数理解析研究所講究録
第726巻 1990年 5-12

複数の母集団を持つTwo-armed Bandit Problem について

（両方のarmが未知の場合）

姫路短期大学　濱田年男　(Toshio Hamada)

1．緒言

各期においてm個の条件$C_1, C_2, \cdots, C_m$のいずれか1つが成立し，その条件のもとで2つの行動a_1，a_2のいずれかが実行可能であるとする．条件C_iのもとでa_k (k=1,2)をおこなったときに分布$F(z \mid u_{ki})$にしたがう観察値を得るものとする．ここで，パラメータu_{ki}は未知であり，事前知識として共役事前分布$G(u_{ki} \mid x_{ki})$が与えられているものとする．また条件C_iのもとでa_kを行って観察値zを得たときの期待利得はh(z)で与えられるものとする．また，ある期に条件C_iが成立しているときに，次の期に条件C_jが成立する確率は，選択した行動に無関係にp_{ij}であるとする．一期先の利得の割引率β（$0<\beta\leqq 1$）倍が利得の現在価値であるとする．目的はn期間の期待割引総利得を最大にすることであり，そのためには各期においていずれの実験を行えばよいかを決定することである．

m=1の場合は従来のtwo-armed bandit problemであり，多くの研究がなされてきている．たとえば，Bradt, Johnson and Karlin [2], Zacks [9], Kelley [6], Yakowitz [8], Berry [1], Jones [4], Kalin and Theodorescu [5], Hamada [3], Kolonko and Benzing [7], などがある．

$m \geqq 2$の場合の例として，ある町に転居してきた人が，次の日から毎朝，通勤のため家から駅まで行くものとする．選択できる行動として，バス停まで歩いてバスに乗り駅前で降りるのと，自転車で駅前まで行くという２とおりの方法があるものとする．前者をa_1，後者をa_2とする．天候および道路の混雑している状態に依存して，所用時間が変わってくる．天候としては，C_1が晴れ，C_2が雨，…，C_mが雪，というように m種類あるものとする．C_i $(i=1,2,\cdots,m)$のときa_k $(k=1,2)$を用いたときの所用時間zの分布が$F(z \mid u_{ki})$であるとする．u_{ki}の値は未知であり，近所の人からの情報等により事前知識が与えられているものとする．この人は毎朝天気を見ていずれの行動を選択するかを決定する．

２．動的計画法による定式化

$x_1=(x_{11},x_{12},\cdots,x_{1m})$, $x_2=(x_{21},x_{22},\cdots,x_{2m})$とする．現在の条件が$C_i$で事前分布のパラメータが$(x_1,x_2)=(x_{11},x_{12},\cdots,x_{1m},x_{21},x_{22},\cdots,x_{2m})$である状態を$(i;x_1,x_2)$で表す．また$\Phi_{iz}x_1=(x_{11},x_{12},\cdots,\phi_z x_{1i},\cdots,x_{1m})$，$\Phi_{iz}x_2=(x_{21},x_{22},\cdots,\phi_z x_{2i},\cdots,x_{2m})$とする．さらに，$R(x_{ki})=E[h(Z) \mid x_{ki}]$ $(k=1,2;\ i=1,2,\cdots,m)$とする．ここで次の３つの仮定を行う．

A_1：任意の$x_{ki} \in \Omega$ $(k=1,2;\ i=1,2,\cdots,m)$に対して$R(x_{ki})>0$

A_2：任意の$x_{ki} \in \Omega$ $(k=1,2;\ i=1,2,\cdots,m)$に対して

$$E[R(\phi_z x_{ki}) \mid x_{ki}]=R(x_{ki})$$

A_3：任意の$c>0$，$k=1,2$，$i=1,2,\cdots,m$に対して$0<R(x_{ki})<c$であるようなx_{ki}，および$c<R(x_{ki})$であるようなx_{ki}が存在する．

いま$V_n(i;x_1,x_2)$は現在の状態が$(i;x_1,x_2)$であり，残りn期間ある時に以後最適政策を用いることにより得られる最大期待割引総利得とする．このとき，次のように定式化できる．n=1,2,3,…に対して

$$V_n(i;x_1,x_2)=\max\{V_n^1(i;x_1,x_2),\ V_n^2(i;x_1,x_2)\} \tag{1}$$

但し

$$V_0(i;x_1,x_2)=0 \tag{2}$$

ここに

$$V_n^1(i;x_1,x_2)=E[h(Z)+\beta\sum_{j=1}^{m}p_{ij}V_{n-1}(j;\Phi_{iZ}x_1,x_2)\mid x_{1i}] \tag{3}$$

$$V_n^2(i;x_1,x_2)=E[h(Z)+\beta\sum_{j=1}^{m}p_{ij}V_{n-1}(j;x_1,\Phi_{iZ}x_2)\mid x_{2i}] \tag{4}$$

(3),(4)はそれぞれつぎのように書き直せる．

$$V_n^1(i;x_1,x_2)=R(x_{1i})+\beta\sum_{j=1}^{m}p_{ij}E[V_{n-1}(j;\Phi_{iZ}x_1,x_2)\mid x_{1i}] \tag{5}$$

$$V_n^2(i;x_1,x_2)=R(x_{2i})+\beta\sum_{j=1}^{m}p_{ij}E[V_{n-1}(j;x_1,\Phi_{iZ}x_2)\mid x_{2i}] \tag{6}$$

いま，$V_n^1(i;x_1,x_2)\geqq V_n^2(i;x_1,x_2)$ならば$a_1$が最適である．また$V_n^1(i;x_1,x_2)\leqq V_n^2(i;x_1,x_2)$ならば$a_2$が最適である．さらにまた$V_n^1(i;x_1,x_2)=V_n^2(i;x_1,x_2)$ならば$a_1,a_2$ともに最適である．

$V_n(i;x_1,x_2)$，$V_n^1(i;x_1,x_2)$，$V_n^2(i;x_1,x_2)$に対して次の性質が得られる．

[補題1] $(x_1,x_2)\in\Omega^{2m}$，$1\leqq i\leqq m$，$1\leqq j\leqq m$，$n\geqq 1$に対して，

(i) $E[V_n(i;\Phi_{iZ}x_1,x_2)\mid x_{1i}]\geqq V_n(i;x_1,x_2)$

(ii) $E[V_n(i;x_1,\Phi_{i2}x_2)\mid x_{2i}]\geqq V_n(i;x_1,x_2)$.

[補題２] $(x_1,x_2)\in\Omega^{2m}$, $1\leqq k\leqq 2$, $1\leqq i\leqq m$, $n\geqq 1$に対して,

(i) $V_n^k(i;x_1,x_2)\geqq R(x_{k2i})+\sum_{t=2}^{n}\beta^{t-1}\sum_{j=1}^{m}p\{j^{-1}\}\{R(x_{1j})\vee R(x_{2j})\}$

(ii) $V_n(i;x_1,x_2)\geqq \sum_{t=1}^{n}\beta^{t-1}\sum_{j=1}^{m}p\{j^{-1}\}\{R(x_{1j})\vee R(x_{2j})\}$.

[補題３] $(x_1,x_2)\in\Omega^{2m}$, $1\leqq k\leqq 2$, $1\leqq i\leqq m$, $n\geqq 1$に対して,

(i) $V_n^k(i;x_1,x_2)\leqq R(x_{ki})+\sum_{t=2}^{n}\beta^{t-1}\sum_{j=1}^{m}p\{j^{-1}\}\{R(x_{1j})+R(x_{2j})\}$

(ii) $V_n(i;x_1,x_2)\leqq \sum_{t=1}^{n}\beta^{t-1}\sum_{j=1}^{m}p\{j^{-1}\}\{R(x_{1j})+R(x_{2j})\}$.

[補題４] $(x_1,x_2)\in\Omega^{2m}$, $1\leqq k\leqq 2$, $1\leqq i\leqq m$, $n\geqq 1$に対して,

(i) $\sum_{t=1}^{n}\beta^{t-1}\sum_{j=1}^{m}p\{j^{-1}\}R(x_{1j})<R(x_{2i})$ならば$V_n^1(i;x_1,x_2)<V_n^2(i;x_1,x_2)$

(ii) $\sum_{t=1}^{n}\beta^{t-1}\sum_{j=1}^{m}p\{j^{-1}\}R(x_{2j})<R(x_{1i})$ならば$V_n^1(i;x_1,x_2)>V_n^2(i;x_1,x_2)$.

いま, $\Xi=\{1,2\}\times\{1,2\}\times\cdots\times\{1,2\}$とし, sをm次元行ベクトルとする. また, $V_n^1(i;x_1,x_2)\geqq V_n^2(i;x_1,x_2)$ならば$\delta_n(i;x_1,x_2)=1$, $V_n^1(i;x_1,x_2)<V_n^2(i;x_1,x_2)$ならば$\delta_n(i;x_1,x_2)=2$とする. さらに, $Q_n(s)=\{(x_1,x_2)\mid s=(s_1,s_2,\cdots,s_m),\ \delta_n(i;x_1,x_2)=s_i,\ 1\leqq i\leqq m\}$とする.

[定理１] 任意の$s\in\Xi$, $n\geqq 1$に対して$Q_n(s)\neq\phi$である.

３. 最大期待割引総利得の分割

[補題５] $(x_1,x_2)\in\Omega^{2m}$, $1\leqq i\leqq m$, $n\geqq 1$に対して, $V_n(i;x_1,x_2)$は次のように分割

される.

$$V_n(i;x_1,x_2)=\sum_{j=1}^{m}W_n(i;j;x_{1j},x_{2j})$$

ここに$W_n(i;j;x_{1i},x_{2i})$は次のように再帰的に定義される.

$$W_n(i;i;x_{1i},x_{2j})=\max\{R(x_{1i})+\beta\sum_{s=1}^{m}p_{is}E[W_{n-1}(s;i;\phi_z x_{1i},x_{2i})\mid x_{1i}],$$

$$R(x_{2i})+\beta\sum_{s=1}^{m}p_{is}E[W_{n-1}(s;i;x_{1i},\phi_z x_{2i})\mid x_{2i}]\}$$

および$j\neq i$に対して

$$W_n(i;j;x_{1j},x_{2j})=\beta\sum_{s=1}^{m}p_{is}W_{n-1}(s;j;x_{1j},x_{2j}). \tag{7}$$

いま,

$$W_n^1(i;i;x_{1i},x_{2i})=R(x_{1i})+\beta\sum_{s=1}^{m}p_{is}E[W_{n-1}(s;i;\phi_z x_{1i},x_{2i})\mid x_{1i}], \tag{8}$$

$$W_n^2(i;i;x_{1i},x_{2i})=R(x_{2i})+\beta\sum_{s=1}^{m}p_{is}E[W_{n-1}(s;i;x_{1i},\phi_z x_{2i})\mid x_{2i}]\} \tag{9}$$

とおくと，次の補題が成立する.

[補題６] $(x_1,x_2)\in\Omega^{2m}$, $1\leqq i\leqq m$, $n\geqq 1$に対して,

$$V_n^1(i;x_1,x_2)-V_n^2(i;x_1,x_2)=W_n^1(i;i;x_{1i},x_{2i})-W_n^2(i;i;x_{1i},x_{2i})$$

この補題により，$V_n^1(i;x_1,x_2)-V_n^2(i;x_1,x_2)$の符号を調べる代わりに,

$W_n^1(i;i;x_{1i},x_{2i})-W_n^2(i;i;x_{1i},x_{2i})$の符号を調べれば十分であることがわかる.

４．最適政策

いま，f_i^tは条件C_iが成立してからt期後に初めて条件C_iが再び成立する確率を

表すものとする．このとき，次の補題が成立する．

［補題７］ $1 \leqq i \leqq m$, $(x_{1i}, x_{2i}) \varepsilon \Omega^2$, $n \geqq 1$に対して，

$$(i)\quad W_n^1(i;i;x_{1i},x_{2i}) = R(x_{1i}) + \sum_{t=2}^{n} \beta^{t-1} f_i^{(t-1)} E[W_{n-t+1}(i;i;\phi_z x_{1i}, x_{2i}) \mid x_{1i}]$$

$$(ii)\quad W_n^2(i;i;x_{1i},x_{2i}) = R(x_{2i}) + \sum_{t=2}^{n} \beta^{t-1} f_i^{(t-1)} E[W_{n-t+1}(i;i;x_{1i}, \phi_z x_{2i}) \mid x_{2i}].$$

いま，

$$D_n(i;x_1,x_2) = W_n^1(i;i;x_{1i},x_{2i}) - W_n^2(i;i;x_{1i},x_{2i})$$

とおくと，$D_n(i;x_{1i},x_{2i}) \geqq 0$ ならばa_1が最適であり，$D_n(i;x_{1i},x_{2i}) \leqq 0$ ならばa_2が最適であり，$D_n(i;x_{1i},x_{2i}) = 0$ ならばa_1, a_2ともに最適である．いま

$$D_n^+(i;x_{1i},x_{2i}) = \max\{D_n(i;x_{1i},x_{2i}),\ 0\}$$

$$D_n^-(i;x_{1i},x_{2i}) = \min\{D_n(i;x_{1i},x_{2i}),\ 0\}$$

とおくと

$$W_n^1(i;i;x_1,x_2) = W_n(i;i;x_{1i},x_{2i}) + D_n^-(i;x_{1i},x_{2i}) \qquad (10)$$

$$W_n^2(i;i;x_1,x_2) = W_n(i;i;x_{1i},x_{2i}) - D_n^+(i;x_{1i},x_{2i}) \qquad (11)$$

［補題８］ $1 \leqq i \leqq m$, $(x_{1i}, x_{2i}) \varepsilon \Omega^2$, $n \geqq 1$に対して，

$$D_n(i;x_{1i},x_{2i}) = \{R(x_{1i}) - R(x_{2i})\}\{1 - \sum_{t=2}^{n} \beta^{t-1} f_i^{(t-1)}\}$$

$$+ \sum_{t=2}^{n} \beta^{t-1} f_i^{(t-1)} \{E[D_{n-t+1}^+(i;i;\phi_z x_{1i}, x_{2i}) \mid x_{1i}]$$

$$+ E[D_{n-t+1}^-(i;i;x_{1i}, \phi_z x_{2i}) \mid x_{2i}]\}.$$

[補題９] $1 \leqq i \leqq m$, $(x_{1i}, x_{2i}) \in \Omega^2$, $n \geqq 1$に対して,

$$R(x_{1i}) - c_n R(x_{2i}) \leqq D_n(i; x_{1i}, x_{2i}) \leqq c_n R(x_{1i}) - R(x_{2i})$$

ここにc_nは1以上の正の定数であり，次のように再帰的に定義される.

$$c_1 = 1$$

であり，$n \geqq 2$に対して,

$$c_n = 1 + \sum_{t=2}^{n} \beta^{t-1} f_i^{(t-1)} (2c_{n-t+1} - 1)$$

である.

[定理２] $1 \leqq i \leqq m$, $(x_{1i}, x_{2i}) \in \Omega^2$, $n \geqq 1$に対して,

(i) $R(x_{1i}) > c_n R(x_{2i})$ならば $D_n(i; x_{1i}, x_{2i}) > 0$

(ii) $R(x_{1i}) < c_n^{-1} R(x_{2i})$ならば $D_n(i; x_{1i}, x_{2i}) < 0$.

この定理により，現在の状態が$(i; x_1, x_2)$であり，残り期間がnのときに，もし$R(x_{1i})/R(x_{2i}) > c_n$ならば$a_1$が最適であり，またもし$R(x_{1i})/R(x_{2i}) < c_n^{-1}$ならば$a_2$が最適である．しかし$c_n^{-1} \leqq R(x_{1i})/R(x_{2i}) \leqq c_n$のときには，$a_1, a_2$のいずれが最適であるかはわからない.

参考文献

[1] Berry, D. A. (1972). A Bernoulli two-armed bandit. Ann. Math. Statist. 43, 871-897.

[2] Bradt, R. N., Johnson, S. M. and Karlin, S. (1956). On sequential designs for maximizing the sum of n observations. Ann. Math. Statist.

27, 1060-1074.

[3] Hamada, T. (1984). On a uniform two-armed bandit problem. J. Japan Statist. Soc. 14, 179-197.

[4] Jones, P. (1976). Some results for the two-armed bandit problem. Math. Operationsforsch. u. Statist. 7, 471-475.

[5] Kalin, D. and Theodorescu, R. (1982). A note on structural properties of the Bernoulli two-armed bandit problem. Math. Operationsforsch. Statist., Ser. Optimization 13, 469-472.

[6] Kelley, T. A. (1974). A note on the Bernoulli two-armed bandit problem. Ann. Statist. 2, 1056-1062.

[7] Kolonko, M. and Benzing, H. (1985) On monotone optimal decision rules and stay-on-a-winner rule for the two-armed bandit. Metrika 32, 395-407.

[8] Yakowitz, S. J. (1969). Mathematics of Adaptive Control Processes. American Elsevier Publishing Co., New York.

[9] Zacks, S. (1967). Bayes sequential strategies for crossing a field containing absorption points. Naval Res. Logist. Quart. 14, 329-343.