数理解析研究所講究録
1043巻 1998年 163-168

ベクトル値マルコフ決定過程における値空間の構造

長岡工業高等専門学校　涌田和芳（Kazuyoshi WAKUTA）

1. はじめに

ベクトル値マルコフ決定過程（VMDP）は，利得ベクトルと同じ次元の値空間を持ち，一般に多くの解が存在する．そのために，値空間の構造に関心が生ずる．値空間の形は？ 各初期状態に対する値空間の間の関係は？ 有効な解を得るために十分な政策のクラスは？ これらの問いは，スカラー値の場合には起こらなかった．すでに，各値空間は確定的定常政策の値により生成される凸多面体であることがわかっている．最初に，VMDPの最適政策の特徴付けのために値空間の間の関係について調べ，VMDPにおける最適性の連結定理（linking property of optimality)を証明する．(MDPの場合は、Schweitzer & Gavish(1976)が証明している）．このことから，各値空間が密接に関係していることがわかる．これと関連して，半定常（semi-stationary）政策という新しい政策のクラスについて述べる（cf. Wakuta (1996), Liu et al.(1996)）．次に，すべての解を記述するために確定的定常政策の確率化政策（randomization）の値の位置を調べる．一般に値空間の面（face）の任意の点は，その面を生成する確定的定常政策の確率化政策の値で表されることが期待される．それが正しいことは，容易に証明される．逆に，確定的定常政策の値が値空間の面を生成すれば，それらの確率化政策の値は同じ面上にあることが予想される．しかし，これは一般には成り立たないが，適当な確定的定常政策を選べば，与えられた面を生成し，かつそれらの政策のすべての確率化政策の値がその面に一致することがわかる．最後に，政策改良法で現れるマルコフ政策の値の位置について議論する．これと関連して，Feinberg & Shwartz (1996)の結果についても言及する．

2. ベクトル値マルコフ決定過程

ベクトル値マルコフ決定過程（VMDP）

$S=\{1,2,\ldots,N\}$：状態空間，$A(i)=$有限集合：行動空間

$p(j|i,a),\ i,j\in S, a\in A(i)$：推移確率

$r(i,a)=(r^1(i,a),\ldots,r^m(i,a))$：利得ベクトル

$\beta(0\leqq\beta<1)$：割引因子

Π：すべての政策の集合，Π_D：すべての確定的定常政策の集合

$$I_\pi(i_1)=E_\pi\left[\sum_{n=1}^{\infty}\beta^{n-1}r(i_n,a_n)|i_1\right],\ i_1\in S.$$

$$V(i_1)=\bigcup_{\pi\in\Pi}\{I_\pi(i_1)\},\ i_1\in S,\quad V_D(i_1)=\bigcup_{f^\infty\in\Pi_D}\{I_{f^\infty}(i_1)\},\ i_1\in S.$$

ここで，$V(i_1) = co\, V_D(i_1)$．ある $i_1 \in S$ に対して $I_{\pi^*}(i_1) \in e(V(i_1))$ であるとき，π^* は i_1－最適であるといい，すべての $i_1 \in S$ に対して i_1－最適ならば，最適であるという．ただし，$U \subset R^m$ に対して $e(U) = \{x \in U \mid$ ある $y \in U$ に対して $x \leqq y$ ならば $y = x\}$．

3. 値空間の間の関係

$c \in R^m$ に対して，$H_c = \{x \in R^m \mid < c, x > \leqq 0\}$ とおく．

[命題 3.1] f^∞ が i_1－最適であるための必要条件は，次のようなベクトル $c(i_1) > 0$ が存在することである：$p_{f^\infty}\{i_n \mid i_1\} > 0$ なるすべての $(i_n, a_n) \in GrA$ に対して

$$r(i_n, a_n) + \beta \sum_{j \in S} p(j \mid i_n, a_n) I_{f^\infty}(j) - I_{f^\infty}(i_n) \in H_{c(i_1)}$$

[命題 3.2] f^∞ が i_1－最適であるための十分条件は，次のようなベクトル $c(i_1) > 0$ が存在することである：ある π に対して $p_\pi\{i_n \mid i_1\} > 0$ であるすべての $(i_n, a_n) \in GrA$ に対して

$$r(i_n, a_n) + \beta \sum_{j \in S} p(j \mid i_n, a_n) I_{f^\infty}(j) - I_{f^\infty}(i_n) \in H_{c(i_1)}$$

これらの条件は，線形不等式系 (S_1) で記述できる．更にＬＰ問題 $P(S_1)$ に変換して，最適性の判定に使う．ここで，i－最適と j－最適の間の関係を調べてみる．通常のＭＤＰの場合は，Schweitzer-Gavish(1976)が議論している．

[定理 3.1] ある $i \in S$ に対して $I_{f^\infty}(i) \in e(V(i))$ で，ある $n \geqq 1$ に対して $p_{f^\infty}\{i_n = j \mid i\} > 0$ ならば，$I_{f^\infty}(j) \in e(V(j))$．

（証明）

$< c, I_{f^\infty}(i) > \geqq < c, I_\pi(i) >$ なる $c \in R^m, c > 0$ が存在する．$< c, r(i,a) >$ を利得とするＭＤＰを考えると，すべての $\pi \in \Pi$ に対して $J^c_{f^\infty}(i) \geqq J^c_\pi(i)$．任意の π に対して

$$\pi' = (f, \ldots, f, \sigma) \quad \text{ただし，} \sigma = \begin{cases} \pi & (i_1 = j) \\ f^\infty & otherwise \end{cases}$$

とおく．このとき，

$$J^c_{f^\infty}(i) - J^c_{\pi'}(i)$$

$$= \beta^{n-1} \sum_{k \in S} p_{f^\infty}\{i_n = k \mid i\} (J^c_{f^\infty}(k) - J^c_\sigma(k))$$

$$= \beta^{n-1} p_{f^\infty}\{i_n = j \mid i\} (J^c_{f^\infty}(j) - J^c_\pi(j)).$$

$$\therefore J^c_{f^\infty}(j) \geqq J^c_\pi(j)$$

f^∞ は，ＭＤＰで j－最適．故に $I_{f^\infty}(j) \in e(V(j))$．

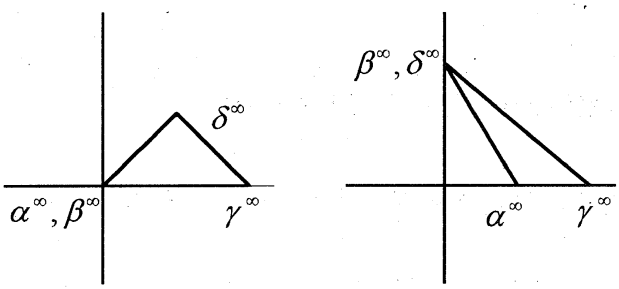
次の例を考える.

[例]

$S=\{1,2\}$, $A=A(1)=A(2)=\{1,2\}$

$p(1|1,1)=p(2|1,2)=p(1|2,1)=p(2|2,2)=1$

$r(1,1)=(0,0), r(1,2)=(1,0), r(2,1)=(1,0), r(2,2)=(0,1)$



一般に，各$i_1 \in S$に対して$V(i_1)$は異なる．また，各$i_1 \in S$に対して最適な端点を指定するとき，これに対応する最適な確定的定常政策は存在するとは限らない．このために，各$i_1 \in S$に対して，確定的定常政策を選択する政策の新しいクラスを導入する．例えば，$(\gamma^\infty, \delta^\infty)$：$i_1=1$なら$\gamma^\infty$，$i_1=2$なら$\delta^\infty$を選択する政策を考える．この政策は，$f(i_1,i_n) \in A(i_n)$なので，半定常(semi-stationary)政策と呼ぶことにする(Wakuta(1996))．Liu他(1996)は，この政策をsub-stochastic stationaryと呼んでいる．

4．確率化政策の値の位置

$\pi=(\pi_1,\pi_2,\ldots)$が$(f^l)^\infty, l=1,\ldots,k$ の確率化政策であるとは，π_n が状態$i \in S$において，確率$t^l_{n,i}$でf^lを選択することを意味する．$t^l_{n,i}=t^l_i$のとき，定常確率化政策という．

[定理 4.1] Fを$V(i)$の面，pをFの任意の点とする．このとき，Fを生成する任意の確定的定常政策は，pをそれらの確率化政策の値として表すことができる．

(証明)

Kallenberg(1983)のfrequency spaceの結果を使う．

$$x_{ja}[\pi]=\sum_i \alpha_i \sum_{t=1}^{\infty} \beta^{t-1} P_\pi\{i_t=j, a_t=a | i_1=i\}$$

$$x[\pi]=(x_{ja}[\pi])$$

$$K=\{x[\pi] | \pi \in \Pi\},\quad K(S)=\{x[\pi] | \pi \in \Pi_S\},\quad K(D)=\{x[\pi] | \pi \in \Pi_D\}$$

とおく．このとき

$$coK(D)=K(S)=K$$

$$I_\pi(i_1)=\sum_{j,a} x_{ja}[\pi](r^1(j,a),\ldots,r^m(j,a))$$

$$=(<x[\pi], r^1>,\ldots,<x[\pi], r^m>)$$

この関係を使って，Kの構造を$V(i_1)$の構造に写す．

$K=co\{y^1,\ldots,y^n\}$

$c^i=(<y^i,r^1>,\ldots,<y^i,r^m>)$

$V(i_1)=co\{c^1,\ldots,c^n\}$

$F=co\{c^{j_1},\ldots,c^{j_k}\}\leftrightarrow K'=co\{y^{j_1},\ldots,y^{j_k}\}$

$p\in F\leftrightarrow x\in K'$

$x=x[\pi^*](i_1)$

$p=I_{\pi^*}(i_1)$

π^*は，K'の端点に対応する確定的定常政策の確率化政策で表される．

逆に，Fを生成する任意の確定的定常政策の定常確率化政策の値は，そのF上の点とは限らない．

[反例]

$S=\{1,2\}$, $A(1)=A(2)=\{1,2\}$

$p(1|1,1)=1,\ p(2|1,1)=0, p(1|1,2)=0,\ p(2|1,2)=1$

$p(1|2,1)=1,\ p(2|2,1)=0, p(1|2,2)=0,\ p(2|2,2)=1$

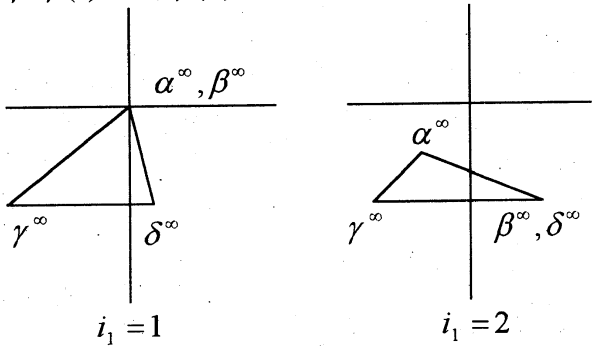
$r(1,1)=(0,0),\ r(1,2)=(-1,-1)$

$r(2,1)=(-1,-1), r(2,2)=(1,-1)$

$0<\beta<1$.

α: $\alpha(1)=1, \alpha(2)=1$; β: $\beta(1)=1, \beta(2)=2$

γ: $\gamma(1)=2, \gamma(2)=1$; δ: $\delta(1)=2, \delta(2)=2$



ただし，$\beta>1/2$

各$i\in S$に対して，確率1/2でα，確率1/2でδを選択する政策をπとする．

$I_\pi(1)=(-(2-\beta)/4(1-\beta),-(2+\beta)/4(1-\beta))$

$I_\pi(2)=(-\beta/4(1-\beta),-(4-\beta)/(1-\beta))$.

$I_\pi(1)$は，$I_{\alpha^\infty}(1)$と$I_{\delta^\infty}(1)$を結ぶ線分上にはない．

[条件 F]

(i) $(f^l)^\infty, l=1,\ldots,k$は，$F$を生成する；

(ii) $(f^l)^\infty, l=1,\ldots,k$は，ベクトル$c\in R^m$でスカラー化したMDPで最適で，
$h(v)=<c,v>$は，F上だけで$V(i_1)$の最大値をとる．

[**定理** 4.2] 与えられた$i_1 \in S$に対して，$V(i_1)$をFの任意の面とし，$(f^l)^\infty, l=1,...,k$は，条件Fを満たすとする．このとき$(f^l)^\infty, l=1,...,k$の任意の確率化政策の値は，F上にある．

（証明）

スカラー化してChitgopekar (1975)の結果を使う．「MDPで，最適な確定的定常政策の確率化政策は最適である」．

[**定理** 4.3] 与えられた$i_1 \in S$に対して，$V(i_1)$をFの任意の面とする．このとき，条件Fを満たす$(f^l)^\infty, l=1,...,k$が存在する．

（証明）

F上だけで$V(i_1)$の最大値をとる$h(v)=<c,v>$が存在する（例えば，Stoer-Witzgall(1970)）．次のような$h(v)=<c_n,v>$が存在する．

- $c_n \to c$
- Fの端点e^lだけで$V(i_1)$の最大値をとる．

e^lに対応する確定的定常政策$(f^l)^\infty$が存在する．$(f^l)^\infty$は，ベクトル$c \in R^m$でスカラー化したMDPでも最適である．

5．マルコフ政策(g^n, f^∞)の位置

VMDPの政策改良法で，一般に

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

が成り立つ．支配する政策が存在していても，政策改良が実行出来ないことがあることに注意する．しかし，$(g^n, f^\infty) \to I_{g^\infty}\ (n \to \infty)$が成り立つので，$(g^n, f^\infty)$の値の位置はどこかという問題が生ずる．なお，Feinberg-Schwartz(1996)は，制約付きマルコフ決定過程の研究で，VMDPにおいて，最適なm－randomized stationary policy（ある確定的定常政策の他に高々m個の行動しか使わない政策）と最適な(m,n)－policy（マルコフ政策で，ある確定的マルコフ政策の他に高々m個の行動しか使わず，かつ$n \geqq N$に対しては確定的定常政策を使う）の存在を示している．

参考文献

Chitgopekar S S (1975) Denumerable state Markovian sequential control processes: On randomizations of optimal policies. Naval Res Logistic Quart 22: 567-573

Feinberg E A, Shwartz A (1996) Constrained discounted dynamic programming. Math. Oper. Res. 21:922-945.

Kallenberg L C M (1983) Linear Programming and Finite Markovian Control Problems, Mathematical Centre Tracts No.148. Mathematisch Centrum, Amsterdam

Liu J, Huang S, Hu G (1996) On discounted Markov decision programming with multi-vector constraints. Chinese Sci Bull.41:202-207

Schweitzer P J, Gavish B (1976) An optimality principle for Markovian decision processes. J Math Anal Appl 54:173-184.

Wakuta K (1996) A new class of policies in vector-valued Markov decision processes. J Math Anal Appl 202: 623-628