数理解析研究所講究録
第1939巻 2015年 79-87

マルコフ決定過程における学習プロセスと決定について

中井　達

千葉大学教育学部

1　はじめに

Nakai[3, 5, 6] などにおいて、状態空間が $(-\infty, \infty)$ の部分観測可能なマルコフ決定過程における学習過程と最適政策・最適値との関係を考えた。ここでは、状態に関する情報を、状態空間 $(-\infty, \infty)$ 上の確率変数で表し、状態 $s \in (-\infty, \infty)$ が大きくなれば、良い状態と考えた。アウトカムの指標を状態とする確率過程を考える。

アウトカムの指標を状態とする確率過程を考え、マルコフ過程での多段決定問題として支出モデルを定式化する。状態は確率的に推移するとともに、追加支出によっても変化する。アウトカムを改善するためにどのくらい支出すれば良いかを決める。

2　逐次支出問題

状態空間が $(-\infty, \infty)$ のマルコフ過程を考え、状態 s をアウトカムを表す指標とする。指標は s が大きくなるにしたがって良くなると考え、この指標を改善するために支出を行う。状態は、支出によるだけでなく、マルコフ過程の推移法則にしたがって推移する。アウトカムを良くするために、どのくらい支出すれば良いかを決定する問題であり、この問題をマルコフ過程における多段決定問題として定式化する。$(-\infty, \infty)$ を状態空間、s を状態、x を支出額とし、$f(s, x)$ を状態が s のとき、決定 x により移る状態とする。$C(x)$ を決定 x に伴う費用、$u(s)$ を状態が s のときの終端利得とする。$\boldsymbol{P} = (p_s(t))$ をマルコフ過程の推移法則とし、$T(s)$ を任意の状態 s に対して、$p_s(t)$ を密度関数とする確率変数とする。n が決定期間で、状態が s のとき、$v_n(s)$ を最適値とし、$x_n^*(s)$ を最適政策とする。

3 劣モジュラー関数と確率的凸性

3.1 劣モジュラー関数と優モジュラー関数

定義 1 s と x の関数 $f(s,x)$ が、$x<y$ および $s<t$ となる任意の x,y と s,t に対して

$$f(t,y)-f(t,x)\leq(\geq)f(s,y)-f(s,x) \tag{1}$$

のとき、劣モジュラ関数 (優モジュラ関数)(*submodular* (*supermodular*) *function*) という。

任意の $\boldsymbol{x},\boldsymbol{y}\in\mathcal{R}^n$ に対して $\boldsymbol{x}\vee\boldsymbol{y},\boldsymbol{x}\wedge\boldsymbol{y}$ を定義する。

$$\begin{aligned}\boldsymbol{x}\vee\boldsymbol{y} &= (\max\{x_1,y_1\},\cdots,\max\{x_n,y_n\}),\\ \boldsymbol{x}\wedge\boldsymbol{y} &= (\min\{x_1,y_1\},\cdots,\min\{x_n,y_n\})\end{aligned}$$

定義 2 X を $\mathcal{R}^n$ の部分集合で、n 変数関数を $f(\boldsymbol{x})$ とする。この関数 $f(\boldsymbol{x})$ が U で優モジュラー *(supermodular)* であるとは、任意の $\boldsymbol{x},\boldsymbol{x}'\in X$ に対して

$$f(\boldsymbol{x})+f(\boldsymbol{x}')\leq f(\boldsymbol{x}\vee\boldsymbol{x}')+f(\boldsymbol{x}\wedge\boldsymbol{x}'), \tag{2}$$

となることである。ただし、$\boldsymbol{x}\vee\boldsymbol{x}',\boldsymbol{x}\wedge\boldsymbol{x}'\in X$ とする。f が *strictly supermodular* であるとは、不等式 (2) が *strictly* に成り立つ $\boldsymbol{x}$ と $\boldsymbol{x}'$ が存在することである。関数 $f(\boldsymbol{x})$ が (*strictly*) 劣モジュラー *(submodular)* であるとは、$-f(\boldsymbol{x})$ が (*strictly*) 優モジュラー *(supermodular)* となることである。

任意の $i,j\in\{1,2,\cdots,n\}$ に対して

$$\hat{\boldsymbol{x}}_{ij}=(x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_{j-1},x_{j+1},\cdots,x_n)\in\mathcal{R}^{n-2},$$

とおき、任意の関数 r $f:\mathcal{R}^n\to\mathcal{R}$ にたいし

$$f_{\hat{\boldsymbol{x}}_{ij}}(x_i.x_j)=f(x_1,\cdots,x_{i-1},x_i,x_{i+1},\cdots,x_{j-1},x_j,x_{j+1},\cdots,x_n)$$

とおく。

$y\leq y'(y<y')$ となる任意の $y,y'\in\mathcal{R}$ に対し、$f(x,y')-f(x,y)$ が x に関して (strictly) increasing のとき、この 2 変数関数 f は (strictly) increasing differences を持つという。

$f_{\hat{\boldsymbol{x}}_{ij}}(x_i.x_j)$ が、任意の $i,j\in\{1,2,\cdots,n\}$ と $\hat{\boldsymbol{x}}_{ij}\in\mathcal{R}^{n-2}$ に対して increasing differences を持つとき、この n 変数関数 $f(\boldsymbol{x})$ は increasing differences を持つと定義する。n 変数関数 $f(\boldsymbol{x})$ が decreasing differences を持つ場合も同様に定義する。

定理 1 (Simchi-Levi, Chen, Bramel [9]) n 変数関数 $f(\boldsymbol{x})$ が *(strictly) supermodular* であることと、$f(\boldsymbol{x})$ が *(strictly) increasing differences* を持つことは同値である。

3.2 陰関数定理

X を $\mathcal{R}^3$ の部分集合とし、$F(x,y,z)$ を C^2 級関数とする。(a,b,c) を X の点で $F(a,b,c)=0$ となるものとする。$F_z(a,b,c)\neq 0$ ならば、(a,b) を含む X の部分集合で $\psi(a,b)=c$ となる C^2 級の陰関数 $z=\psi(x,y)$ が存在する。陰関数定理より

$$\frac{\partial z}{\partial x}=-\frac{F_x}{F_z},\frac{\partial z}{\partial y}=-\frac{F_y}{F_z}$$

および

$$\begin{aligned}\frac{\partial^2 z}{\partial x^2} &= -\frac{F_{xx}F_z^2-2F_{xz}F_xF_z+F_{zz}F_x^2}{F_z^3}\\ \frac{\partial^2 z}{\partial y^2} &= -\frac{F_{yy}F_z^2-2F_{yz}F_yF_z+F_{zz}F_y^2}{F_z^3}\\ \frac{\partial^2 z}{\partial x\partial y} &= -\frac{F_{xy}F_z^2-F_{xz}F_{yz}F_z-F_{yz}F_xF_z+F_{zz}F_xF_y}{F_z^3},\end{aligned}$$

となる。ただし $F_x(x,y,z)$, $F_y(x,y,z)$, $F_z(x,y,z)$, $F_{xx}(x,y,z)$, $F_{yy}(x,y,z)$, $F_{zz}(x,y,z)$, $F_{xy}(x,y,z)$, $F_{yz}(x,y,z)$, $F_{zx}(x,y,z)$ を、$F_x, F_y, F_z, F_{xx}, F_{yy}, F_{zz}, F_{xy}, F_{yz}, F_{zx}$ とおく。

C^2 級関数 $f(s,x)$ が s,x の凹関数で、$f(s,x)$ が厳密な意味で s,x の増加関数とする。このとき、$f_{ss}(s,x)<0$ である。また、$f(s,x)$ が厳密な意味で s,x の増加関数であり、$f_s(s,x)>0$ および $f_x(s,x)>0$ となる。このような C^2 級関数 $f(s,x)$ に対して $F(s,x,t)=f(s,x)-t$ とおく。(s,x,t) を X に含まれる $F(s,x,t)=0$ となる点とする。適当な条件の下で、C^2 級の陰関数 $s=\psi(x,t)$ が存在し

$$\frac{\partial s}{\partial x}=-\frac{F_x}{F_s}=-\frac{f_x}{f_s},\frac{\partial s}{\partial t}=-\frac{F_t}{F_s}=\frac{1}{f_s}$$

および

$$\begin{aligned}\frac{\partial^2 s}{\partial x^2} &= -\frac{f_{xx}f_s^2-2f_{xs}f_xf_s+f_{ss}f_x^2}{f_s^3}>0,\\ \frac{\partial^2 s}{\partial t^2} &= -\frac{f_{ss}}{f_s^3}>0,\quad \frac{\partial^2 s}{\partial x\partial t}=\frac{f_{ss}f_x}{f_s^3}<0\end{aligned}$$

となる。

すなわち、陰関数定理より C^2 級関数 $s=\psi(x,t)$ が存在し、$\psi_{xt}(x,t)<0$ および $\psi_x(x,t)<0, \psi_t(x,t)>0$ となる。

3.3 尤度比順序

X と Y を 2 つの確率変数とする。

定義 3 (TP_2) 確率密度関数 $f_X(x)$ と $f_Y(x)$ を持つ 2 つの確率変数 X と Y に対して、$x \geq y$ となる任意の x と y に対して、$f_X(y)f_Y(x) \leq f_X(x)f_Y(y)$ であるとき、X は Y より尤度比の意味で大きいといい、$X \geq_{LRD} Y$ あるいは $X \succeq Y$ と表す。

つぎの補題 1 は、よく知られた性質である。(Kijima and Ohnishi[1] など)

補題 1 $X \succeq Y$ ならば、非減少非負関数 $h(x)$ に対して $E[h(X)] \geq E[h(X)]$ となる。

3.4 確率的凸性と凹性

$\{X(s)|s \in \Theta\}$ を s をパラメータとする確率変数列とする。ただし、Θ は $(-\infty,\infty)$ または $(-\infty,\infty)$ に含まれる凸集合とする。

(1) $\{X(s)|s \in (-\infty,\infty)\}$ が SI(stocahstically increasing) であるとは、任意の増加 関数 $u(s)$ に対して、$E[u(X(s))]$ が、s の増加関数となることである。

(2) $\{X(s)|s \in (-\infty,\infty)\}$ が SICX(stocahstically increasing and convex) であるとは、任意の増加凸関数 $u(s)$ に対して、$E[u(X(s))]$ が、s の増加凸 関数となることである。

(3) $\{X(s)|s \in (-\infty,\infty)\}$ が SICV(stocahstically increasing and concave) であるとは、任意の増加凹関数 $u(s)$ に対して、$E[u(X(s))]$ が、s の増加凹関数となることである。

つぎに、$s_1 \leq s_2 \leq s_3 \leq s_4$ で $s_1+s_4=s_3+s_2$ のとき、$X_i=X(s_i)$ とおく $(i=1,2,3,4)$。$(s_4-s_3=s_2-s_1)$

定義 4 (1) $\{X(s)|s \in \Theta\}$ が *SICX(sp)(stocahstically increasing and convex in sample path sense)* であるとは、$\max\{X_2,X_3\} \leq X_4$ であり *(a.s.)*、$X_2+X_3 \leq X_1+X_4$ となることである。

(2) $\{X(s)|s \in \Theta\}$ が *SICV(sp)(stocahstically increasing and concave in sample path sense)* であるとは、$X_1 \leq \max\{X_2, X_3\}$ であり *(a.s.)*、$X_2 + X_3 \geq X_1 + X_4$ となることである。

補題 2 (1) $\{X(s)|s \in \Theta\}$ が *SICX(sp)* ならば、*SICX* である。
(2) $\{X(s)|s \in \Theta\}$ が *SICV(sp)* ならば、*SICV* である。

4 逐次支出問題と最適政策

4.1 支出問題とマルコフ決定過程

n を決定期間、s を過程の状態とし、$u(s)$ を終端利得、$C(x)$ を決定 x に対する費用とする。$v_n(s)$ を最適値とし、$x_n^*(s)$ を最適決定とすれば、最適方程式は、

$$v_n(s) = \max_{x \geq 0}\{-C(x) + E[v_{n-1}(T(f(s,x)))]\}, \tag{3}$$

となる。ただし、$v_1(s) = \max_{x \geq 0}\{-C(x) + E[u(T(f(s,x)))]\}$ とする。

仮定 1 (1) $u(s)$ は s の増加凹関数
(2) $C(x)$ は x の増加凸関数で $C(0) = 0$ とする
(3) $f(s,x)$ は s, x の劣モジュラー凹関数で $f(s,0) = s$ とし、s と x の厳密な増関数
(4) $s \leq s'$ となる s, s' に対して $T(s') \succeq T(s)$ とする
(5) $\{T(s)|s \in (-\infty, \infty)\}$ は SICV とする

決定と推移の順序は、次のように考える。状態が s のとき、決定 x をとり、状態はこの決定により $f(s,x)$ となる。つぎに、推移法則 $\boldsymbol{P}$ にしたがって状態が推移し、状態は $T(f(s,x))$ となる。

$$\overline{v}(s) = \max_{x \geq 0}\{-C(x) + u(f(s,x))\}$$

とおく。このとき、$u(s)$ が s の増加関数であれば、$\overline{v}(s)$ も増加関数である。$C(x)$ が凸関数のとき、$u(s)$ が凹関数ならば、$\overline{v}(s)$ も凹関数である。ただし、$C(x)$ は増加関数とする。また、$v_n(s)$ は、s に関する増加凹関数である。

性質 1 仮定 1 のもとで、$x_n^*(s)$ は s に関して減少する。

仮定 2 $t \geq s$のとき凹関数$u(s)$に対し、$E[u(T(t))] - E[u(T(s))] \leq u(t) - u(s)$である。

仮定2より、任意の$n \geq 1$に対して

$$E[v_n(T(t))] - E[v_n(T(s))] \leq E[v_{n-1}(T(t))] - E[v_{n-1}(T(s))] \quad (4)$$

性質 2 仮定2のもとで、$x_n^*(s)$ は nに関して減少する。

5 部分観測可能なマルコフ過程

5.1 部分観測可能なマルコフ過程と情報

状態空間を$(-\infty, \infty)$とするマルコフ過程を考える。状態は部分観測可能なマルコフ過程にしたがって推移する。それぞれの状態sに対して確率変数Y_sを考え、観測課程とし、これらの確率変数を通して情報を得る。これらのYを観測し、ベイズ学習にしたがって情報を改良する。推移法則$\boldsymbol{P} = (p_s(t))_{s,t\in(-\infty,\infty)}$は$Y_s$とは独立であり、情報の集合を$\mathcal{S}$とする。

$\boldsymbol{\mu}$を事前情報として、状態空間上の確率分布とするとき、$\overline{\boldsymbol{\mu}}$を推移法則に従って推移したあとの状態空間上の確率分布とし、$\boldsymbol{\mu}_y$をyを観測したあと、ベイズの定理にしたがって改良した事後情報、$\boldsymbol{\mu}^x$を決定xを取ったあとの状態空間上の確率分布とする。

情報のあいだには、LRD $(\geq_{LRD})$に基づく順序関係を仮定する。さらに、Y_sの分布関数を$F_s(y)$とし、$s \leq t$ならば、$Y_t \geq_{LRD} Y_s$である$(s, t \in (-\infty, \infty))$。

5.2 事前事後情報

観測、決定、推移の順序は、次のように考える。事前情報が$\boldsymbol{\mu}$のとき、値yを観測し、ベイズの定理にしたがって情報を$\boldsymbol{\mu}_y \in \mathcal{S}$と改良する。つぎに、決定$x$をとり、情報は$\boldsymbol{\mu}_y^x$となる。最後に、状態が推移し、事後情報は$\overline{\boldsymbol{\mu}_y^x}$となる。これら学習、決定、推移の順序を変えても同様の結果が得られる。

定義 5

任意の$s \in \Re$と$x \in \Re$の非負集合値関数$\boldsymbol{h}(x) = (h(x,s))_{s\in(-\infty,\infty)}$に対して、任意の$t$と$s$$(s \leq t$かつ$s, t \in (-\infty, \infty))$について、$x < y$ならば$\boldsymbol{h}(y) \geq_{LRD} \boldsymbol{h}(x)$とする。すなわち$h(x,t)\,h(y,s) \leq h(x,s)\,h(y,t)$である。このとき、関数$\boldsymbol{h}(x,s)$を$x$に関する増加関数という。

事前情報 $\boldsymbol{\mu}$ と事後情報 $\overline{\boldsymbol{\mu}^x}$ の関係について、Nakai[2, 4] などより、補題3が仮定1の (4) のもとで成り立つことが知られている。

補題 3 $\boldsymbol{\mu} \geq_{LRD} \boldsymbol{\nu}$, $y < y'$ とする。$\overline{\boldsymbol{\mu}} \succ \overline{\boldsymbol{\nu}}$ である。任意の y に対して、$\boldsymbol{\mu}_y \geq_{LRD} \boldsymbol{\nu}_y$ および $\overline{\boldsymbol{\mu}_y} \geq_{LRD} \overline{\boldsymbol{\nu}_y}$ である。任意の $\boldsymbol{\mu}$ に対して、$\boldsymbol{\mu}_{y'} \geq_{LRD} \boldsymbol{\mu}_y$ および $\overline{\boldsymbol{\mu}_{y'}} \geq_{LRD} \overline{\boldsymbol{\mu}_y}$ である。

5.3 正規分布の場合

不完備情報のマルコフ決定問題として最適支出問題を正規分布の場合に考える。状態に関する情報を $N(\mu, \sigma^2)$ とすれば、密度関数は

$$\mu(s) = \phi_{\mu,\sigma^2}(s)$$

である。ここで、$\phi_{\mu,\sigma^2}(x)$ を正規分布 $N(\mu, \sigma^2)$ の密度関数とおく。

それぞれの状態 s に確率変数 X_s が対応し、情報過程とする。その密度関数を

$$f_s(x) = \phi_{s,\sigma_1^2}(x)$$

とする。このとき、観測値 y が得られたときにベイズ学習にしたがえば、事後情報の密度関数は

$$\mu_y(s) = \psi\left(s \middle| \frac{\sigma^2 y + \sigma_1^2 \mu}{\sigma_1^2 + \sigma^2}, \frac{\sigma^2 \sigma_1^2}{\sigma_1^2 + \sigma^2}\right)$$

となる。推移法則 $\boldsymbol{P} = (p_s(t))_{s,t\in(-\infty,\infty)}$ を $p_s(t) = \phi_{s,\sigma_2^2}(t)$ とすれば、事前情報が $\mu(s) = \phi_{\mu,\sigma^2}(s)$ のとき、推移によって情報は

$$\overline{\mu}(t) = \phi_{\mu,\sigma_2^2+\sigma^2}(t)$$

となる。補題3の性質はこれらの状況で成り立つ。状態に関する情報が $\boldsymbol{\mu}$ のとき、x を追加して支出すれば、情報は

$$\mu^x(t) = \phi_{\mu,\sigma^2}(\psi(t, x))$$

となる。ここで、$f(\psi(t,x), x) = t$ とする。ここで、$x' > x, t' > t$ のとき、$(\psi(x', t') - \psi(x', t))(\psi(x', t) + \psi(x', t') - 2\mu) < (\psi(x, t') - \psi(x, t))(\psi(x, t) + \psi(x, t') - 2\mu)$ である。よって、

$$\frac{\mu^{x'}(t')}{\mu^{x'}(t)} > \frac{\mu^x(t)}{\mu^x(t)}$$

すなわち、$\boldsymbol{\mu}^{x'} \geq \boldsymbol{\mu}^x$ が成り立つ。

性質 3 この節の条件の下で、$y > y'$ ならば任意の x に対して $\boldsymbol{\mu}_y^x \succeq \boldsymbol{\mu}_{y'}^x$ および $\overline{\boldsymbol{\mu}_y^x} \succeq \overline{\boldsymbol{\mu}_{y'}^x}$ である。$\boldsymbol{\mu} \succeq \boldsymbol{\nu}$ ならば、任意の y に対して $\boldsymbol{\mu}_y \succeq \boldsymbol{\nu}_y$、$\boldsymbol{\mu}_y^x \succeq \boldsymbol{\nu}_y^x$、$\overline{\boldsymbol{\mu}_y^x} \succeq \overline{\boldsymbol{\nu}_y^x}$ となる。$x > x'$ ならば、任意の y に対して $\boldsymbol{\mu}_y^x \succeq \boldsymbol{\mu}_y^{x'}$ および $\overline{\boldsymbol{\mu}_y^x} \succeq \boldsymbol{\mu}_y^{x'}$ である。

6 部分観測可能なマルコフ過程での多段支出問題

状態空間を $(-\infty, \infty)$ とするマルコフ過程を考え、状態 s はアウトカムの指標とする。状態は部分観測可能なマルコフ過程にしたがって推移し、それぞれの状態 s $(s \in (-\infty, \infty))$ に対して確率変数 Y_s が存在し、観測課程とする。この観測できない状態に関する情報は、これらの確率変数 Y を観測し、ベイズ学習にしたがって情報を改良する。指標を改良するため、支出を行う。指標を改良するため、どのくらい支出を行えば良いかを決定する問題である。

n を計画期間とし、x を決定 $(0 \leq x \leq K)$ とする。$c(x)$ を決定 x に伴う費用とし、$f(s,x)$ を状態が s のとき、決定 x を取ったあとでの状態を表す関数とする。$\boldsymbol{\mu}$ を事前情報とし、このとき最適に振る舞って得られる最適値を $v_n(\boldsymbol{\mu})$ とする。

このとき、最適方程式は

$$\begin{aligned} v_n(\boldsymbol{\mu}) &= E[v_n(\boldsymbol{\mu}|Y)] \\ v_n(\boldsymbol{\mu}|y) &= \max_{x \geq 0} \left\{ -c(x) + v_{n-1}(\overline{\boldsymbol{\mu}_y^x}) \right\} \end{aligned} \tag{5}$$

となる。ただし、S を決定過程の状態を表す確率変数とすれば、$v_0(\boldsymbol{\mu}) = E_{\boldsymbol{\mu}}[u(S)]$ とする。

このとき、Nakai[3] や [4] と同様の仮定の下で、性質 3 が成り立ち、次の性質が得られる。

性質 4 $\boldsymbol{\mu} \geq_{LRD} \boldsymbol{\nu}$ ならば $v_n(\boldsymbol{\mu}) \geq v_n(\boldsymbol{\nu})$ である。

参考文献

[1] M. Kijima and M. Ohnishi: Stochastic Orders and Their Applications in Financial Optimization, *Math. Methods of Oper. Res.*, **50**, 351–372, (1999).

[2] T. Nakai, A Generalization of Multivariate Total Positivity of Order Two with an Application to Bayesian Learning Procedure, *Journal of Information & Optimization Sciences*, vol. 23, 163–176, 2002.

[3] T. Nakai, A Sequential Expenditure Problem for Public Sector Based on the Outcome, *Recent Advances in Stochastic Operations Research* (Eds. T. Dohi, S. Osaki and K. Sawaki), World Scientific Publishing, 277–295, 2007.

[4] T. Nakai, A Sequential Decision Problem based on the Rate Depending on a Markov Process, *Recent Advances in Stochastic Operations Research 2* (Eds. T. Dohi, S. Osaki and K. Sawaki), World Scientific Publishing, 11–30, 2009.

[5] 中井　達, 多段決定問題と Stochastic Convexity について, 京都大学数理解析研究所講究録「不確実・不確定環境下における数理的意思決定とその周辺」, vol. 1802, 193–199, 2012.7.

[6] 中井　達, 投資モデルに基づく逐次決定問題について, 京都大学数理解析研究所講究録「決定過程に関わる数理モデルの新たな展開と応用」, vol. 1857, 109–120, 2013.10.

[7] Shaked, M. and Shanthikumar, J. G., *Stochastic Orders and Their Applications* (Probability and mathematical statistics : a series of monographs and textbooks), Academic Press, Boston, Massachusetts, 1994.

[8] White, D. J. Structural properties for contracting state partially observable Markov decision processes. J. Math. Anal. Appl. 186 (1994), 486–503

[9] David Simchi-Levi, Xin Chen, Julien Bramel, *Convexity and Supermodularity*, The Logic of Logistics, Theory, Algorithms, and Applications for Logistics and Supply Chain Management, Springer Series in Operations Research, 2005, pp 13-32