マルコフ決定過程

——その解説的展望——

小山昭雄

1 はじめに

系の状態が時の経過とともに変化し，しかもその変化が偶然の作用によっておこるような現象は，確率論の分野で，確率過程の名のもとに古くから研究されている。このような現象は，自然科学の研究対象となるような系についてのみならず，社会科学の分野でもしばしば見られる。そして，社会科学の対象となる確率過程の多くは，マルコフ過程である。

社会科学における問題を自然科学におけるそれと区別する重要な点は，前者においては人間の意思決定が関わってくるということと，結果に対して何らかの利得が付随するということをあげることができる。これらをマルコフ過程に結びつけたものが，標題に掲げたマルコフ決定過程にほかならない。

マルコフ決定過程は1950年代に，R. Bellman によって初めて採り上げられた。しかし，それが広く人々の関心を引くようになったのは，1960年に出版されたR. A. Howard の名著 "Dynamic Programming and Markov Process" によるといってよいだろう。この書物で Howard は，マルコフ決定問題に対する完全な解答を与えたが，彼の方法はいかにもエンジニア的であって，論理のエレガンシーを要求する数学者の目から見ると，いささか気持の悪い面も少なくない。そのような面は，しかし，1962年，1965年の D. Blackwell の論文によって，完全に取り除かれた。一方，A. S. Manne, F. D'Epenoux 等によって，マルコフ決定過程の問題が線型計画の問題に結びつけられ，線型計画問題を解くことによって，マルコフ決定問題を解くことができるようになった。

R.A.Howard は前記の書物の中で，時間をあらわすパラメータが離散的な場合のみならず，連続的な場合についても完全な解を与えている。後に，同じHowardによって，さらに一般のセミ・マルコフ決定過程が考察された。この場合の線型計画的定式化は，尾崎・三根両氏によって試みられている。

以下において，上述の線に沿った理論の解説的展望を試みることにしたい。

2 マルコフ過程

系のとりうる状態は 1, 2, ……, N とし，時点 t において系の状態が i であったときに次の時点 $t+1$ において状態 j になる確率を p_{ij} とかく。p_{ij} は i と j のみに依存し，t には無関係に定まるものとする。p_{ij} を i から j への推移確率 (transition probability) とよぶ。むろん

$$0 \leqq p_{ij} \leqq 1$$

$$\sum_{j=1}^{N} p_{ij} = 1 \qquad i = 1, 2, \cdots, N$$

である。

系の状態の推移確率が p_{ij} $i,j=1,2,\cdots\cdots,N$ で与えられるような確率過程をマルコフ過程(Markov process)という。正確には有限マルコフ連鎖(finite Markov chain)というが，われわれはより一般のマルコフ過程はここでは考えないから，マルコフ過程というよび名で通すことにする。p_{ij} を i 行 j 列の元とする正方行列 $\boldsymbol{P}$ を推移確率行列 (transition probability matrix) という。

$$\boldsymbol{P}=\begin{pmatrix} p_{11}\ p_{12}\cdots\cdots p_{1N} \\ p_{21}\ p_{22}\cdots\cdots p_{2N} \\ \cdots\cdots \\ p_{N1}\ p_{N2}\cdots\cdots p_{NN} \end{pmatrix} \qquad \cdots\cdots(1)$$

マルコフ過程は $\boldsymbol{P}$ によって完全に記述される。

時点 t において，系の状態が i である確率を $\omega_i(t)$ とし，初期状態確率ベクトルを $\boldsymbol{\omega}(t)=(\omega_1(t),\ \omega_2(t),\cdots\cdots,\ \omega_N(t))\cdots\cdots(2)$ とする。そのとき，$t+1$ 時点における系の状態確率ベクトル $\boldsymbol{\omega}(t+1)$ は

$$\boldsymbol{\omega}(t+1)=\boldsymbol{\omega}(t)\boldsymbol{P} \qquad t=0,1,2,\cdots\cdots(3)$$

で与えられる。この関係を繰り返し適用すると

$$\boldsymbol{\omega}(t)=\boldsymbol{\omega}(0)\boldsymbol{P}^t \quad t=0,1,2,\cdots\cdots \qquad (4)$$

が得られる。

$\boldsymbol{P}^t$ は $t\to\infty$ のときにある極限行列に収束する場合としない場合とがある。収束する場合に極限行列を $\boldsymbol{P}_0$ とする。

$$\lim_{t\to\infty}\boldsymbol{P}^t=\boldsymbol{P}_0 \qquad (5)$$

このとき（4）の両辺で $t\to\infty$ とすると右辺は収束するから左辺も収束し

$$\lim_{t\to\infty}\boldsymbol{\omega}(t)=\boldsymbol{\omega} \qquad (6)$$

が存在する。そして（4），（5）から

$$\boldsymbol{\omega}=\boldsymbol{\omega}(0)\boldsymbol{P}_0 \qquad (7)$$

である。ところで（3）式の両辺で $t\to\infty$ とすると

$$\boldsymbol{\omega}=\boldsymbol{\omega}\boldsymbol{P} \qquad (8)$$

が得られる。これと $\boldsymbol{\omega}=(\omega_1,\ \omega_2,\cdots\cdots,\omega_N)$ の成分和が1であること，すなわち

$$\omega_1+\omega_2+\cdots\cdots+\omega_N=1 \qquad (9)$$

とから，$\omega_1,\omega_2,\cdots,\ \omega_N$ の値が決まる。一方

$$\boldsymbol{P}^{t+1}=\boldsymbol{P}^t\boldsymbol{P}$$

の両辺で $t\to\infty$ とすると（5）から

$$\boldsymbol{P}_0=\boldsymbol{P}_0\boldsymbol{P} \qquad (10)$$

が得られ，これから，$\boldsymbol{P}_0$ の任意の行ベクトルは，たとえば第 i 行ベクトル $\boldsymbol{a}_i$ は

$$\boldsymbol{a}_i=\boldsymbol{a}_i\boldsymbol{P} \qquad (11)$$

をみたし，しかも $\boldsymbol{a}_i$ の成分和は1である。(8)式と(11)式は同じものだから，ベクトル $\boldsymbol{\omega}$ と $\boldsymbol{a}_i$ $i=1,2,\cdots,N$ は同じものである。すなわち，(5)から定まる $\boldsymbol{P}_0$ の行ベクトルは，すべて同じであって，それは $\boldsymbol{\omega}$ と等しい。$\boldsymbol{a}_i$ は $\boldsymbol{\omega}(0)$ とは無関係だから，当然，$\boldsymbol{\omega}$ も，$\boldsymbol{\omega}(0)$ には依存しない。$\boldsymbol{P}^t$ が収束しない場合については，ここでは触れない。

〔例〕 オモチャ・メーカーの問題

Howard の書物にある有名な例である。オモチャ・メーカーの売り出す製品が，評判が良い場合と悪い場合とがある。良い方を状態1とし，悪い方を状態2とする。そして今期に状態1にあるとき，次の期にふたたび状態1にある確率は1/2，状態2に移る確率が1/2とする。また，今期に状態2にあれば，メーカーは新製品の開発等の努力をする。その結果次の期に状態1になる確率は2/5，依然として状態2のままである確率は3/5であるとする。この場合は

$$p_{11}=1/2,\ p_{12}=1/2,\ p_{21}=2/5,\ p_{22}=3/5$$

となるから，推移確率行列は

$$\boldsymbol{P}=\begin{pmatrix} 1/2 & 1/2 \\ 2/5 & 3/5 \end{pmatrix}$$

である。そして $\lim_{t\to\infty}\boldsymbol{P}^t$ は存在する。いま状態の確率分布ベクトルの極限を

$$\boldsymbol{\omega}=(\omega_1,\omega_2)$$

とすれば，(8)と(9)の関係から

$$(\omega_1,\omega_2)\begin{pmatrix} 1/2 & 1/2 \\ 2/5 & 3/5 \end{pmatrix}=(\omega_1,\omega_2)$$

$$\omega_1+\omega_2=1$$

である。これを解いて

$\omega_1=4/9 \quad \omega_2=5/9$

が得られる。したがって

$$\lim_{t\to\infty} \boldsymbol{P}^t=\lim_{t\to\infty}\begin{pmatrix}1/2 & 1/2\\ 2/5 & 3/5\end{pmatrix}^t=\begin{pmatrix}4/9 & 5/9\\ 4/9 & 5/9\end{pmatrix}$$

である。

3　利得の導入

状態 i から j への推移に伴って r_{ij} だけの利得が得られるものとする。r_{ij} を i 行 j 列の元とする行列 $\boldsymbol{R}$ を利得行列という。利得は必ずしも金額である必要はなく，問題によっていろいろな形をとりうる。たとえば生産量であることも可能である。

マルコフ過程に利得を導入することによって利得の系列が生ずる。現在の状態が i であるとき，n 期後までの間の利得の期待値を $v_i(n)$ とすれば，$v_i(n)$ について次の漸化式が成り立つ。

$$v_i(n)=\sum_{j=1}^{N} p_{ij}[r_{ij}+v_j(n-1)] \qquad (12)$$

なぜなら，現在状態 i で次の期に j になれば，この推移に伴って r_{ij} だけの利得が得られ，その後の $n-1$ 期間に $v_j(n-1)$ だけの期待利得が得られる。したがって i から j に移行した場合の向う n 期間の期待利得は $r_{ij}+v_j(n-1)$ であり，これが得られる確率は p_{ij} だからである。いま

$$r_i=\sum_{j=1}^{N} p_{ij}r_{ij} \qquad (13)$$

とおき，r_i を直接期待利得とよぶことにすれば，(12)は

$$v_i(n)=r_i+\sum_{j=1}^{N} p_{ij}\,v_j(n-1) \qquad (14)$$

とかくことができる。ベクトル，行列記号を用いてまとめてかくと

$$\boldsymbol{v}(n)=\boldsymbol{r}+\boldsymbol{P}\boldsymbol{v}(n-1) \qquad (15)$$

ここで

$$\boldsymbol{v}(n)=\begin{pmatrix}v_1(n)\\ v_2(n)\\ \vdots\\ v_N(n)\end{pmatrix},\quad \boldsymbol{r}=\begin{pmatrix}r_1\\ r_2\\ \vdots\\ r_N\end{pmatrix}$$

ある。

オモチャ・メーカーの例で利得行列を

$$\boldsymbol{R}=\begin{pmatrix}9 & 3\\ 3 & -7\end{pmatrix}$$

とすれば

$r_1=9\times0.5+3\times0.5=6$

$r_2=3\times0.4-7\times0.6=-3$

となるから

$$\boldsymbol{r}=\begin{pmatrix}6\\ -3\end{pmatrix}$$

である。

4　さまざまな手段の導入

オモチャ・メーカーの問題で，メーカーは状態1（評判が良い）の場合に，つぎの2つの手段のうちのいずれかをとることができる。

手段1　特別には何の手もうたない。このときは，前述のように，$p_{11}=0.5$，$p_{12}=0.5$，$r_{11}=9$，$r_{12}=3$ であるとする。

手段2　評判が下らないように広告宣伝をする。このときは，$p_{11}=0.8$，$p_{12}=0.2$ となって，次の期も引きつづいて良い評判を保つ確率は上昇するが，しかし利得は，$r_{11}=4$，$r_{12}=4$ と変化するものとする。

つぎに，状態2（評判が悪い）の場合もメーカーは次の2つの手段のうちのいずれかをとることができる。

手段1　特別には何の手もうたない。このときは前述のように $p_{21}=0.4$，$p_{22}=0.6$，$r_{21}=3$，$r_{22}=-7$ であるとする。

手段2　評判を回復するために，研究開発に努力する。この場合は $p_{21}=0.7$，$p_{22}=0.3$ となって，評判をとりもどす確率は上昇する

が，研究開発費の影響で利得は減って $r_{21}=1$，$r_{22}=-19$ となる。

この考え方を一般化しよう。

利得を伴ったマルコフ過程で，状態 i のときにとることのできる手段がいくつか（有限個）あり，それらの手段の集合を $\Omega(i)$ とかく。とりうる手段のうちの1つ，たとえば $k\epsilon\Omega(i)$，をとれば，次の期に j に移行する確率は $p_{ij}{}^k$ となり，そのときの利得は $r_{ij}{}^k$ となる。すなわち，とる手段によって推移確率が変り，また推移に伴う利得も変るのである。

ここまできて，マルコフ決定過程が定まったわけである。

5 最適政策の決定・有限期間の場合

前節で述べたマルコフ決定過程において，第 n 期までの利得の期待値を最大にするには，毎期どのような手段をえらんだらよいか，という問題を考える。

系の状態が i のときに，とりうる手段の集合 $\Omega(i)$ のなかの1つの手段 k をえらぶわけであるが，ここで，各 i に対して $\Omega(i)$ の中の1つの k を対応させる関数を考え，これを決定関数 (decision function) とよび，f であらわす。すなわち，各 i に対して $f(i)\epsilon\Omega(i)$ である。決定関数の系列を政策 (policy) とよぶ。とくに，同一の決定関数のみからなる政策を定常政策 (stationary policy) とよぶ。n 期間にわたる総期待利得を最大にするような政策を最適政策 (optimal policy) という。

さて，

$v_i(n)$＝状態 i からスタートして n 期間にわたる最適政策のもとでの総期待利得

とおく。そのときダイナミック・プログラミングの考え方を (12) 式に適用すると

$$v_i(n)=\max_{k\epsilon\Omega(i)}\sum_{j=1}^{N} p_{ij}{}^k\,[r_{ij}{}^k+v_j(n-1)] \tag{16}$$

が成立する。$n=1$ のときは

$$v_i(1)=\max_{k\epsilon\Omega(i)}\sum_{j=1}^{N} p_{ij}{}^k r_{ij}{}^k \tag{17}$$

である。(17) から始めて，(16) の漸化式をつぎつぎに解けば $v_i(n)$ が求まり，それぞれの段階での最適な決定関数が決まる。

6 最適政策の決定・無限期間の場合

計画期間が無限になると，毎期の期待利得の無限の期間にわたる総計は，一般に無限大になる。このような場合に普通採用される処理方法としては，将来の利得は割引いて考え，その上で，割引現在価値の総計を考察の対象にするやり方と，総期待利得の代りに1期当りの平均期待利得を考察の対象とするやり方とがある。しかし，後者の平均期待利得を考える場合は，実は，無限の期間を考慮した場合の1期当りの平均利得というものが，そもそも，定義できない，といった場合も生じてきて，その取り扱いはかなり数学的にも面倒な面が多い。そこで，ここでは前者の，将来利得は割引いて考えるという立場を採用し，無限の将来にわたる利得の割引現在価値を考察の対象とする。割引率を β $(0<\beta<1)$ としよう。そうすると (16) 式はつぎのようになる。

$$v_i(n)=\max_{k\epsilon\Omega(i)}\sum_{j=1}^{N} p_{ij}{}^k[r_{ij}{}^k+\beta v_j(n-1)] \tag{18}$$

あるいは

$$r_i{}^k=\sum_{j=1}^{N} p_{ij}{}^k\ r_{ij}{}^k$$

とおけば

$$v_i(n)=\max_{k\epsilon\Omega(i)}[r_i{}^k+\beta\sum_{j=1}^{N} p_{ij}{}^k v_j(n-1)] \tag{20}$$

となる。割引率を考慮に入れているわれわれの場合には，$\lim_{n\to\infty} v_i(n)=v_i$ が存在すること

が示される。したがって(20)の両辺で $n\to\infty$ とすると

$$v_i=\max_{k\in\Omega(i)}\left[r_i^k+\beta\sum_{j=1}^{N}p_{ij}^k\ v_j\right] \qquad (21)$$

が得られる。

一方，次の定理の成り立つことが示される。

定理1　定常政策で最適なものが存在する。

この定理によれば，われわれは考察する政策の範囲を定常政策に限定し，その中から最適なものを求めればよい。定常政策であるから，期の如何を問わず，状態 i のみに依存して $k\in\Omega(i)$ が決まることになる。最適政策は次の定理で示されるアルゴリズムによって求められる。

定理2　最適政策決定のアルゴリズム

（a）　数値決定演算 (value determination operation)

与えられた決定関数 f_n に対して，次の連立方程式を v_i について解く

$$v_i=r_i^k+\beta\sum_{j=1}^{N}p_{ij}^k\ v_j \qquad i=1,2,\cdots,N \qquad (22)$$

ただし，$k=f_n(i)$

（b）　政策改良ルーチン

上で求められた $v_1,v_2,\cdots\cdots,v_N$ を使って，各 i に対して

$$r_i^k+\beta\sum_{j=1}^{N}p_{ij}^k\ v_j \qquad (23)$$

を最大ならしめるような k を対応させる決定関数 f_{n+1} をつくる。$f_n\neq f_{n+1}$ なら，いま求めた f_{n+1} に対して（a）の演算から繰り返す。$f_n=f_{n+1}$ なら，つねに f_n を採用するという政策が最適政策になる。

最初の決定関数 f_1 は，各 i に対して r_i^k を最大にする k を対応させるものをとればよい。

この定理の証明はしないけれども，(21)式がアルゴリズムの妥当性を示唆していることは見てとれる。

例。まえにあげたオモチャ・メーカーの問題を $\beta=0.9$ として解いてみよう。データを改めて整理してかいておく。

状態1のとき手段1をとれば

$p_{11}^1=0.5,\ p_{12}^1=0.5,\ r_{11}^1=9,\ r_{12}^1=3$

手段2をとれば

$p_{11}^2=0.8,\ p_{12}^2=0.2,\ r_{11}^2=4,\ r_{12}^2=4$

状態2のとき手段1をとれば

$p_{21}^1=0.4,\ p_{22}^1=0.6,\ r_{21}^1=3,\ r_{22}^1=-7$

手段2をとれば

$p_{21}^2=0.7,\ p_{22}^2=0.3,\ r_{21}^2=1,\ r_{22}^2=-19$

したがって

$r_1^1=6,\ r_1^2=4,\ r_2^1=-3,\ r_2^2=-5$

である。

まず，f_1 として，各 i に対して r_i^k を最大にする k を対応させるものをえらぶと，上のデータから

$f_1(1)=1,\ f_1(2)=1$

である。この f_1 に対してつくられた連立方程式 (22) はつぎのようになる。

$$\begin{cases}v_1=6+0.9(0.5v_1+0.5v_2)\\ v_2=-3+0.9(0.4v_1+0.6v_2)\end{cases}$$

これを解いて $v_1=15.5,\ v_2=5.6$ が得られる。

つぎに，この $v_1,\ v_2$ を使って（b）の政策改良ルーチンをおこなう。各 i に対して (23) 式の値を計算すると下のようになる。

i	k	$r_i^k+\beta\sum_j p_{ij}^k v_j$
1	1	$6+0.9(0.5\times15.5+0.5\times5.6)=15.5$
	2	$4+0.9(0.8\times15.5+0.2\times5.6)=\boxed{16.2}$
2	1	$-3+0.9(0.4\times15.5+0.6\times5.6)=5.6$
	2	$-5+0.9(0.7\times15.5+0.3\times5.6=\boxed{6.3}$

ここで $r_i^k+\beta\sum_j p_{ij}^k\ v_j$ の値の大きい方が□で囲んである。この結果，政策改良ルーチンによって得られる f_2 は

$f_2(1)=2,\ f_2(2)=2$

となる。

この f_2 に対して（a）による（22）の連立方程式をつくれば

$$\begin{cases} v_1=4+0.9(0.8v_1+0.2v_2) \\ v_2=-5+0.9(0.7v_1+0.3v_2) \end{cases}$$

となり，これを解いて $v_1=22.2$，$v_2=12.3$ が得られる。そこで，この v_1, v_2 を用いて政策改良ルーチンの計算を実行すると，結果として得られる f_3 は f_2 と同じものであることが確かめられる。よって，つねに f_2 を採用するということ，すなわち，評判の良いときでも広告宣伝をし，評判の悪いときは研究開発をするというのが最適政策なのである。

7 線型計画法による定式化

マルコフ決定過程における最適政策を，定理2のアルゴリズムではなく，線型計画法によって解くことも可能である。それをここで説明する。ただし線型計画法についての解説は省略し，それについての知識は仮定した上で話を進めてゆく。

線型計画法で問題を定式化するためには，決定関数の概念を拡張しなくてはならない。

いままでは，各状態 i に対して手段の集合 $\Omega(i)$ の中から1つの手段 k をえらんで対応させる関数を決定関数とよんだが，ここでは，それを，つぎのように拡張する。

状態 i のときに $\Omega(i)$ の中から手段 k をえらぶ確率を考え，これを q_i^k として，各 i に対して $\Omega(i)$ 上の確率分布 $\{q_i^k\}$ を対応させる関数を定義するのである。この関数を混合決定関数とよぶことにし，それをあらわすのに，いままでと同じ記号 f を用いることにする。混合決定関数に対して，従来の意味での決定関数を純粋決定関数という。純粋決定関数が混合決定関数の特殊の場合であることはいうまでもない。混合決定関数の系列を混合政策という。同一の混合決定関数だけから成る政策を定常混合政策という。

純粋政策のみを考察の対象とする場合には，定理1で述べたように，定常な最適政策が存在する。しかし，混合政策にまで概念を拡張した場合であっても，やはり定常な最適政策が存在することが示される。この事実を定理としてあげておく。

定理3　混合政策まで考慮した場合でも，最適な定常政策が存在する。

この定理により，最適政策を求めるに当って定常政策だけに考察の範囲をしぼってよい。そこで以後は定常政策だけを考えることにし，決定関数と政策は特に記号上の区別はせず，f とかいたら，これは決定関数でもあり，また同時に（$f, f, \cdots\cdots$）という定常政策をもあらわすものとする。

なお，ここでは，考察期間は無限であるとする。

ここでつぎのような記号を導入する。

$v_i(f)$＝定常政策 f のもとで，状態 i からスタートした場合の，無限の将来にわたる全期待利得の割引現在価値

$p_{ij}^{(n)}(f)$＝定常政策 f のもとで，状態 i からスタートして n 期後に状態 j になる確率

$r_j(f)$＝定常政策 f のもとで状態 j からスタートしたときの直接期待利得

混合決定関数 f が各 i に対して $\Omega(i)$ 上の確率分布 $\{q_i^k\}$ で定義されているとき

$$f=\{q_i^k\,;\ k\in\Omega(i), i=1,2,\cdots\cdots,N\} \tag{24}$$

とかくことにする。この f に対して

$$r_i(f)=\sum_{k\in\Omega(i)} q_i^k r_i^k \tag{25}$$

$$p_{ij}(f)=\sum_{k\in\Omega(i)} q_i^k p_{ij}^k$$

一般に

$$p_{ij}^{(n)}(f)=\sum_{\nu=1}^{N} p_{i\nu}^{(n-1)}(f)p_{\nu j}(f) \tag{26}$$

となり，そして

$$v_i(f)=\sum_{n=0}^{\infty}\beta^n\left(\sum_{\nu=1}^{N}p_{i\nu}{}^{(n)}(f)\,r_\nu(f)\right) \tag{27}$$

である。

われわれの目的は，各 i に対して $v_i(f)$ を最大にするような f を求めることにある。

線型計画法によって問題を定式化するために，つぎの線型計画問題を考える。

主問題

$$\begin{aligned}
&v_1\geqq r_1{}^k+\beta\sum_{j=1}^{N}p_{1j}{}^k\,v_j \quad k\in\Omega(1)\\
&v_2\geqq r_2{}^k+\beta\sum_{j=1}^{N}p_{2j}{}^k\,v_j \quad k\in\Omega(2)\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
&v_N\geqq r_N{}^k+\beta\sum_{j=1}^{N}p_{Nj}{}^k v_j \quad k\in\Omega(N)
\end{aligned} \tag{28}$$

のもとで

$$\min\ a_1v_1+a_2v_2+\cdots+a_Nv_N \tag{29}$$

ここで目的関数の係数 a_1, a_2, $\cdots$, a_N はいずれも正であって，和が1になるように与えられている。$(a_1, a_2, \cdots, a_N)$ は，系の初期の状態確率分布と考えてよい。$\Omega(i)$ の元の個数を K_i とすれば，制約式は全部で $K_1+K_2+\cdots\cdots+K_N$ 個ある。なお，変数の非負条件がないことに注意。

この問題の双対問題は次の形になる。

双対問題

$$\sum_{i=1}^{N}\sum_{k\in\Omega(i)}(\delta_{ij}-\beta p_{ij}{}^k)w_i{}^k=a_j \qquad j=1,2,\cdots,N \tag{30}$$

$$w_i{}^k\geqq 0,\quad i=1,2,\cdots,N,\quad k\in\Omega(i) \tag{31}$$

のもとで

$$\max\sum_{i=1}^{N}\sum_{k\in\Omega(i)}r_i{}^k w_i{}^k \tag{32}$$

まず双対問題の方から考える。双対問題に対して，その実行可能解と混合決定関数の間に1対1の対応がつけられることを示すつぎの定理が成立する。

定理4　任意の混合決定関数 f を

$$f=\{q_i{}^k\,;\,k\in\Omega(i),\ i=1,2,\cdots,N\}$$

とする。そのとき

$$\overline{w}_i{}^k=\left(\sum_{\nu=1}^{N}a_\nu\sum_{n=0}^{\infty}\beta^n p_{\nu i}{}^{(n)}(f)\right)q_i{}^k \qquad k\in\Omega(i),\ i\in S \tag{33}$$

によって定義される $\{\overline{w}_i{}^k\}$ は双対問題の実行可能解を与える。

逆に，双対問題の任意の実行可能解 $\{w_i{}^k\}$ に対して

$$q_i{}^k=w_i{}^k\Big/\sum_{l\in\Omega(i)}w_i{}^l \tag{34}$$

とおけば，この $\{q_i\,;\,k\in\Omega(i)\}$ は $\Omega(i)$ 上の確率分布を与え，したがって，各 i に対して $\{q_i{}^k\ k\in\Omega(i)\}$ を対応させることによって1つの混合決定関数 f が定義される。しかも，このようにして定義された f を用いて (33) 式によって $\overline{w}_i{}^k$ を定義すれば，$\overline{w}_i{}^k=w_i{}^k$ となる。すなわち，変換 (34) は双対問題の実行可能解と混合決定関数の間の1対1の対応を与える。

つぎに，双対問題の最適解とマルコフ決定過程の最適政策の間には，つぎの定理が成立する。

定理5　双対問題の最適解を $\{w^*{}_i{}^k\}$ とすれば，

$$q^*{}_i{}^k=w^*{}_i{}^k\Big/\sum_{l\in\Omega(i)}w^*{}_i{}^l$$

によって定まる混合決定関数 f^* は最適政策を与える。

最適政策を与えるような決定関数は，実は純粋決定関数の中からえらぶことが可能である。すなわち

定理6　最適な純粋定常政策が存在し，それは，双対問題をシンプレクス法で解くことによって得られる。

以上の諸定理は双対問題にかかわるものであったが，主問題に関してつぎの定理が成立する。

定理7　主問題の最適解を v_1^*, v_2^*, $\cdots$, v_N^* とすれば，これに対して

$$v_i^*=\max_{k\in\Omega(i)}\left\{r_i^k+\beta\sum_{\nu=1}^{N}p_{i\nu}^k k_\nu^*\right\} \qquad (35)$$

が成立する。

ところで定理2のアルゴリズムで示したことからわかるように，(35) の関係が成立するような $v_1^*,\cdots v_N^*$ は，実は最適政策が採用された場合の v_i の値にほかならない。したがって，双対問題の最適解から最適政策が得られ，主問題の最適解として，最適政策に対応する最適な $v_i(f)$ の値が得られるのである。

なお，主問題の最適解 $v_1^*,\cdots, v_N^*$ が得られれば，それから最適政策を求めることができる。すなわち

定理8 主問題の最適解を $v_1^*, v_2^*,\cdots, v_N^*$ としたとき，各 i に対して

$$v_i^*=r_i^k+\beta\sum_{\nu=1}^{N}p_{i\nu}^k v_\nu^*$$

を成立せしめるような k を対応させる決定関数を f とすれば，f は最適な決定関数である。

以上の一連の定理により，マルコフ決定過程と線型計画法との間の関係は完全に明らかになった。

例。オモチャ・メーカーの問題を線型計画法で定式化してみよう。データは

$p_{11}^1=0.5$, $p_{12}^1=0.5$　　$r_1^1=6$

$p_{11}^2=0.8$, $p_{12}^2=0.2$　　$r_1^2=4$

$p_{21}^1=0.4$, $p_{22}^1=0.6$　　$r_2^1=-3$

$p_{21}^2=0.7$, $p_{22}^2=0.3$　　$r_2^2=-5$

であるから，主問題および双対問題はつぎのようになる。ただし，$\beta=0.9$, $a_1=a_2=0.5$ とする。

主問題

$v_1\geqq 6+0.9(0.5v_1+0.5v_2)$

$v_1\geqq 4+0.9(0.8v_1+0.2v_2)$

$v_2\geqq -3+0.9(0.4v_1+0.6v_2)$

$v_2\geqq -5+0.9(0.7v_1+0.3v_2)$

のもとで

$\min\ 0.5v_1+0.5v_2$

制約式の変数を左辺にまとめると

$0.55v_1-0.45v_2\geqq 6$

$0.28v_1-0.18v_2\geqq 4$

$-0.36v_1+0.46v_2\geqq -3$

$-0.63v_1+0.73v_2\geqq -5$

のもとで

$\min\ 0.5v_1+0.5v_2$

という形になる。したがって双対問題はつぎの形をとる。

双対問題

$0.55w_1^1+0.28w_1^2-0.36w_2^1-0.63w_2^2 \leqq 0.5$

$-0.45w_1^1-0.18w_1^2+0.46w_2^1+0.73w_2^2 \leqq 0.5 \quad w_i^k\geqq 0$

のもとで

$\max 6w_1^1+4w_1^2-3w_2^1-5w_2^2$

これらをシンプレクス法で解けば，数値決定演算，政策改良ルーチンで求めたのと同じ結論が得られる。

参考文献

小河原正巳，坂本武司『マルコフ過程』共立出版，1967.

後藤昌司「マルコフ型逐次決定過程」『オペレーションズ・リサーチ』13巻，第2～第8号 日科技連，1968.

尾崎俊治「線型計画とマルコフ決定過程」『経営科学』第14巻，第1号，1970.

小山昭雄『マルコフ過程とその周辺』東洋経済 近刊

R. Bellmann「Dynamic Programming」Princeton. 1957.

D. Blackwell. "Discrete Dynamic Programming." Ann. Math, Stat. 33, 1962.

D. Blackwell. "Discounted Dynamic Programming" Ann, Math. Stat. 36, 1965.

F. D' Epenoux "A Probabilistic Production and Inventory Problem." Mgt. Sci. 10, 1963.

C. Derman "On Sequential Decisions and Markov Chanis" Mgt, Sci. 9.1962.

C. Derman. "Finite Morkovian Decision Processes" Academic Press. 1970.

R. A. Howard「Dynamic Programming

and Markov Process」M. I. T. Press. 1960.
R. A. Howard. "Research in Semi-Markovian Decision Structures." J. O. R. S. J. 6, 1964.
A. S. Manne "Linear Programming and Sequential Decisions." Mgt, Sci. 6, 1960.
S. Osaki and H. Mine. "Linear Programming Algorithms for Semi-Markovian Decision Processes." J. Math Anal. Appl. 22 1968.
P. Wolfe and G. B. Dantzig "Linear Programming in a Markov Chain." J. O. R. S. A. 10, 1962.