数理解析研究所講究録
第835巻 1993年 144-152

平均利得基準をもつベクトル値マルコフ決定過程：最適な確定的定常政策の特徴づけ

長岡高専　涌田和芳（Kazuyoshi Wakuta）

§1. 序

Thomas [7]は，完全エルゴーディックな場合の平均利得基準ベクトル値マルコフ決定過程について研究し，Furukawa[2]が割引利得基準について与えた政策改良法を修正して，すべてのパレート最適な政策を求めるためのアルゴリズムを提案した．Iki & Furukawa [5]は，政策の有効性（efficiency）は少し異なるが，多重連鎖の場合についての政策改良法を示した．これらの論文では，政策は確定的（deterministic）定常政策だけに限定されている．しかし，すべての確定的定常政策の中では最適であるが，（確率的定常政策を含む）すべての定常政策の中では最適ではない確定的定常政策が存在する可能性がある（§6を参照）．したがって，確定的定常政策だけに限定するのは得策でない．最近 Ghosh [3]はすべての定常政策を扱い，完全エルゴーディックな場合のパレート最適解の構

造を調べた.

本報告では, 完全エルゴード性を仮定し, すべての定常政策を考慮する. また最適性は凸錐で決定される有効性によって定義する. そして, スカラー化の手法を用いてすべての定常政策の中で最適な確定的定常政策を特徴づける.

§2. 準備

$K \subset R^p$: nontrivial, closed convex cone

$x \leqq_K y$ ($y \geqq_K x$) $\Leftrightarrow$ $y - x \in K$

$\forall U \subset R^p$ に対して

$e(U) = \{ x \in U \mid x \leqq_K y \text{ for some } y \in U \text{ implies } x = y \}$

$U^* = \{ u^* \in R^p \mid \langle u^*, u \rangle \geqq 0, \forall u \in U \}$

Assumption 2.1. $\operatorname{int} K^* \neq \emptyset$

§3. 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.

$S = \{1, 2, \cdots, N\}$: 状態空間

A : 有限行動空間, $A(i)$: 状態 $i \in S$ で実行可能な行動集合

$p(j \mid i, a)$, $i, j \in S$, $a \in A(i)$: 推移確率

$r(i, a) = (r^1(i, a), \cdots, r^p(i, a)) \in R^p$: 利得関数

Assumption 3.1. 任意の確定的定常政策によって生ずるマルコフ連鎖で、すべての状態は1つのエルゴード集合に属する.

Π はすべての定常政策の全体, Π_D はすべての確定的定常政策の全体を表す.

$\pi \in \Pi$ に対して,

$\phi_\pi(i_1) = \lim_{T\to\infty} \frac{1}{T} E_\pi\left[\sum_{t=1}^{T} r(i_t, a_t) \mid i_1\right]$: 政策 π の平均利得

$P_\pi = (p_{ij}(\pi))$, ただし, $p_{ij}(\pi) = \sum_{a\in A(i)} p(j|i,a)\,\pi(a|i)$

$P_\pi^* = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P_\pi^t$

μ_π : 不変確率測度

$r_\pi(i) = (r_\pi^1(i), \cdots, r_\pi^p(i))$, $i = 1, \cdots, N$

$r_\pi^k = (r_\pi^k(1), \cdots, r_\pi^k(N))'$, $k = 1, \cdots, p$

ただし, $r_\pi^k(i) = \sum_{a\in A(i)} r^k(i,a)\,\pi(a|i)$

このとき, $\phi_\pi(i_1) = \sum_{i\in S} \mu_\pi(i)\, r_\pi(i)$ (これを ϕ_π とかく)

$V = \bigcup_{\pi\in\Pi} \{\phi_\pi\}$

$f \in \Pi_D$ に対して (f^∞ は単に f と表す),

$D_f = (I - P_f + P_f^*)^{-1} - P_f^*$

$u_f^k = D_f r_f^k$, $k = 1, \cdots, p$

$u_f(i) = (u_f^1(i), \cdots, u_f^p(i))$, $i = 1, \cdots, N$ ただし, $u_f^k(i)$ は u_f^k の第 i 成分

π^* は VMDP に対して平均最適 $\Longleftrightarrow$ $\phi_{\pi^*} \in e(V)$

§4. 平均最適政策

$c \in R^p$ を選び，利得関数 $r^c(i,a) = \langle c, r(i,a) \rangle$ をもつスカラー値マルコフ決定過程（MDP(c)）を考える．

$J_\pi(i_1) = \lim_{T\to\infty} \frac{1}{T} E_\pi\left[\sum_{t=1}^{T} r^c(i_t, a_t) \mid i_1\right]$：政策 π の平均利得

このとき，$J_\pi(i_1) = \langle c, \phi_\pi \rangle$（これを J_π とかく）．$f \in \Pi_D$ に対して，

$w_f = D_f r_f^c$，ただし，$r_f^c = (r_f^c(1, f(1)), \cdots, r^c(N, f(N))'$．

π^* は MDP(c) に対して平均最適 $\Longleftrightarrow$ $J_{\pi^*} \geq J_\pi$，$\forall \pi \in \Pi$

Proposition 4.1 (cf. Ghosh [3]). V は，すべての確定的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凸集合である．

Proposition 4.2 (cf. Benson [1]). $\pi^* \in \Pi$ が VMDP に対して平均最適であるための必要十分条件は，それがあるベクトル $c \in (\text{int } K^*)$ をもつ MDP(c) に対して平均最適であることである．

Proposition 4.3. VMDP において平均最適な確定的定常政策が存在する．

Proposition 4.4 (cf. Veinott [8]). $f^* \in \Pi_D$ が MDP(c), $c \in R^p$ に対して平均最適であるための必要十分条件は, J_{f^*} と w_{f^*} が次の最適方程式を満たすことである:

$$J_{f^*} + w_{f^*}(i) = \max_{a \in A(i)} \{ r^c(i,a) + \sum_{j \in S} p(j|i,a) w_{f^*}(j) \}, \quad i \in S.$$

Theorem 4.1. $f^* \in \Pi_D$ が VMDP に対して平均最適であるための必要十分条件は, ある $c \in (\text{int } K^*)$ に対して

$$r(i,a) + \sum_{j \in S} p(j|i,a) u_{f^*}(j) - \phi_{f^*} - u_{f^*}(i) \in H_c, \quad i \in S, \ a \in A(i)$$

となることである. ここで, $H_c = \{ x \in R^p \mid \langle c, x \rangle \leqq 0 \}$.

§5. パレート最適

$K = R^p_+$ とする. $f \in \Pi_D$ に対して

$$g_f(i,a) = r(i,a) + \sum_{j \in S} p(j|i,a) u_f(j) - \phi_f - u_f(i), \quad i \in S, \ a \in A(i)$$

Q_f : $g_f(i,a)$, $i \in S$, $a \in A(i)$ を行ベクトルとする行列

Corollary 5.1. $f^* \in \Pi_D$ がパレートの意味で平均最適であるための必要十分条件は, $Q_{f^*} x \leqq 0$, $x > 0$ が解をもつことである.

Remark 5.1. 線形不等式系 (S): $x > 0$, $-Q_{f^*} x \geqq 0$ はフーリエ消去法 (cf. Stoer & Witzgall [6]) によって解をもつかどうか判定できる.

§6. 数値例

$K=R_+^2$, $S=\{1,2\}$, $A=\{1,2\}$, $A(1)=A(2)=A$

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

$p(1|1,2)=\frac{1}{2}$, $p(2|1,2)=\frac{1}{2}$

$p(1|2,1)=\frac{1}{4}$, $p(2|2,1)=\frac{3}{4}$

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

$r(1,1)=(-1,-4)$, $r(1,2)=(-7,4)$

$r(2,1)=(-1,1)$, $r(2,2)=(8,1)$

次の4つの確定的定常政策がある： f_{11} : $f_{11}(1)=1$, $f_{11}(2)=1$; f_{12} : $f_{12}(1)=1$, $f_{12}(2)=2$; f_{21} : $f_{21}(1)=2$, $f_{21}(2)=2$; f_{22} : $f_{22}(1)=2$, $f_{22}(2)=2$. $f=f_{11}$ を判定する.

<Theorem 4.1による判定>

最適方程式

$$\tilde{\phi}_f + \tilde{u}_f(i) = r(i,f(i)) + \sum_{j\in S} p(j|i,f(i))\,\tilde{u}_f(j),\quad i\in S$$

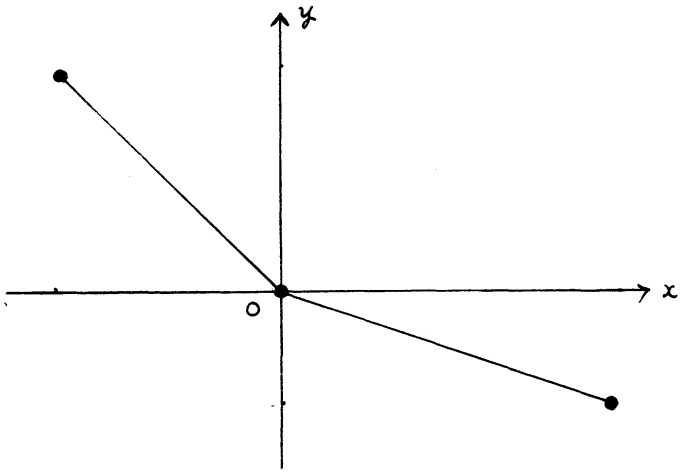
$$\tilde{u}_f(1)=(0,0)$$

を解いて $\tilde{\phi}_f$ と $\tilde{u}_f(i)$, $i\in S$ を求める. このとき, $\tilde{\phi}_f=\phi_f$ で $\tilde{u}_f^{k_0}(i)$ は $u_f^{k_0}(i)$ とは定数しか違わない (cf. Howard [4]). したがって, この解を用いても g_f は同じである.

$\phi_f=(-1,0)$, $\tilde{u}_f(1)=(0,0)$, $\tilde{u}_f(2)=(0,4)$

$g_f(1,1)=(0,0)$, $g_f(1,2)=(-6,6)$

$g_f(2,1)=(0,0)$, $g_f(2,2)=(9,-3)$



Th. 4.1 を満たす半空間は存在しない. しがって, f は最適ではない.

<Corollary 5.1 による判定>

$$(S): \begin{cases} Ix > 0 \\ -Qfx \geqq 0 \end{cases}, \quad Qf = \begin{pmatrix} 0 & 0 \\ -6 & 6 \\ 0 & 0 \\ 9 & -3 \end{pmatrix}$$

すなわち,

$$(S): \begin{cases} x_1 > 0 \\ x_2 > 0 \\ 6x_1 - 6x_2 \geqq 0 \\ -9x_1 + 3x_2 \geqq 0 \end{cases}$$

まずフーリエ消去法により x_2 を消去する.

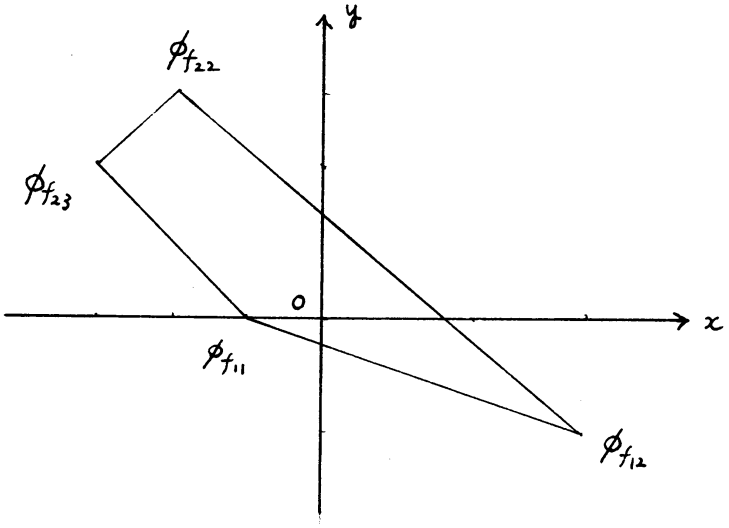
$$(S'): \begin{cases} x_1 > 0 \\ 6x_1 > 0 \\ -12x_1 \geqq 0 \end{cases}$$

更に x_1 を消去して

$$(S^2): \begin{cases} 0x_1 > 0 \\ 0x_1 > 0 \end{cases}$$

これは正しくない zero relation なので，(S) は解をもたない．故に，f は平均最適ではない．

次に政策の利得の集合 V について考える．4つの確定的定常政策の利得は，$\phi_{f_{11}} = (-1, 0)$，$\phi_{f_{12}} = (\frac{7}{2}, -\frac{3}{2})$，$\phi_{f_{22}} = (-3, 2)$，$\phi_{f_{22}} = (-2, 3)$ なので，すべての定常政策の利得の集合は下図のとおりである．



この図より f_{11} は確定的定常政策の中では平均最適であるが，すべての定常政策の中では平均最適ではない．

参考文献

[1] H. P. Benson, An improved definition of proper efficiency for vector maximization with respect to cones, JMAA 71 (1979), 232-242.

[2] N. Furukawa, Vector-valued Markov decision processes with countable state space, Recent Development in Markov Decision Processes, pp. 205-223, Ed. by D. J. White et al. Academic Press, 1980.

[3] M. K. Ghosh, Markov decision processes with multiple costs, OR Letter 9 (1990), 257-260.

[4] R. A. Howard, Dynamic Programming and Markov Processes, Wiley, 1960.

[5] T. Iki & N. Furukawa, Vector-valued Markov decision processes with average reward criterion, Mem. Fac. Edu. Miyazaki Univ. Nat. Sci. 54·55 (1984), 1-10.

[6] J. Stoer & C. Witzgall, Convexity and Optimization in Finite Dimensions I, Springer-Verlag, 1970.

[7] L. C. Thomas, Constrained Markov decision process as multi-objective problems, Multi-objective Decision Making, pp. 77-94, Ed. by D. J. White et al. Academic Press, 1983.

[8] A. F. Veinott, Jr. On finding optimal policies in discrete dynamic programming with no discounting, AMS 37 (1966), 1284-1294.