経営科学（日本オペレーションズ・リサーチ学会邦文機関誌）　**第14巻　第3号**（1970年12月）

完全エルゴード・マルコフ決定過程に関する一考察†

尾　崎　俊　治*

1. 序

マルコフ決定過程において，どのような政策に対しても，マルコフ連鎖がエルゴード的になるような決定過程，すなわち，完全エルゴード・マルコフ決定過程について議論する．

Blackwell [1] は同じ平均期待利得をもつ政策が2つ以上存在する場合には，修正項ベクトルを最大にする最適政策が存在することを示した．さらに，Veinott [4] は一般の場合に，この最適政策を求める政策反復アルゴリズムを求めた．

ここでは，完全エルゴード過程に対して，上に述べた意味の最適政策を求める簡単な線形計画あるいは政策反復アルゴリズムを開発し，その数値例を与える．

2. 準　　備

問題の設定および記号は Blackwell [1] にしたがうとする．以後の議論に必要な従来の結果を記す．

[**補題1**] (Blackwell [1])　Qを$S\times S$確率行列とすると，$(I+Q+\cdots\cdots+Q^N)/(N+1)$は$N\to\infty$のとき確率行列$Q^*$に収束する．そのとき，

(1)　　$\text{rank}\,(I-Q)+\text{rank}\,Q^*=S$

となる．

[**定理2**] (Blackwell [1])　任意の政策$f\in F$に対し，$Q(f)$に対応する行列を$Q^*(f)$とする．そのとき，

(2)　　$V_\beta(f)=[x(f)/(1-\beta)]+y(f)+\varepsilon(\beta,\ f)$

となる．ここで，$x(f)$は

(3)　　$(I-Q(f))\,x=0,\ Q^*(f)\,x=Q^*(f)\,r(f)$

の一意の解であり，$y(f)$ は

(4)　　$(I-Q(f))\,y=r(f)-x,\ Q^*(f)\,y=0$

---

† 1968年4月5日受理．

* 京都大学工学部，現在広島大学工学部．

の一意の解であり，$\beta\to 1$ のとき $\varepsilon(\beta, f)\to 0$ となる．

これらの結果より，完全エルゴード過程に対して，つぎの2つの系を得る．

［系3］ 完全エルゴード過程において，政策 $f_a\in F(a=1,2,\cdots\cdots,\nu)$ に対して

$$(5)\qquad x(f_a)=Q^*(f_a)\,r(f_a)=g\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}=g\cdot\mathbf{1}$$

すなわち，平均期待利得がすべて同じならば，

$$(6)\qquad (I-Q(f))\,y(f)=r(f)-x(f)$$

の解 $y(f)$ のうち，$y_S(f)=0$ とおくことによって得られる相対解 $y(f)$ はすべての $f_a\in F$ に対して同じである．

［証明］ 文献[3]で示したように完全エルゴード・マルコフ決定過程においては，$y_S(f)=0$ とおいた相対解 $y(f)$ と $x(f)=g\cdot\mathbf{1}$ は対応する線形計画問題の単体乗数となる．さて，ある基底解に対して文献[3]の式 (2.23) を満たす単体乗数が存在する．この単体乗数を用いて(2.23)式が成立するすべての政策 $f_a\in F$ $(a=1,2,\cdots\cdots,\nu)$ も同じ $x(f_a)=Q^*(f_a)\,r(f_a)=g\cdot\mathbf{1}$ となる．式 (2.23) を書き直せば，この系の式(6)と一致する．ただし，$y(f)$ は $y_S(f)=0$ とおいた相対解である．すなわち，相対解 $y(f)$ はすべての $f_a\in F$ に対して同じである．

［系4］ 式(6)の相対解を $y(f)=[v_i]$，正確な解を $y(f)=[v_i(f)]$ とすれば

$$(7)\qquad v_s(f)-v_S(f)=v_s\qquad (s=1,2,\cdots\cdots,S-1)$$

$$(8)\qquad v_S(f)=-\sum_{s=1}^{S-1}\pi_s(f)v_s$$

となる．ここで，$\pi_s(f)$ は $Q^*(f)$ の第 s 列要素，すなわち，状態 s の極限確率である．

［証明］ 式(7)は相対解の意味より明らかである．式(7)を式(4)の第2式に代入すれば，式(8)を得る．

3. アルゴリズム

完全エルゴード過程に対しては，つぎの線形計画問題

$$(9)\qquad \text{Max}\sum_{s=1}^{S}\sum_{k\in A}i(s,k)\,Z_s{}^k$$

subject to

$$(10)\qquad \sum_{k\in A}Z_{s'}{}^k-\sum_{s=1}^{S}\sum_{k\in A}q(s'\mid s,k)\,Z_s{}^k=0\qquad (s'=1,2,\cdots\cdots,S)$$

$$(11)\qquad \sum_{s=1}^{S}\sum_{k\in A}Z_s{}^k=1$$

$$(12)\qquad Z_s{}^k\geqq 0\qquad (s=1,1,\cdots\cdots,S,\ k\in A)\qquad A：\text{有限決定空間}$$

を解くことによって平均最適政策を得る．この線形計画問題の制限式(10)のうち1つは冗長であるから，$s'=S$ に関する制限式を除く．このとき，双対変数は $(v_1,v_2,\cdots\cdots,v_{S-1},g)$ となる．

したがって，平均最適政策 $f_a \epsilon F$ はこの線形計画問題より得られる．ただし，最終の単体判定基準において0になる項はすべて同じ平均最適政策となることに注意しなければならない．

つぎの定理はよく知られている．

［**定理5**］ 上の線形計画問題において，最適解の中には，各 $s \epsilon S$ に対し唯1つの $Z_s^k > 0$ で，他の k に対しては0となるものが存在する．

この定理により，$Z_s^k > 0$ ならば $Z_s^k = \pi_s(f)$ となる．すなわち，主変数 Z_s^k は極限確率を与えている．この事実と系4より，平均最適政策に対するすべての主および双対変数を求めれば，直ちに $v(f)$ を最大にする最適政策が求められる．あるいは，式(8)を目的関数と考えれば，式(9)の $i(s,k)$ のかわりに $-v_s$ を用いた線形計画問題を解くことによって，最適政策を得る．

一方，線形計画と政策反復アルゴリズムの関係は既に三根，尾崎[3]によって明らかにされているから，対応する政策反復アルゴリズムも直ちに確立できる．まず，通常の方法で，平均最適政策 $f_a \epsilon F$ $(a=1,\cdots\cdots,\nu)$ を求める．これらの平均最適政策 $(\nu \geqq 2)$ の中で，$v_S(f)$ を最大にする政策 $f \epsilon F$ を求めることになる．既に述べたように，$r(f)$ のかわりに $-y(f) = -[v_i]$ を用いて，$y_S(f) = 0$ とおいて，$y(f)$，g についての政策反復アルゴリズムを用いれば，最適の g が $v_S(f) = 0$ を与える．

4. 数 値 例

可能な政策 $f_a \epsilon F$ $(a=1,2,3,4,5)$ はつぎのように与えられる．

$$Q(f_a) = \begin{bmatrix} 0 & 1 \\ \dfrac{a}{5} & \dfrac{5-a}{5} \end{bmatrix}, \quad r(f_a) = \begin{bmatrix} 2 \\ 3+\dfrac{a}{5} \end{bmatrix}$$

線形計画問題の最適解の主および双対変数はつぎのようになる．

主変数 Z_s^k　　双対変数 v_1, g

$Z_1^1 = 0.1666$, $Z_2^1 = 0.833$, $v_1 = -1$, $g = 3$

$Z_1^1 = 0.285$, $Z_2^2 = 0.714$, $v_1 = -1$, $g = 3$

$Z_1^1 = 0.375$, $Z_2^3 = 0.625$, $v_1 = -1$, $g = 3$

$Z_1^1 = 0.444$, $Z_2^4 = 0.555$, $v_1 = -1$, $g = 3$

$Z_1^1 = 0.5$, $Z_2^5 = 0.5$, $v_1 = -1$, $g = 3$

したがって，$f_a \epsilon F$ $(a=1,2,3,4,5)$ はいずれも平均最適政策となり，特に f_5 が $v(f)$ を最大にする最適政策となる．また，

$$\pi_1(f_5) = 0.5, \quad \pi_2(f_5) = 0.5$$

であるから，式(7)および式(8)より

$$y(f_5) = \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix}$$

となる.

前節で述べた線形計画を用いれば，これらの政策はすべて同じ平均最適で，相対解 $v_1=-1$, $v_2=0$ となる．よって

$$\text{Max } Z_1^1$$

subject to

$$Z_1^1-\frac{1}{5}Z_2^1-\frac{2}{5}Z_2^2-\frac{3}{5}Z_2^3-\frac{4}{5}Z_2^4-\frac{5}{5}Z_2^5=0$$

$$Z_1^1+Z_1^2+Z_2^2+Z_2^3+Z_2^4+Z_2^5=1$$

$$Z_1^1,\ Z_2^k\geqq 0 \qquad (k=1,2,3,4,5)$$

なる線形計画問題を得る．この最適解は

$$Z_1^1=Z_2^5=0.5$$

$$Z_1^1=v_S(f_5)=0.5$$

となるから，上と同じ結果を得る.

また，政策反復アルゴリズムを用いるならば，例えば初期政策を f_1 とすれば，

$$v_1+g=1,\ -\frac{1}{5}v_1+g=0$$

より，$v_1=5/6$, $g=1/6$ を得る．PIR (Policy Improvement Routine) によって，つぎの政策は f_5 となる.

$$v_1+g=1,\ -\frac{5}{5}v_1+g=0$$

より，$v_1=1/2$, $g=1/2$ を得る．これが最適政策である.

5. 結　　論

以上述べたように，平均最適政策の中で，$y(f)$を最大にする最適政策を求める線形計画および政策反復アルゴリズムを確立した.

したがって，まず通常の方法で平均最適政策を求め，２つ以上の平均最適政策が存在すれば，もう一度ここで述べた線形計画あるいは政策反復アルゴリズムを用いることにより，$y(f)$を最大にする最適政策を求めることができる.

最後に，日頃御指導頂きます京都大学工学部三根久教授に厚く感謝します.

参 考 文 献

[1] Blackwell, D., "Discrete Dynamic Programming," *Ann. Math. Stat.*, **33** (1962), 719-726.
[2] Derman, C., "On Sequential Decisions and Markov Chains," *Management Sci.*, **9** (1962), 16-24.
[3] 尾崎俊治，"線形計画とマルコフ決定過程，" 経営科学，第14巻，第１号 (1970年７月)，17-33.
[4] Veinott, A. F., Jr., "On Finding Optimal Policies in Discrete Dynamic Programming with No Discounting," *Ann. Math. Stat.*, **34** (1966), 1284-1294.