数理解析研究所講究録
第798巻 1992年 171-179

ベクトル値マルコフ決定過程における最適定常政策

長岡高専　涌田和芳（Kazuyoshi Wakuta）

§1. 序

割引因子をもつベクトル値マルコフ決定過程において，最適定常政策の特徴付けを考える．Furukawa [1980] は，この問題について最初に議論し，次のような結果を発表している：ある連続性条件の仮定の下で，定常政策が定常政策全体の中で最適であるための必要十分条件は，その利得が最適方程式

$$u(i) \in e\left\{ \bigcup_{a \in A} \sum_{j} \left(r_{ij}^{a} + \beta u(j) \right) q_{ij}^{a} \right\}$$

の maximal な解となることである．一方，White [1982] は，どんな定常政策によっても支配されない非定常政策の存在を示している．ここでは，すべての (randomized, history-dependent な) 政策の中で最適定常政策を特徴付けることを考える．

§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）は，次のもので定義される．

S：空でないボレル集合，状態空間

A：　　　〃　　　，行動空間

F：行動制約集合関数 $F(s) \subset A$ $(s \in S)$，$G_r F = \{(s,a) \mid s \in S, a \in F(s)\}$ は，$S \times A$ のボレル集合で，S から A への1つのボレル可測写像のグラフを含むとする

$q \in Q(S \mid G_r F)$：システムの運動法則，ただし，$Q(Y \mid X)$ は X から Y への推移確率の全体を表わす

$r \in M^p(G_r F)$：利得関数，ただし，$M^p(X)$ は X から R^p $(p \geqq 1)$ への有界ボレル可測関数の全体を表わす

β $(0 \leqq \beta < 1)$：割引因子

政策 π の利得は，

$$I(\pi)(s_1) = E_\pi \left[\sum_{i=1}^{\infty} \beta^{i-1} r(s_i, a_i) \,\middle|\, s_1 \right], \quad s_1 \in S$$

で定義し，

$$V(s_1) = \bigcup_{\pi \in \Pi} \{ I(\pi)(s_1) \}, \quad s_1 \in S$$

とおく．ただし，Π は政策全体の集合を表わす．そして，

$e(V(s_1))$, $s_1 \in S$ を，nontrivial, closed convex cone K を domination cone とする $V(s_1)$, $s_1 \in S$, の properly efficient な元の集合とする．

定義　$I(\pi^*)(s_1) \in e(V(s_1))$, $s_1 \in S$, が成り立つとき，π^* は最適であるという．

§3. 最適定常政策

各 $s_1 \in S$ に対して $c(s_1) \in R^p$ を選び，利得関数 $r(s_1, s_n, a_n) = \langle c(s_1), r(s_n, a_n) \rangle$ をもつ非定常動的計画（$NDP(c)$）を考える．ただし，$\langle \cdot, \cdot \rangle$ は内積を表わす．その他は，VMDPと同じとする．$NDP(c)$ における政策 π の利得は，

$$J(\pi)(s_1) = E_\pi\left[\sum_{i=1}^{\infty} \beta^{i-1} r(s_1, s_n, a_n) \middle| s_1\right], \quad s_1 \in S$$

で定義し，$J(\pi^*)(s_1) \geqq J(\pi)(s_1)$, $s_1 \in S$, $\pi \in \Pi$ が成り立つとき π^* は最適であるという．

仮定　K の dual cone $K^* = \{ k^* \in R^p \mid \langle k^*, k \rangle \geqq 0,\ k \in K \}$ に対して，$\mathrm{int}\, K^* \neq \phi$ と仮定する．

命題 3.1　π^* がある $c(s_1) \in (\mathrm{int}\, K^*)$, $s_1 \in S$, 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．

証明　$V(s_1)$ は S に関して convex であること (cf. Schäl [1979]) を用いて，Benson [1979] の Theorem 4.2 を適用する□

定常政策 f^∞ に対して

$$R_f(s) = E_{f^\infty}\left[\sum_{i=1}^{\infty}\beta^{i-1} r(s_i, a_i) \mid s\right]$$

$$L_f(s,a) = r(s,a) + \beta\int_S R_f(s')\,dq(s' \mid s,a)$$

$$H_c = \{x \in R^p \mid \langle c, x\rangle \leqq 0\},\quad c \in R^p$$

$P_\pi(s_1)$：政策 π によって生ずる $A\times(S\times A)^\infty$ 上の確率測度

と定義する．

<u>定理 3.1</u>　定常政策 $(f^*)^\infty$ が VMDP に対して最適ならば，各 $s_1 \in S$ に対して，

$$L_{f^*}(s_n, a_n) - R_{f^*}(s_n) \in H_{c(s_1)},\quad P_{(f^*)^\infty}(s_1)-\text{a.s. } s_n, a_n \in F(s_n), n\geqq 1 \tag{3.1}$$

が成り立つような $c(s_1) \in (\text{int } K^*)$, $s_1 \in S$ が存在する．

証明　命題 3.1 より，$(f^*)^\infty$ が最適ならば $(f^*)^\infty$ はある $c(s_1) \in (\text{int } K^*)$, $s_1 \in S$ をもつ $\text{MDP}(c)$ に対して最適である．次の条件付期待値を考える．

$$J(\pi)(h_n) = E_\pi\left[\sum_{i=n}^{\infty}\beta^{i-n} r(s_1, s_i, a_i) \mid h_n\right]$$

ここで，$E_\pi[\cdot \mid h_n]$ は $\pi_n\, q\, \pi_{n+1}\, q \cdots$ による条件付期待値を

表わす（cf. Hinderer [1970], p.80 および Appendix 3）.

特に，$\pi = (f^*)^\infty$ ならば

$$J(\pi)(h_n) = J((f^*)^\infty)(s_1, s_n) = \langle c(s_1), R_{f^*}(s_n) \rangle \tag{3.2}$$

また，$\pi = \{f^*, \cdots, f^*, f_n, f^*, \cdots\}$ ならば

$$J((f^*)^\infty)(s_1, s_n) \geqq J(\pi)(h_n), \quad P_{(f^*)^\infty}(s_1)\text{-a.s. } h_n \tag{3.3}$$

が成り立ち，$a_n = f(s_n)$ とおいて右辺を書き改めると，

$$J((f^*)^\infty)(s_1, s_n) \geqq r(s_1, s_n, a_n) + \beta \int_S J((f^*)^\infty)(s_1, s_n)\, dq(s_{n+1} | s_n, a_n),$$
$$P_{(f^*)^\infty}(s_1)\text{-a.s. } s_n, a_n \in F(s_n) \tag{3.4}$$

(3.2)を使って（3.4）を書き改めると結果を得る□

<u>定理 3.2</u>　各 $s_1 \in S$ に対して

$$L_{f^*}(s_n, a_n) - R_{f^*}(s_n) \in H_{c(s_1)}, \quad P_\pi(s_1)\text{-a.s. } s_n, a_n \in F(s_n),\ n \geqq 1,\ \pi \in \Pi \tag{3.5}$$

が成り立つような $c(s_1) \in (\mathrm{int}\, K^*)$, $s_1 \in S$ が存在すれば、定常政策 $(f^*)^\infty$ は VMDP に対して最適である.

証明

$$r(s_1, s_n, a_n) = \langle c(s_1), r(s_n, a_n) \rangle$$

を利得関数にもつ NDP(c) を考える. $u(s_1, s_n) = J((f^*)^\infty)(s_1, s_n)$ とおくと，(3.5)より

$$u(s_1, s_n) \geqq r(s_1, s_n, a_n) + \beta \int_S u(s_1, s_{n+1})\, dq(s_{n+1} \mid s_n, a_n),$$

$$P_\pi(s_1) - \text{a.s. } s_n,\ a_n \in F(s_n),\ n \geqq 1,\ \pi \in \Pi \qquad (3.6)$$

一般に,条件付期待値の性質より

$$E_\pi\left[\ \sum_{i=1}^{N} \left\{ \beta^i u(s_1, s_{i+1}) - E_\pi\left[\beta^i u(s_1, s_{i+1}) \mid \bar{h}_i \right] \right\} \Big| s_1 \right] = 0 \qquad (3.7)$$

が成り立つ．ただし，$\bar{h}_i = (s_1, a_1, \cdots, s_i, a_i)$．ここで，(3.6)より

$$E_\pi\left[\beta^i u(s_1, s_{i+1}) \mid \bar{h}_i \right]$$

$$= \beta^i \int_S u(s_1, s_{i+1})\, dq(s_{i+1} \mid s_i, a_i)$$

$$= \beta^{i-1} \left\{ r(s_1, s_i, a_i) + \beta \int_S u(s_1, s_{i+1})\, dq(s_{i+1} \mid s_i, a_i) \right\} - \beta^{i-1} r(s_1, s_i, a_i)$$

$$\leqq \beta^{i-1} u(s_1, s_i) - \beta^{i-1} r(s_1, s_i, a_i).$$

これを (3.7)に代入して，$N \to \infty$ とすると，

$$J((f^*)^\infty)(s_1) = u(s_1, s_1) \geqq E_\pi\left[\sum_{i=1}^{\infty} \beta^{i-1} r(s_1, s_n, a_n) \Big| s_1 \right] = J(\pi)(s_1),$$

$$s_1 \in S.$$

π は任意なので，$(f^*)^\infty$ は NDP(C) に対して最適である．したがって，VMDPに対しても最適である □

§4. 例

次のようなVMDPを考える

$K = R_+^2 = \{(x, y) \mid x \geqq 0, y \geqq 0\}$

$S = \{1, 2, 3\}$, $A = \{1, 2, 3\}$, $F(1) = F(2) = \{1, 2\}$, $F(3) = \{1\}$

$q(\{1\} \mid 1, 1) = q(\{1\} \mid 1, 2) = 1$, $q(\{2\} \mid 2, 1) = q(\{3\} \mid 2, 2) = 1$,

$q(\{3\} \mid 3, 1) = 1$

$r(1, 1) = (2, 2)$, $r(1, 2) = (1, 4)$, $r(2, 1) = (2, 2)$, $r(2, 2) = (3, 1)$,

$r(3, 1) = (3, 1)$

このVMDPにおいて，定常政策 γ^∞: $\gamma(1) = \gamma(2) = \gamma(3) = 1$

は最適であることを定理3.2を用いて示す.

$I(\gamma^\infty)(1) = I(\gamma^\infty)(2) = (2/1-\beta, \ 2/1-\beta)$

$I(\gamma^\infty)(3) = (3/1-\beta, \ 1/1-\beta)$

各 $s_1 \in S$ に対して

$$D_{s_1} = \{ L_\gamma(s_n, a_n) - R_\gamma(s_n) \mid P_\pi(s_1)(s_n) > 0, a_n \in F(s_n), \pi \in \Pi \}$$

を求める.

$$D_1 : \begin{cases} r(1,1) + \beta I(\gamma^\infty)(1) - I(\gamma^\infty)(1) = (0, 0) \\ r(1,2) + \beta I(\gamma^\infty)(1) - I(\gamma^\infty)(1) = (-1, 2) \end{cases}$$

$$D_2 : \begin{cases} r(2,1) + \beta I(\gamma^\infty)(2) - I(\gamma^\infty)(2) = (0, 0) \\ r(2,2) + \beta I(\gamma^\infty)(2) - I(\gamma^\infty)(2) = (1, -1) \\ r(3,1) + \beta I(\gamma^\infty)(3) - I(\gamma^\infty)(3) = (0, 0) \end{cases}$$

$$D_3 : r(3,1) + \beta I(\gamma^\infty)(3) - I(\gamma^\infty)(3) = (0,0)$$

図1より，各 $s_1 \in S$ に対して (3.5) を満たすベクトル $c_{s_1} \in (\mathrm{int}\, K^*)$ が存在することがわかる．故に，γ^∞ は最適である．

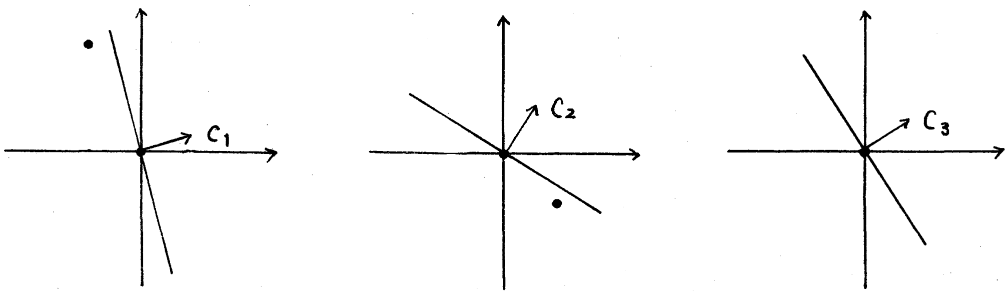

図 1

<u>注意</u> スカラー化という方法でVMDPを考えるとき，通常はベクトル $c(s_1)$ は初期状態には依存せず一定である．しかし、そのような方法では上述の例の定常政策 γ^∞ の最適性は判定されない．

参考文献

[1] H. P. Benson, An improved definition of proper efficiency for vector maximization with respect to cones, J. Math. Anal. Appl. 71 (1979), 232−241

[2] N. Furukawa, Characterization of optimal policies in vector-valued Markovian decision processes, Math. Oper. Res. 5 (1980), 271−279

[3] K. Hinderer, *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter* (Springer-Verlag, Berlin, 1970)

[4] M. Schäl, On dynamic programming and statistical decision theory, Ann. Probab. 7 (1979), 432-445

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