# マルコフ過程とマルコフ決定過程

渡邊直樹

2021 年 9 月 29 日

## 1　マルコフ過程の例

マルコフ過程（Markov process）とは，ある状態から別の状態への遷移確率が与えられたとき，その遷移が数期に渡って続いた結果，初期状態がどのような状態に変化するかを考察する確率過程（stochastic process）のことであり，遷移の行き着いた状態（あるいは漸近する状態）が初期状態とは無関係に決まる性質を持つ．

市場構造が明確であり，取引される価格と数量の予想が得られるならば，市場占有率（シェア）から状態 $i$ のときの意思決定者の利得を割り出すことができるだろう．ここでは，フォーマルな記述はスキップして，マルコフ過程を用いたシェアの遷移に関する簡単な例を示す．

ある財が A 社と B 社によって市場に供給されているとしよう．A 社のシェアは 70%であり，B 社のそれは 30%であるが，今期 B 社は新モデルの開発に成功したため，A 社の顧客の 20%が B 社の顧客となった．一方，A 社も B 社に対抗して新サービスの投入し，B 社の顧客の 20%を獲得した．ここでは，単純化のため新規顧客はいないとしておく．A 社の今期のシェアを $a_0$，B 社のそれを $b_0$ で表すと，次の期のシェア $a_1$ と $b_1$ は

$$\begin{aligned} a_1 &= 0.8a_0 + 0.2b_0 \\ b_1 &= 0.2a_0 + 0.8b_0 \end{aligned}$$

と表せる．これを行列を用いて書き直すと，

$$\begin{pmatrix} a_1 \\ b_1 \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix} \begin{pmatrix} a_0 \\ b_0 \end{pmatrix}$$

となる．同様に，この遷移が恒久的に続く場合，$t$ 期後の A 社と B 社のシェアを $a_t$ と $b_t$ で表すと，

$$\begin{pmatrix} a_t \\ b_t \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}^t \begin{pmatrix} a_0 \\ b_0 \end{pmatrix}$$

となる．これにより，$a_0 = 0.7$，$b_0 = 0.3$ のとき，

$$a_1 = 0.620, \quad b_1 = 0.380$$
$$a_2 = 0.572, \quad b_2 = 0.428$$

となり，シェアは $a_\infty = 0.500$，$b_\infty = 0.500$ に漸近していくことが判る．このように，遷移確率が各期に依存しないとき，それを時間的に斉次なマルコフ過程といい，遷移確率が当該期から見て有限個の過去の状態にも依存して決まる場合には，それを多重マルコフ過程という．

## 2　マルコフ決定過程

マルコフ決定過程（Markov decision process）とは，ある状態から別の状態への遷移確率が当該期における状態と行動によって決まるとき，各期の状態に依存してどのような行動を取るべきかを考察する意思決定理論である．

- $f(i) = k$: 状態（state）$i$ に直面したときに行動（action）$k$ をとることを表す（意思）決定関数（decision function）．
- $f_m = (f_m(1), f_m(2))$: 状態の個数が 2 つのとき，$m$ 期における意思決定関数．
- $s = (f_1, f_2, \ldots)$: 政策（policy）または戦略（strategy）．
- $f^\infty = (f, \ldots, f, \ldots)$: 定常政策（stationary policy）．毎期同じ政策関数を採用．
- $r_{ij}^k$: 状態 $i$ のときに行動 $k$ をとって状態 $j$ が実現した当該期の利得．
- $p_{ij}^k$: 状態 $i$ のときに行動 $k$ をとったとして，状態 $j$ が実現する確率．
- $r_i^k$: 状態 $i$ のときに行動 $k$ をとった当該期における期待利得．
  $r_i^k = p_{i1}^k r_{i1}^k + p_{i2}^k r_{i2}^k + \cdots$.

- $\beta \in (0,1)$: 割引率.
- $v_i(s)$: 初期状態が $i$ のとき，政策 $s$ を選択（choice）した場合の将来にわたる期待利得（expected reward or expected payoff）の割引現在価値の総和.
- $v(s) = (v_1(s), v_2(s))$: 状態の個数が 2 つのときに，それぞれの状態 $i$ を初期状態とする $v_i(s)$ を第 $i$ 要素とするベクトル.
- $v(s^*) \geq v(s)$ なる $s^*$: 最適政策（optimal policy）.

以下では，例を用いて，最適な定常政策の導出過程を示す．また，割引率は $\beta = 0.9$ とし，各数値は小数点以下第 2 位を四捨五入し，小数点以下第 1 位までの概算を用いている．また，ここで紹介する解法は動的計画法と呼ばれている．次の数値例は Howard (1960) に記されている有名な例の 1 つである.

- 数値決定演算（value determination operation）

ここで示される解法は Blackwell (1960, 1965) の考え方に沿ったものである．最初の決定関数 $f_0$ を，各状態 $i$ について，当該期の期待利得 $r_i^k$ を最大にする行動 $k$ を対応させるようにとる．つまり，初期状態 $i$ において，今期の期待利得を最大にするような「近視眼的な（myopic）」政策をまずは候補とする.

表 1: Transition and Data

| 状態 | 行動 | 推移確率 | | 利得 | | 当該期の期待利得 |
|---|---|---|---|---|---|---|
| $i$ | $k$ | $p_{i1}^k$ | $p_{i2}^k$ | $r_{i1}^k$ | $r_{i2}^k$ | $r_i^k = p_{i1}^k r_{i1}^k + p_{i2}^k r_{i2}^k$ |
| 1 | 1 | 0.8 | 0.2 | 4.5 | 2 | 4 |
| | 2 | 0.5 | 0.5 | 9 | 3 | 6 |
| 2 | 1 | 0.7 | 0.3 | 1 | -19 | -5 |
| | 2 | 0.4 | 0.6 | 3 | -7 | -3 |

表 1 より，$f^0 = (f^0(1), f^0(2))$ を

$$f^0(1) = 2, f^0(2) = 2$$

と定める．この $f^0$ に対して，割引現在価値の総和は次の再帰的方程式で記述できる．

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

これを解いて，$v_1 = 15.5$，$v_2 = 5.6$．

- 政策改良ルーチン（policy improvement routine）

このようにして求めた $v_1 = 15.5$ と $v_2 = 5.6$ の下で，$m = 0$ 期において各状態 $i$ において行動 $k$ をとったときの期待利得の割引現在価値の総和は表 2 にまとめられている．

表 2: Policy Improvement Routine 1

| 状態 $i$ | 行動 $k$ | 割引現在価値の総和 $r_i^k + \beta \sum_\nu p_{i\nu}^k v_\nu$ |
|---|---|---|
| 1 | 1 | 4+0.9(0.8×15.5+0.2×5.6)=<u>16.2</u> |
| | 2 | 6+0.9(0.5×15.5+0.5×5.6)=15.5 |
| 2 | 1 | -5+0.9(0.7×15.5+0.3×5.6)=<u>6.3</u> |
| | 2 | -3+0.9(0.4×15.5+0.6×5.6)=5.6 |

表 2 は，状態 1 のときは行動 1 を，状態 2 のときも行動 1 をとった方が期待利得の割引現在価値の総和が大きくなることを示している．よって，$f^1 = (f^1(1), f^1(2))$ を

$$f^1(1) = 1, f^1(2) = 1$$

と定める．この $f^1$ に対して，割引現在価値の総和は次の再帰的方程式で記述できる．

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

これを解いて，$v_1 = 22.2$，$v_2 = 12.3$．

このようにして求めた $v_1 = 22.2$ と $v_2 = 12.3$ の下で，各状態 $i$ において行動 $k$ をとったときの期待利得の割引現在価値の総和は表 3 にまとめられている．

以上より，毎期 $f^1$ をとる $f^\infty = (f^1, \ldots, f^1, \ldots)$ が定常最適政策であることが判る．なお，表 3 で，$f^1$ よりもよい政策が見つかれば，それを新たな $f^1$ として更新し，同様の計算を繰り返す．

表 3: Policy Improvement Routine 2

| 状態 $i$ | 行動 $k$ | 割引現在価値の総和 $r_i^k + \beta \sum_\nu p_{i\nu}^k v_\nu$ |
|---|---|---|
| 1 | 1 | 4+0.9(0.8×22.2+0.2×12.3)=<u>22.2</u> |
| | 2 | 6+0.9(0.5×22.2+0.5×12.3)=15.5 |
| 2 | 1 | -5+0.9(0.7×22.2+0.3×12.3)=<u>12.3</u> |
| | 2 | -3+0.9(0.4×22.2+0.6×12.3)=5.6 |

# 補論：線形計画法による解法

第2章の例において，たとえば，初期状態が状態1である確率を $a_1 = 0.5$，状態2である確率を $a_2 = 0.5$ とすると，政策改良ルーチンを繰り返し適用することで求めた解を線形計画法で求めることができる．まず，$q_i^k$ を状態 $i$ のとき，行動 $k$ をとる確率としよう．各状態 $i$ に対して，そこで取りうる行動の集合上の確率分布 $\{q_i^k\}$ を対応させる関数を**混合決定関数**という．これに対して，第1章で定義された決定関数は**純粋決定関数**ということもある[1]．簡単な線形計画問題であれば，Excel のアドインに組み込まれている「ソルバー」でも解くことができる．

**主問題**

$$
\begin{aligned}
\max \quad & 0.5v_1 + 0.5v_2 \\
s.t. \quad & (1 - 0.9 \times 0.8)v_1 - 0.9 \times 0.2v_2 \geq 4 \\
& (1 - 0.9 \times 0.5)v_1 - 0.9 \times 0.5v_2 \geq 6 \\
& 0.9 \times 0.7v_1 - (1 - 0.9 \times 0.3v_2) \geq -5 \\
& 0.9 \times 0.4v_1 - (1 - 0.9 \times 0.6v_2) \geq -3
\end{aligned}
$$

つまり，

$$
\begin{aligned}
\min \quad & 0.5v_1 + 0.5v_2 \\
s.t. \quad & 0.28v_1 - 0.18v_2 \geq 4 \\
& 0.55v_1 - 0.45v_2 \geq 6 \\
& -0.63v_1 + 0.73v_2 \geq -5 \\
& -0.36v_1 + 0.46v_2 \geq -3
\end{aligned}
$$

[1] ここでは詳しくは述べないが，各期における混合決定関数 $f_m$ の系 $s = (f_1, f_2, \ldots)$ を**混合政策**という．

双対問題

$$\begin{array}{ll} \min & 4w^1 + 6w_1^2 - 5w_2^1 - 3w_2^2 \\ s.t. & 0.28w_1^1 + 0.55w_1^2 - 0.63w_2^1 - 0.36w_2^2 = 0.5 \\ & -0.18w_1^1 - 0.45w_1^2 + 0.73w_2^1 + 0.46w_2^2 = 0.5 \\ & w_1^2 \geq 0, w_1^2 \geq 0, w_2^1 \geq 0, w_2^2 \geq 0 \end{array}$$

シンプレックス法で双対問題を解くと，

$$w_1^1 = 7.45, w_1^2 = 0, w_2^1 = 2.52, w_2^2 = 0$$

となるので，ここから定まる $\{q_i^k\}$ は

$$q_1^1 = 1, q_1^2 = 0, q_2^1 = 1, q_2^2 = 0.$$

こうして，$f^\infty = (f_1, \ldots, f_1, \ldots)$ が定常最適政策であることが判る．双対問題の最適解の正の成分（$w_1^1$ と $w_2^1$）に対応する主問題の制約式はその最適解によって等号で満たされる．よって，

$$\begin{array}{rcl} 0.28v_1 - 0.18v_2 & = & 4 \\ -0.63v_1 + 0.73v_2 & = & -5 \end{array}$$

より，$v_1 = 22.2$，$v_2 = 12.3$ が求められる．

# 参考文献


[1] Howard, R. A. (1960) “Repositioning Dynamics and Pricing Strategy,” *Dynamic Programming and Markov Process*, The MIT Press.

[2] Blackwell, D. “Discrete Dynamic Programming,” *Annals of Mathematical Statistics* 33, 719-726.

[3] Blackwell, D. “Discounted Dynamic Programming,” *Annals Mathematical. Statistics* 36, 226-235.