

平均利得基準をもつベクトル値マルコフ決定過程 ：多重連鎖の場合

長岡工業高等専門学校 涌田和芳 (Kazuyoshi WAKUTA)

1. はじめに

割引利得型ベクトル値マルコフ決定過程 (VMDP) については多くの文献がある. 一方, 平均利得型 VMDP の研究は少なく, 十分には行はれていない: Thomas [7] は, Furukawa [3] の方法を修正し, 完全エルゴード性な場合に最適な確定的定常政策を求める政策反復法を与えた. Iki & Furukawa [4] は異なった最適性 (bias optimality) のもとで, 多重連鎖過程の場合の政策反復法について議論した. Durinovic et al. [2] は, 多重連鎖の場合を多目的 LP 問題として定式化し, 最適な確定的定常政策を特徴づけた. Novák [4] は, 完全エルゴード性な場合を多目的 LP 法を用いて解いた.

最近著者は, 割引利得型および完全エルゴード性な場合の平均利得型 VMDP について, 最適な確定的定常政策を線形不等式系で特徴づけ, それに基づいた政策反復法を提案した [8] [9] [10]. 本論の目的は, 多重連鎖の場合の平均利得型 VMDP について同様なアプローチが可能であることを示すことである.

2. ベクトル値マルコフ決定過程

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

$$a \geq b \Leftrightarrow a_k \geq b_k, k = 1, \dots, m$$

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

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

$U \subset R^m$ に対して

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

ベクトル値マルコフ決定過程

$S = \{1, \dots, N\}$: 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

$GrA = \{(i, a) \mid i \in S, a \in A(i)\}$

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

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

政策 π の期待平均利得を $\phi_\pi(i_1) = (\phi_\pi^1(i_1), \dots, \phi_\pi^m(i_1))$ とする. ただし,

$$\phi_\pi^k(i_1) = \liminf_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \sum_{j,a} P_\pi \{i_t = i, a_t = a \mid i_1\} r^k(j, a).$$

$$x_{ja}^T[\pi](i_1) = \frac{1}{T} \sum_{t=1}^T P_\pi \{i_t = j, a_t = a \mid i_1\}, (j, a) \in Gr A$$

: T 期までの状態一行動の期待頻度

$X[\pi](i_1) : \{x^T[\pi](i_1), T = 1, 2, \dots\}$ の極限点 $x[\pi](i_1)$ の集合, とおく. このとき

$\phi_\pi(i_1) = (\phi_\pi^1(i_1), \dots, \phi_\pi^m(i_1))$ の各成分は,

$$\phi_\pi^k(i_1) = \liminf_{T \rightarrow \infty} \sum_{j,a} x_{ja}^T[\pi](i_1) r^k(j, a)$$

とかける.

Π : すべての政策の集合; $\pi = \{\pi_1, \pi_2, \dots\}$, $\pi_n = \pi_n(\cdot | h_n)$

$\Pi_1(i_1)$: $X[\pi](i_1)$ が一点だけからなる政策の集合

Π_D : すべての確定的定常政策の集合, $f: S \rightarrow A$ s.t. $f(i) \in A(i)$

とおく.

政策 $\pi \in \Pi_1(i_1)$ に対しては, $\phi_\pi(i_1) = \sum_{j,a} x_{ja}[\pi](i_1) r(j, a)$ である.

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

$$V_1(i_1) = \bigcup_{\pi \in \Pi} \{\phi_\pi(i_1)\}, \quad V_D(i_1) = \bigcup_{f \in \Pi_D} \{\phi_f(i_1)\},$$

$$W(i_1) = \bigcup_{\pi \in \Pi} \left\{ \sum_{j,a} x_{ja}[\pi](i_1) r(j, a) \mid x[\pi](i_1) \in X[\pi](i_1) \right\}$$

とおく.

定義 2.1. すべての $i_1 \in S$ に対して $\phi_{\pi^*}(i_1) \in e(V(i_1))$ であるとき, π^* は最適であるという.

命題 2.1. (Derman [1], Kallenberg [5]) $W(i_1) = V_1(i_1) = co V_D(i_1)$, $i_1 \in S$.

命題 2.2. $e(V(i_1)) = e(V_1(i_1))$, $i_1 \in S$.

3. 最適な確定的定常政策

$B^m(S)$ を S 上の m 値関数の集合とし, $r^c(i_1, i_n, a_n) = \langle c(i_1), r(i_n, a_n) \rangle$, $c \in B^m(S)$ を利得関数にもつ非定常動的計画 (NDP(c)) を考える. そして, 政策 π の期待平均利得を

$$J_\pi(i_1) = \sum_{j,a} x_{ja}[\pi](i_1) r^c(i_1, j, a), \quad \pi \in \Pi_1(i_1), i_1 \in S$$

とする.

定義 3.1. 各 $i_1 \in S$ に対して, $J_{\pi^*}(i_1) \geq J_\pi(i_1)$, $\pi \in \Pi_1(i_1)$ であるとき, π^* は NDP(c) で最適であるという.

命題 3.1 (Yu[11]) 政策 $\pi^* \in \bigcap_{i_1 \in S} \Pi_1(i_1)$ が最適ならば, ある $c \in B^m(S)$, $c > 0$ に対して π^* は NDP(c) で最適であり, 逆も成り立つ.

$$S(\pi, i_1) = \{j \in S \mid P_\pi\{i_t = j \mid i_1\} > 0 \text{ for some } t \geq 1\}, \quad \pi \in \Pi_1(i_1), i_1 \in S,$$

$$S(i_1) = \bigcup_{\pi \in \Pi_1(i_1)} S(\pi, i_1), \quad i_1 \in S \text{ とおく.}$$

定理 3.1. f^∞ が最適ならば, 各 $i_1 \in S$ について, 任意の $(i_t, a_t) \in \text{Gr } A$, $i_t \in S(f^\infty, i_1)$ に対して

$$\langle c(i_1), \phi_{f^\infty}(i_t) \rangle \geq \sum_{j=1}^N p(j|i_t, a_t) \langle c(i_1), \phi_{f^\infty}(j) \rangle \quad (3.1)$$

$$\begin{aligned} \langle c(i_1), \phi_{f^\infty}(i_t) \rangle + \langle c(i_1), u(i_t) \rangle \\ \geq \langle c(i_1), r(i_t, a_t) \rangle + \sum_{j=1}^N p(j|i_t, a_t) \langle c(i_1), u(j) \rangle \end{aligned} \quad (3.2)$$

なる $u \in B^m(S)$ と $c \in B^m(S)$, $c > 0$ が存在する.

定理 3.2. 各 $i_1 \in S$ について, 任意の $(i_t, a_t) \in \text{Gr } A$, $i_t \in S(f^\infty, i_1)$ に対して, (3.1) (3.2) が成り立てば, f^∞ は最適である.

系 3.1. 各 $i_1 \in S$ について, 任意の $(i_t, a_t) \in \text{Gr } A$ に対して, (3.1) (3.2) が成り立てば, f^∞ は最適である.

4. 線形不等式系による最適性の特徴付け

$i_1 \in S$ を固定し, $x_k = c^k(i_1)$, $k=1, 2, \dots, m$, とおき, (i_t, a_t) を (i, a) で置き換えると定理 3.1 の条件は, 次のようになる.

$$(S_{i_1}) : \begin{cases} \sum_{k=1}^m \phi_{f^\infty}^k(i) x_k \geq \sum_{k=1}^m \sum_{j=1}^N p(j|i, a) \phi_{f^\infty}^k(j) x_k \\ \sum_{k=1}^m \phi_{f^\infty}^k(i) x_k + \sum_{k=1}^m u^k(i) x_k \geq \sum_{k=1}^m r^k(i, a) x_k + \sum_{k=1}^m \sum_{j=1}^N p(j|i, a) u^k(j) x_k, \\ (i, a) \in \text{Gr } A, i \in S(f^\infty, i_1) \\ x_k > 0, k=1, \dots, m. \end{cases}$$

定理 4.1. f^∞ が最適ならば, 各線形不等式系 $(S_1), \dots, (S_N)$ は解をもつ.

定理 4.2. 各線形不等式系 $(T_1), \dots, (T_N)$ が解をもてば, f^∞ は最適である. ただし, $(T_1), \dots, (T_N)$ は, 定理 3.2 に対応する線形不等式系である.

系 4.1. 各線形不等式系 $(U_1), \dots, (U_N)$ が解をもてば, f^∞ は最適である. ただし, $(U_1), \dots, (U_N)$ は, 系 3.1 に対応する線形不等式系である.

次の LP 問題を考える.

$P(S_{i_1})$: Max z

subject to

$$\left\{ \begin{array}{l} x_1 \geq z, \dots, x_m \geq z \\ \sum_{k=1}^m \phi_{f^\infty}^k(i) x_k \geq \sum_{k=1}^m \sum_{j=1}^N p(j|i, a) \phi_{f^\infty}^k(j) x_k \\ \sum_{k=1}^m \phi_{f^\infty}^k(i) x_k + \sum_{k=1}^m u^k(i) x_k \geq \sum_{k=1}^m r^k(i, a) x_k + \sum_{k=1}^m \sum_{j=1}^N p(j|i, a) u^k(j) x_k, \\ x_k \geq 0, k=1, \dots, m, z \geq 0. \end{array} \right. \quad (i, a) \in Gr A, i \in S(f^\infty, i_1)$$

定理 4.3.

- (i) $P(S_{i_1})$ の最大値が正のとき, またその時に限り (S_{i_1}) は解を持つ. このとき, $P(S_{i_1})$ は非有界である.
(ii) $P(S_{i_1})$ の最大値が0のとき, またその時に限り (S_{i_1}) は解を持たない.

$P(T_{i_1})$, $P(U_{i_1})$ についても同様な定理が成り立つ.

5. 数値例

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

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

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

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

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

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

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

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

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

[γ の最適性の判定]

$$\Phi_\gamma = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}, \quad u_\gamma = \begin{bmatrix} 0 & -3 \\ 0 & 0 \end{bmatrix}$$

● $i=1, a=1$ のとき

$$x_1 + 2x_2 \geq x_1 + 2x_2$$

$$x_1 + 2x_2 - 3x_2 \geq -3x_2$$

● $i=1, a=2$ のとき

$$x_1 + 2x_2 \geq x_1 + 2x_2$$

$$x_1 + 2x_2 - 3x_2 \geq x_1 - x_2$$

- $i=2, a=1$ のとき

$$x_1 + 2x_2 \geq x_1 + 2x_2$$

$$x_1 + 2x_2 \geq 2x_1 + x_2 - 3x_2$$
- $i=2, a=2$ のとき

$$x_1 + 2x_2 \geq x_1 + 2x_2$$

$$x_1 + 2x_2 \geq x_1 + 2x_2$$

$P(S_0)$: $Max\ z$

subject to

$$\begin{cases} x_1 \geq z, x_2 \geq z \\ -x_1 - 2x_2 \leq 0 \\ x_1 - 4x_2 \leq 0 \\ x_1 \geq 0, x_2 \geq 0, z \geq 0. \end{cases}$$

これを解いて $Max\ z = \infty$. すなわち γ は最適である.

参考文献

- [1] C. Derman, Finite State Markovian Decision Processes (Academic Press, New York, 1970).
- [2] S. Durinovic, H. M. Lee, M. N. Katehakis, and J. A. Filar, Multiobjective Markov decision process with average reward criterion, Large Scale Systems 10 (1986) 215-226.
- [3] N. Furukawa, Vector-valued Markovian decision processes with countable state space, in: R. Hartley, L. C. Thomas, and D. J. White, eds., Recent Developments in Markov Decision Processes (Academic Press, New York, 1980) pp.205-223.
- [4] T. Iki and N. Furukawa, Vector-valued Markov decision processes with average criterion, Mem. Fac. Edu. Miyazaki Univ. Nat. Sci. 54-55 (1984) 1-10.
- [5] L. C. M. Kallenberg, Linear Programming and Finite Markovian Control Problems (Mathematisch Centrum, Amsterdam, 1983).
- [6] J. Novák, Linear programming in vector criterion Markov and semi-Markov decision processes, Optimization 20 (1989) 651-670.
- [7] L. C. Thomas, Constrained Markov decision processes as multi-objective problems, in: S. French, R. Hartley, L. C. Thomas, and D. J. White, eds., Multi-Objective Decision Making (Academic Press, New York, 1983) pp. 77-94.
- [8] K. Wakuta, Vector-valued Markov decision processes and the systems of linear inequalities, Stochastic Process. Appl. 56(1995) 159-169.
- [9] K. Wakuta and K. Togawa, A solution procedure for multiobjective Markov decision processes, Preprint.
- [10] K. Wakuta and K. Togawa, A solution procedure for multiobjective Markov decision processes : Average reward case, Preprint.
- [11] P. L. Yu, Multiple-Criteria Decision Making (Plenum Press, New York, 1985).