数理解析研究所講究録
899巻 1995年 181-188

ON A GAMMA TWO-ARMED BANDIT PROBLEM WITH ONE ARM KNOWN

神戸商科大学　　濱田年男
(Toshio Hamada)

1．緒言

次のような問題を考える．２つの実験A_0とA_1があり，A_0またはA_1のいずれかを行うことにより，パラメータ$(\beta,1)$および(γ,u)のガンマ分布に従う観察値が得られるものとする．βとγの値は既知であるが，uの値は未知であり，uの真の値に対する事前分布として，パラメータ(w,α)のガンマ分布が与えられているものとする．n期間の問題を考え，各期においてA_0またはA_1のいずれかを行い，観察値を得た後に次の期には，またA_0またはA_1のいずれかを行うものとする．ある期においてA_1を行うことにより，uの値に対する事前分布は事後分布へと更新される．目的はn期間の観察値の総和の期待値を最大にすることである．そのためには，各期において，その前の期までに得られている情報を用いて，いずれの実験を行えばよいかを決定することである．この問題を$(P_{\beta,\gamma})$と呼ぶことにする．

この種の問題は２腕のバンディット問題と呼ばれ，これまでに多くの研究結果が報告されている．たとえば，ガンマ分布の代わりにベルヌーイ分布を用いた場合には，事前分布としてベータ分布を仮定し，過去40年余りにわたって研究されてきている．たとえば，Bradt, Johnson and Karlin [3], Bellman [1], Yakowitz [11], Berry and Fristedt [2], Gittins [4], 等，また区間(0,1)と(0,u)の一様分布の場合には，未知のuに対して，Pareto分布が事前分布として仮定され，Woodroofe [10], Kalin and Theodorescu [9], Hamada [5,6,7] 等において研究されている．特に[5]においては，動的計画法により定式化し，再帰方程式を解くことにより，最適解を与える境界値の求め方と，それを用いた最適解の構造が与えられている．また，A_0とA_1が指数分布に従う実験の場合にも，未知パラメータの事前分布としてガンマ分布を仮定する場合には，一様分布の場合よりはやや複雑であるが，同様な再帰方程式を解くことができ，最適解を得られることがわかっている[6]．しかし，A_0とA_1がガンマ分布に従う実験の場合には，再帰方程式を解くのはかなり複雑になり，不可能のように思われた．

一方，濱田 [8] において，問題(Q_m)として，２種類の実験e_0とe_1のいずれかを行う

ことにより，パラメータ１及びｕの指数分布から観察値を得るものとし，ｕの値は未知で事前分布として，パラメータ(w, α)のガンマ分布を仮定できる場合に，e_0またはe_1のいずれか一方のみをｍ回行うことができる場合を考え，$m=2$の場合の解を与えた．

本研究では，２節において問題$(P_{\beta,\gamma})$を動的計画法により定式化し，最適解の構造を明らかにする．また，３節では問題(Q_m)の動的計画法による定式化と，最適解の諸性質について論じ，４節では最適解を表す境界値の求め方について論じる．５節では$\beta=\gamma=m$の場合に問題$(P_{\beta,\gamma})$と問題(Q_m)が等価であることを証明し，さらに６節では問題$(P_{\gamma,\gamma})$の近似解を与える．

２．問題$(P_{\beta,\gamma})$の動的計画法による定式化

問題$(P_{\beta,\gamma})$を動的計画法によって定式化する．まず

$G_n(\beta,\gamma,w,\alpha)$：残りｎ期間あり，ｕについての事前分布がパラメータ(w,α)のガンマ分布のとき，以後最適政策を用いて得られる観察値の和の期待値の最大値

とおくと，

$$G_n(\beta,\gamma,w,\alpha)=\max\{G_n^0(\beta,\gamma,w,\alpha),\ G_n^1(\beta,\gamma,w,\alpha)\} \tag{1}$$

$$(n=1,2,3,\cdots;\ G_0(\beta,\gamma,w,\alpha)=0)$$

ここに

$$G_n^0(\beta,\gamma,w,\alpha)=\beta+G_{n-1}(\beta,\gamma,w,\alpha) \tag{2}$$

および

$$G_n^1(\beta,\gamma,w,\alpha)=\int_0^\infty\int_0^\infty\{z+G_{n-1}(\beta,\gamma,w+z,\alpha+\gamma)\}\phi(z|u,\gamma)\phi(u|w,\alpha)dzdu \tag{3}$$

である．ここに

$$\phi(u|w,\alpha)=\{\Gamma(\alpha)\}^{-1}w^\alpha u^{\alpha-1}e^{-wu}$$

である．ここで，(3)は

$$G_n^1(\beta,\gamma,w,\alpha)=\gamma w(\alpha-1)+\int_0^\infty G_{n-1}(\beta,\gamma,w+z,\alpha+\gamma)\phi(z|w,\alpha,\gamma)dz \tag{4}$$

となる．ここに，

$$\phi(z|w,\alpha,\gamma)=\{\Gamma(\alpha)\Gamma(\gamma)\}^{-1}\Gamma(\alpha+\gamma)w^\alpha z^{\gamma-1}(w+z)^{-\alpha-\gamma}$$

である.

［定理１］状態 $(n,\beta,\gamma,w,\alpha)$において，$A_0$が最適ならば，以後$A_0$ばかり行うのが最適である．よって，$G_n(\beta,\gamma,w,\alpha)=G_n^0(\beta,\gamma,w,\alpha)$ならば，$G_n(\beta,\gamma,w,\alpha)=n\beta$である.

この定理により，(1)式は次のように書き換えることができる.

$$G_n(\beta,\gamma,w,\alpha)=\max\{n\beta,\ G_n^1(\beta,\gamma,w,\alpha)\} \tag{5}$$

さらに次の補題が得られる.

［補題１］$G_n(\beta,\gamma,w,\alpha)$は次の性質を満たす.

(i) wについて連続で，単調増加である.

(ii) αについて連続で，単調減少である.

(iii) γについて連続で，単調増加である.

(iv) βについて連続で，非減少である.

（証明）帰納法による．□

［補題２］$G_n(\beta,\gamma,w,\alpha)\leqq n\beta+n\gamma w(\alpha-1)^{-1}$

（証明）帰納法による．□

［補題３］

(i) $0<w<(\alpha-1)\beta/(n\gamma)$ ならば $G_n^1(\beta,\gamma,w,\alpha)<n\beta$

(ii) $(\alpha-1)\beta/n<w$ ならば $G_n^1(\beta,\gamma,w,\alpha)>n\beta$

（証明）(4),(5)式，および補題２による．□

これらより次の定理が得られる.

［定理２］

(i) wについての方程式 $G_n(\beta,\gamma,w,\alpha)=n\beta$ は唯一の根$r_n(\alpha,\gamma,\beta)$を持つ.

(ii) $r_n(\alpha,\gamma,\beta)$はαについて連続で単調増加である.

(iii) $r_n(\alpha,\gamma,\beta)$はγについて連続で単調減少である.

（証明）(i)は補題１(i)および補題３より明らか．(ii)は補題１の(i)および(ii)より明らか．(iii)は補題１の(i)および(iii)より明らか．□

これより，最適政策は，“残り期間がｎであり，A_1について事前知識が(w,α)のとき，

$$w\left\{\begin{array}{c}>\\=\\<\end{array}\right\}r_n(\alpha,\gamma,\beta)\ \text{ならば}\left\{\begin{array}{c}A_1\\A_0\text{または}A_1\\A_0\end{array}\right\}\text{を行うのが最適である．”}$$

3．問題(Q_m)の動的計画法による定式化

問題(Q_m)においては，2種類の実験e_0またはe_1のいずれか一方のみをm回行うことができる．e_0をm回行うことをE_0，e_1をm回行うことをE_1とする．e_0またはe_1のいずれかを1回行うことにより，パラメータ1及びuの指数分布から観察値を得る．uの値は未知で事前分布として，パラメータ(w, α)のガンマ分布が仮定される．この問題については，濱田[7]において動的計画法によって定式化され，$m=2$の場合には最適解が求められている．再帰方程式は

$F_{n,m}(w, \alpha)$：残りn期間あり，uについての事前分布がパラメータ(w, α)のガンマ分布のとき，以後最適政策を用いて得られる観察値の和の期待値の最大値

$F^1_{n,m}(w, \alpha)$：残りn期間あり，uについての事前分布がパラメータ(w, α)のガンマ分布のとき，まずE_1を行い，以後は最適政策を用いる場合に得られる観察値の和の期待値の最大値

とおくと，

$$F_{n,m}(w, \alpha)=\max\{mn,\ F^1_{n,m}(w, \alpha)\} \tag{6}$$

$$(n=1,2,3,\cdots;\ F_{0,m}(w, \alpha)=0)$$

で与えられる．ここに

$$F^1_{n,m}(w, \alpha)=\int_0^\infty\int_0^\infty\cdots\int_0^\infty\{\sum_{i=1}^m z_i+F_{n-1}(w+\sum_{i=1}^m z_i, \alpha+m)\}\prod_{i=1}^m \phi(z_i\ u)\,\phi(u\ w, \alpha)\,dzdu \tag{7}$$

となる．(7)は

$$F^1_{n,m}(w, \alpha)=mw(\alpha-1)^{-1}+\int_0^\infty\int_0^\infty\cdots\int_0^\infty F_{n-1}(w+\sum_{i=1}^m z_i, \alpha+m)\times\phi(z_1, z_2, \cdots, z_m \mid w, \alpha)\,dz_m\cdots dz_2dz_1 \tag{8}$$

と書き換えられる．ここに

$$\phi(z_1, z_2, \cdots, z_m \mid w, \alpha)={}_{\alpha+m}P_m w^\alpha(w-\sum_{i=1}^m z_i)^{-\alpha-m}$$

である．さらに最適解の諸性質として，$F^1_{n,m}(w, \alpha)$がwについて連続で狭義単調増加なことより，wについての方程式

$$F^1_{n,m}(w, \alpha)=mn$$

の解の存在と一意性が得られている.

［定理３］（濱田[7]）

(i) $F^{1}_{n,m}(w,\alpha)$はwについて連続で狭義単調増加，αについて連続で狭義単調減少である.

(ii) wについての方程式 $F^{1}_{n,m}(w,\alpha)=mn$ は唯一の根 $s_{n,m}(\alpha)$をもつ.

４．問題(Q_m)の最適解

問題(Q_2)の最適解を求める一つの方法は，濱田[7]において与えられているが，mの値が３以上の場合にはそのままでは拡張できない．ここでは，一般のmに対する最適解を求めるための方法を与える.

n＝1に対して，

$$F_{1,m}(w,\alpha)=\begin{cases} m, & 0<w<s_{1,m}(\alpha) \\ mw(\alpha-1)^{-1}, & s_{1,m}(\alpha)\leqq w \end{cases}$$

n≧2に対して，

$$F_{n,m}(w,\alpha)$$

$$=\begin{cases} nm, & 0<w<s_{n,m}(\alpha) \\ (n-t+1)mw(\alpha-1)^{-1}+(t-1)m & \\ \quad +mw^{\alpha}\sum_{k=0}^{(n-s+1)m-1}\sum_{i=1}^{s-1}C_{n,k,i,m}(\alpha)\{s_{1,m}(\alpha+(n-1)m)-w\}^{k}/k!, & \\ & s_{n,m}(\alpha+(n-t)m)\leqq w<s_{n-1,m}(\alpha+(n-t+1)m) \quad (2\leqq t\leqq n) \\ nmw(\alpha-1)^{-1}, & s_{1,m}(\alpha+(n-1)m)\leqq w \end{cases}$$

ここに，$C_{n,k,i,m}(\alpha)$は i＝1，$0\leqq k\leqq m-1$に対して

$$C_{n,k,1,m}(\alpha)=-{}_{\alpha+k-1}P_{k}\{s_{1,m}(\alpha+(n-1)m)\}^{-\alpha-k}$$
$$+{}_{\alpha+k-2}P_{k-1}\{s_{1,m}(\alpha+(n-1)m)\}^{-\alpha-k+1},$$

$1\leqq i\leqq n-2$，$0\leqq k\leqq m-1$に対して

$$C_{n,k,i,m}(\alpha)=\sum_{h=k}^{m-1}[-{}_{\alpha+h-1}P_{h}\{s_{i,m}(\alpha+(n-i)m)\}^{-\alpha-h}$$
$$+{}_{\alpha+h-2}P_{h-1}\{s_{i,m}(\alpha+(n-i)m)\}^{-\alpha-h+1}$$
$$-\sum_{j=0}^{(n-1-i)m-1}\sum_{t=1}^{i}C_{n-1,j,t,m}(\alpha)\{s_{1,m}(\alpha+(n-1)m)$$

$$-s_{i,m}(\alpha+(n-i)m\}^{j+m-h}{}_{\alpha+m-1}P_m/(j+m-h)!$$

$$+\sum_{j=0}^{(n-i)m-1}\sum_{t=1}^{i-1}C_{n-1,j,t,m}(\alpha+m)\{s_{1,m}(\alpha+(n-1)m)$$

$$-s_{i,m}(\alpha+(n-i)m\}^{j+m-h}{}_{\alpha+m-1}P_m/(j+m-h)!]$$

$$\times\{s_{i,m}(\alpha+(n-i)m)-s_{1,m}(\alpha+(n-1)m)\}^{h-k}/(h-k)!$$

$i=n-1$, $0\leqq k\leqq m-1$に対して

$$C_{n,k,n-1,m}(\alpha)=\sum_{h=k}^{m-1}[-{}_{\alpha+h-1}P_h\{s_{n-1,m}(\alpha+m)\}^{-\alpha-h}$$

$$+{}_{\alpha+h-2}P_{h-1}\{s_{n-1,m}(\alpha+m)\}^{-\alpha-h+1},$$

$$+\sum_{j=0}^{m-1}\sum_{t=1}^{n-2}C_{n-1,j,t,m}(\alpha+m)\{s_{1,m}(\alpha+(n-1)m)$$

$$-s_{n-1,m}(\alpha+m)\}^{j+m-h}{}_{\alpha+m-1}P_m/(j+m-h)!]$$

$$\times\{s_{n-1,m}(\alpha+m)-s_{1,m}(\alpha+(n-1)m)\}^{h-k}/(h-k)!$$

$1\leqq i\leqq n-2$, $m\leqq k\leqq(n-i)m-1$

$$C_{n,k,i,m}(\alpha)=C_{n-1,k-m,i,m}(\alpha+m)_{\alpha+m-1}P_m$$

さらに

$$s_{1,m}(\alpha)=\alpha-1$$

および，$s_{n,m}(\alpha)$ $(n=2,3,\cdots)$ は次のwについての方程式の唯一の根である．

$$-1+w/(\alpha-1)+w^{\alpha}\sum_{k=0}^{m-1}\sum_{i=1}^{n-1}C_{n,k,i,m}(\alpha)\{s_{1,m}(\alpha+(n-1)m)-w\}^{k}/k!=0$$

5．問題$(P_{m,m})$と問題(Q_m)の同値性

問題$(P_{\beta,\gamma})$において，βとγが正整数で$\beta=\gamma=m$のとき，変数変換

$$z=y_1+y_2+\cdots+y_m$$

$$z_2=\quad y_2$$

$$z_m=\qquad\qquad y_m$$

により次の補題が得られる．

［補題４］

$$\int_0^1 f(z)\,\phi(z\ w,\alpha,\gamma)dz=\int_0^\infty\int_0^\infty\cdots\int_0^\infty f(\sum_{i=1}^{m}y_i)^{\alpha+m-1}P_m w^{\alpha}(w-\sum_{k=1}^{m}y_k)^{-\alpha-m}dy_m\cdots dy_2dy_1$$

この補題を用いて次の定理を得る．

［定理３］ $n\geqq 1$, $m\geqq 2$, $w>0$, $\alpha>1$に対して，

$$G_n(m,m,w,\alpha)=F_{n,m}(w,\alpha)$$

が成立する．

（証明）(4),(5),(6),(8)，および補題４による．□

６．問題$(P_{\gamma,\gamma})$の近似解

定理３により，問題$(P_{m,m})$の解は問題(Q_m)の解と一致する．したがって自然数mに対して，$\gamma=m$ ならば

$$r_n(\alpha,m,m)=s_{n,m}(\alpha)$$

が成立する．任意の実数γに対して，$r_n(\alpha,\gamma,\gamma)$をどのように求めるかは，重要な問題である．再帰方程式(1)と(4)を直接解くのは困難である．そこで，補題１の(iii)を用いて，近似的に次のように求める方法を提案する．$[\gamma]$はγを越えない最小の整数とする．このとき，

$$r_n(\alpha,\gamma,\gamma)=(1-\gamma+[\gamma])s_{n,[\gamma]}(\alpha)+(\gamma-[\gamma])s_{n,[\gamma]+1}(\alpha)$$

により$r_n(\alpha,\gamma,\gamma)$の値を求めて，

“残り期間がｎであり，A_1について事前知識が(w,α)のとき，

$$w\left\{\begin{array}{c}>\\=\\<\end{array}\right\}r_n(\alpha,\gamma,\gamma)\ \text{ならば}\left\{\begin{array}{c}A_1\\A_0\text{または}A_1\\A_0\end{array}\right\}\text{を行うのが最適である”}$$

とすることにより，近似解を得ることができる．

文　献

[1] Bellman, R.(1956). A problem in the sequential design of experiments. Sankhyā A 16, 221-229.

[2] Berry, D. A. and Fristedt, B. (1985). Bandit Problems: Sequential Allocation of Experiments. Chapman and Hall, London and New York.

[3] 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.

[4] Gittins, J. C. (1989). Multi-Armed Bandit Allocation Indices. Jon Wiley & Sons, New York.

[5] Hamada, T. (1978). A uniform two-armed bandit problem: the parameter of one distribution is known. J. Japan Statist. Soc. 8, 29-35.

[6] Hamada, T. (1987). A two-armed bandit problem with one arm known including switching costs and terminal rewards. J. Japan Statist. Soc. 17, 21-30.

[7] Hamada, T. (1992). A discounted uniform one-armed bandit problem. Sequential Analysis, 11, 1-15.

[8] 濱田 (1992), A two-armed bandit problem with one arm known and with batch sampling. 京都大学数理解析研究所講究録798, 213-218.

[9] Kalin, D. and Theodorescu, R. (1990). A uniform two-armed bandit problem with one arm known revisited. J. Japan Statist. Soc. 20, 159-168.

[10] Woodroofe,M.(1976). On the one arm bandit problem. Sankhyā A 38, 79-91.

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