Rightarrow I_{g^\infty}\leq I_{f^\infty}$

(iii) $I_f^g=I_f^g \Rightarrow I_{g^\infty}=I_{f^\infty}$

<u>すべての最適な定常政策を求めるためのアルゴリズム</u>

E_n：すでに最適と判定された定常政策の集合
F_n：すでに最適でないと判定された定常政策の集合
G_n：まだどちらとも判定されない定常政策の集合
とおく．

1. $E_0=F_0=G_0=\phi$ とおき，$(f_1)^\infty$を選ぶ．

2. $(f_n)^\infty$, $n\geqq 1$ に対して

$$v_n(i)=r(i,f_n(i))+\beta\sum_{j\in S}p(j\mid i,f_n(i))\,v_n(j),\ i\in S$$

を解いて，$v_n(i)=I_{(f_n)^\infty}(i),\ i\in S$ を求める．

3. $A_{f_n}=\{g\in F\mid I_{f_n}^g\geq I_{f_n}^{f_n}\}$

$B_{f_n}=\{g\in F\mid I_{f_n}^g\leq I_{f_n}^{f_n}\}$

$C_{f_n}=\{g\in F\mid I_{f_n}^g=I_{f_n}^{f_n}\}$

を求める．

4. $A_{f_n}\neq\phi$ なら

$E_n=E_{n-1}$, $F_n=F_{n-1}\cup B_{f_n}\cup C_{f_n}$, $G_n=G_{n-1}$

とおく．$f_{n+1}\in A_{f_n}$ を選択し，Step 2 へ戻る．

$A_{f_n}=\phi$ なら $(f_n)^\infty$ が最適かどうか判定する.

$(f_n)^\infty$ が最適なら

$E_n=E_n\cup C_{f_n}$, $F_n=F_{n-1}\cup B_{f_n}$, $G_n=G_{n-1}$

とおく.

$(f_n)^\infty$ が最適でないなら

$E_n=E_{n-1}$, $F_n=F_{n-1}\cup B_{f_n}\cup C_{f_n}$, $G_n=G_{n-1}$

とおく.

判定できないなら,

$E_n=E_{n-1}$, $F_n=F_{n-1}\cup B_{f_n}$, $G_n=G_{n-1}\cup C_{f_n}$

とおく.

$f_{n+1}\notin(E_n\cup F_n\cup G_n)$ を選び, Step 2 へ戻る.

5. $E_n\cup F_n\cup G_n=F$ となったら停止し, Step 6 へ行く.

6. $G_n=\phi$ ならば, E_n がすべての最適な定常政策の集合である. $G\neq\phi$ ならば, $F'=E_n\cup G_n$ とおき, G_n の各政策が最適かどうか判定する.

参考文献

N. Furukawa, Characterization of optimal policies in vector-valued Markov decision processes, Math. Oper. Res. 5(1980a)271-279.

————, Vector-valued Markovian decision processes with countable state space, Recent Developments in Markov Decision Processes, pp. 205-223, Ed. by R. Hartley et al., Academic Press, New York, 1980b.

M. Henig, Vector-valued dynamic programming, SIAM J. Control Optim. 21(1983) 490-499.

J. Stoer and C. Witzgall, Convexity and Optimization in Finite Dimensions I, Springer-Verlag, Berlin, 1970.

L. C. Thomas, Constrained Markov decision processes as multi-objective problems, Multi-Objective Decision Making, pp. 77-94, Ed. by S. Fench et al., Academic Press, New York, 1983.

B. Viswanathan et al., Multiple criteria Markov decision processes, TIMS Studies in Management Sciences 6(1977)263-272.

K. Wakuta, Optimal stationary policies in the vector-valued Markov decision processes, Stochastic Process. Appl. 42(1992)149-156.

———, RIMS Kokyuroku 835(1993a)144-152.

———, 日本数学会（春期）予稿集 (1993b)101-102.

C. C. White and K. W. Kim, Solution procedures for vector criterion Markov decision processes, Large Scale Systems 1(1980), 129-140.

D. J. White, Multiple-objective infinite-horizon discounted Markov decision processes, J. Math. Anal. Appl. 89(1982)639-647.

P. L. Yu, Multiple-Criteria Decision Making, Plenum Press, New York, 1985.