数理解析研究所講究録
第726巻 1990年 181-189

Convex Rewardをもつ MDP と Stopping Game Problem

English Tittle: Markov Decision Processes with convex reward and Stopping Game Problems

千葉大・教養 安田正実 (Masami YASUDA)

はじめに

Stopping Game Problem(*SGP*) はよく知られた Optimal Stopping Problem を2人ゼロ和ゲームにした場合で，連続時間，離散時間それぞれにたいして，最適政策の存在定理や具体的な政策の形などが議論されている．これと深く関係する問題として，Singular Stochastic Control(*SSC*) がある．Impulse Control などともよばれる．これは在庫問題や Cash management などの Smoothing problem の問題と共通の構造をもち，一つの典型的な応用である．

しかし，Singular Stochastic Control で取り扱われる問題の多くは，定数係数を仮定している．したがって，当然，定数係数であることを拡張できないものか，もしそれが可能ならば，拡張された *SSC* は，*SGP* とどう関連をもつかが問題となろう．ここでは，まず離散時間の推移システムが線形で記述されるマルコフ決定過程 Markov Decision Process(*MDP*) を議論する．この *MDP* を *SSC* の1つの代表とみなし，一方，離散時間の独立同一分布にたいする最適停止問題を *SGP* の一つとして両者の比較をする．われわれは考える *MDP* のモデルの仮定の中に利得関数が convex であることを仮定する．これが2つの問題 *SGP* と *SSC* を関連づけるカナメであり，得られる最適政策の形について本質的な点である．

結論だけをみれば，いわゆる古典的な在庫問題の two-side S-policy の議論に他ならない．それゆえその拡張である two-side (s, S)policy の場合ではどうなるかが起こってくる．単調な *MDP* ではなく，困難な点が生じるが，それは今後の課題としたい．

1．定式化と仮定

まずここで取り扱う MDP 問題における state space と action space はそれぞれ $S=(-\infty,\infty)$, $A=(-\infty,\infty)$ とする．これの feasible action space も unbounded であることに注意する．遷移法則をあらわす system dynamics は linearly independent disturbance, かつ density をもつ場合とする：

$$x_{n+1} \stackrel{\text{def}}{=} x_n + a_n - \xi_n, \qquad x_n \in S,\ a_n \in A, \tag{1.1}$$

ただし ξ_n は各 n について独立で，すべて ξ と同一分布にしたがうとする．通常の MDP では確率測度で表現するから，これとそろえるために関数 $f(x)$ にたいする 確率変数 ξ の期待値を

$$P^a f(x) \stackrel{\text{def}}{=} Ef(x+a-\xi), \qquad Pf(x) \stackrel{\text{def}}{=} P^0 f(x)$$

で表す．

つぎにこの議論のポイントである reward について，

仮定 1. 2つの関数 $r^{\pm}(x,a)$ が与えられて，利得 $r(x,a), x \in S, a \in A$ は

$$r(x,a) \stackrel{\text{def}}{=} \begin{cases} r^+(x,a), & \text{if } a>0; \\ 0, & \text{if } a=0; \\ r^-(x,a), & \text{if } a<0. \end{cases} \tag{1.2}$$

仮定 2. $r^{\pm}(x,a)$ はそれぞれ2つの関数 $C^{\pm}(x), D^{\pm}(y), y=x+a$ の和の形で表される：

$$r^{\pm}(x,a) \stackrel{\text{def}}{=} C^{\pm}(x) + D^{\pm}(x+a). \tag{1.3}$$

ただし，$r^{\pm}(x,a)$ はともに x について連続，a について convex とし，また $r^+(x,a) \to \infty$ as $a \to \infty$, $r^-(x,a) \to \infty$ as $a \to -\infty$ かつ

$$r^+(x,a) - r^-(x,a) \left\{ \begin{matrix} > \\ = \\ < \end{matrix} \right\} 0 \quad \text{if} \quad a \left\{ \begin{matrix} > \\ = \\ < \end{matrix} \right\} 0 \tag{1.4}$$

と仮定する．

そして　有限期間の割引率のある総期待利得を

$$v_n(x) \stackrel{\text{def}}{=} \inf_{\{f_k\}} E[\sum_{k=1}^{n} \beta^{k-1} r(x_k, f_k) \mid x_1 = x], \quad n = 1, 2, \cdots, x \in S \tag{1.5}$$

また，無限期間の総期待利得を

$$v(x) \stackrel{\text{def}}{=} \lim_{n \to \infty} v_n(x), \quad x \in S \tag{1.6}$$

で定める.

この定式化のもとで，optimality equation は明らかにつぎで与えられる.

有限期間の最適方程式：

$$\begin{aligned} v_0(x) &= \min_{-\infty < a < \infty} r(x, a), \\ v_n(x) &= \min_{-\infty < a < \infty} \{r(x, a) + \beta P^a v_{n-1}(x)\}, \qquad n = 1, 2, \cdots, x \in S \end{aligned} \tag{1.7}$$

無限期間の最適方程式：

$$v(x) = \min_{-\infty < a < \infty} \{r(x, a) + \beta P^a v(x)\}, \qquad x \in S. \tag{1.8}$$

ここで考える MDP は，利得関数が unbounded positive case での minimization problem に相当している．したがって利得関数の形と推移の方程式系の仮定から，Bertsekas(1976) の結果を用いて，つぎを得ることができる.

補題 1.1.　仮定 1, 2 を満たす無限期間の MDP(1.6) について，定常な最適政策が存在し，最適方程式の解は，最適値と等しい.

2．最適政策の表現

われわれはこのような特別な利得関数にたいする MDP 問題の最適政策の形を具体的に求める．有限期間の場合には，帰納的にして求めることができる．すなわち，まず $v_0(x)$ は convex in x,　$|x| \to \infty$，$v_0(x) \to \infty$ は成立している．よって

$$\begin{aligned} w_n(x, a) &\stackrel{\text{def}}{=} r(x, a) + \beta P^a v_n(x) \\ &= C(x) + D(x + a) + \beta P^a v_n(x) \end{aligned} \tag{2.1}$$

も convex in a, $w_n(x,a) \to \infty, a \to \pm\infty$ for fixed x である．したがって feasible action space が $(-\infty,\infty)$ であったから，$-\infty < y < \infty$ についての最小化を考えると，連続性と凸性により

$$\min_{-\infty<y<\infty} \{D^{\pm}(y) + \beta P v_n(y)\} \tag{2.2}$$

が存在する．それぞれの最小となる y の値を L_n, U_n とおくと，$L_n < U_n$ である．さらに

$$\begin{aligned} &\min_{-\infty<a<\infty} \{r(x,a) + \beta P^a v_{n-1}(x)\} \\ &= \min \begin{cases} \min\limits_{a\geq 0} \{r^+(x,a) + \beta P^a v_{n-1}(x)\} \\ \beta P v_{n-1}(x) \\ \min\limits_{a\leq 0} \{r^-(x,a) + \beta P^a v_{n-1}(x)\} \end{cases} \end{aligned} \tag{2.3}$$

と a について場合分けをおこなう．いま $x+y=a$ とおくと，線形性から $P^a v_n(x) = P v_n(x)$ に注意すると，まず $a \geq 0$ では

$$\begin{aligned} &\min_{a\geq 0}\{r^+(x,a) + \beta P^a v_n(x)\} \\ =&\min_{y\geq x}\{D^+(y) + \beta P v_n(y)\} + C^+(x). \end{aligned}$$

よって (2.2) より，最小となるのは $y = L_n$ としているから，

$$y^* \stackrel{\text{def}}{=} x + a^* = \begin{cases} L_n, & \text{if } L_n \geq x; \\ x, & \text{if } L_n \leq x. \end{cases}$$

つまり action a について表すと，

$$a^* \stackrel{\text{def}}{=} \begin{cases} L_n - x, & \text{if } L_n \geq x; \\ 0, & \text{if } L_n \leq x \end{cases} \tag{2.4}$$

のときである．　同様に $a \leq 0$ では

$$\begin{aligned} &\min_{a\leq 0}\{r^-(x,a) + \beta P^a v_n(x)\} \\ =&\min_{y\leq x}\{D^-(y) + \beta P v_n(y)\} + C^-(x) \end{aligned}$$

より

$$y^* \stackrel{\text{def}}{=} x + a^* = \begin{cases} U_n, & \text{if } U_n \leq x; \\ x, & \text{if } U_n \geq x \end{cases}$$

であるから

$$a^* \stackrel{\text{def}}{=} \begin{cases} U_n - x, & \text{if } U_n \leq x; \\ 0, & \text{if } U_n \geq x \end{cases} \tag{2.5}$$

を得ることができる．以上によって，最適政策の具体的な形を求めることができた．

定理 2.1. 与えられた有限計画問題 MDP の最適政策 $\{f_n^*\}$ は

$$f_n^*(x) = \begin{cases} L_n - x, & \text{if } x \leq L_n; \\ 0, & \text{if } L_n \leq x \leq U_n; \\ U_n - x, & \text{if } U_n \leq x \end{cases} \tag{2.6}$$

であり，対応する最適値は

$$v_{n+1}(x) = \begin{cases} r^+(x, L_n - x) + \beta P v_n(L_n), & \text{if } x \leq L_n; \\ \beta P v_n(x), & \text{if } L_n \leq x \leq U_n; \\ r^-(x, U_n - x) + \beta P v_n(U_n), & \text{if } U_n \leq x \end{cases} \tag{2.7}$$

となる．

Neave(1970) が既にこのような在庫問題を取り扱っていて，setup cost があると，政策が単純 (simple) にならない反例を挙げている．つまり，(s, S)-policy が最適とならないことを主張しているが，それには利得の仮定に凸性が入っていない．前にも述べたように利得の仮定により，この具体的な最適政策と最適値を求めることができるのである．

3．無限期間の計画問題

割引き率 $\beta < 1$ を導入した無限期間の問題は，利得関数が有界ではないが，§1 の補題により，最適方程式の解が最適値となるから有限期間の極限：

$$v(x) = \lim_{x \to \infty} v_n(x) \tag{3.1}$$

となっている．よって未知定数 L, U と未知関数 $v(x)$ についての方程式

$$v(x) = \begin{cases} r^+(x, L - x) + \beta P v(L), & \text{if } x \leq L; \\ \beta P v(x), & \text{if } L \leq x \leq U; \\ r^-(x, U - x) + \beta P v(U), & \text{if } U \leq x \end{cases} \tag{3.2}$$

を考えることとなる．仮定から $v_n(x)$ が各 n について convex であり，$v(x)$ も convex となる．また確率変数 ξ が密度をもつとしているから，絶対連続で，$Pv(x) = Ev(x-\xi)$ は x の関数としてすべての点で微分可能である．したがって

$$f(x) \overset{\text{def}}{=} \frac{dv(x)}{dx} \tag{3.3}$$

とおくことができる．境界での値によって

$$\begin{aligned} &f(x) = \beta P f(x), \quad L \le x \le U, \\ &f(L) = -\left.\frac{\partial r^+(L,a)}{\partial a}\right|_{a=0}, \\ &f(U) = -\left.\frac{\partial r^-(U,a)}{\partial a}\right|_{a=0} \end{aligned} \tag{3.4}$$

となる．また

$$\begin{cases} \varphi(x,a) \overset{\text{def}}{=} \dfrac{\partial r^+}{\partial x}(x,a) - \dfrac{\partial r^+}{\partial a}(x,a), \\ \psi(x,a) \overset{\text{def}}{=} \dfrac{\partial r^-}{\partial x}(x,a) - \dfrac{\partial r^-}{\partial a}(x,a) \end{cases} \tag{3.5}$$

とおくと，無限期間での問題をつぎのように書くことができる．

定理 3.1. いま $\varphi(x,a), \psi(x,a)$ を (3.5) のように与えると，最適値 $v(x)$ の微分 (3.3) は，$f(x)$, L, U を求める two phase 型 Stephan 問題:

$$\begin{aligned} &f(x) = \varphi(x, L-x), && x \le L, \\ &f(x) = \beta P f(x), && L \le x \le U, \\ &f(x) = \psi(x, U-x), && U \le x \end{aligned} \tag{3.6}$$

に帰着される．

このような問題はよく知られた変分不等式の形で書くことができる．すなわち，ある関数 $\chi(x,a)$ が与えられたとして，すべての x, y にたいし

$$\begin{aligned} &v(x) - \beta P v(x) \le 0, \\ &v(x) - v(y) - \chi(x, y-x) \le 0 \\ &\{v(x) - \beta P v(x)\}\{v(x) - v(y) - \chi(x, y-x)\} = 0 \end{aligned} \tag{3.7}$$

をみたす $v(x)$ を求めることとなる.

とくに利得関数 $\chi(x,a)$ が x について定数（たとえば，$\chi(x,a)=K$ if $a>0,=H$ if $a<0$ で K,H が x に依らない）であり，考える系を diffusion process(Brown motion) にした場合を Harrison(1983) などが解いている．ここでの MDP 問題の場合では，

$$\chi(x,a) \stackrel{\text{def}}{=} \begin{cases} \varphi(x,a), & \text{if } a>0; \\ 0, & \text{if } a=0; \\ \psi(x,a), & \text{if } a<0. \end{cases} \tag{3.8}$$

が対応している.

4 . Stopping Game Problem(SGP) との関連

基本的な SGP での最適方程式は $\{r_1(x),\ r_2(x); -\infty < x < \infty\}$ を与えた two phase 型 Stephan 問題で，$g(x), l, u$ を求める：

$$\begin{aligned} g(x) &= r_1(x), && x \le l, \\ g(x) &= \beta P g(x), && l \le x \le u, \\ g(x) &= r_2(x), && u \le x \end{aligned} \tag{4.1}$$

の形である．この $r_1(x)$, $r_2(x)$ はそれぞれのプレーヤが早く stop したときに得る payoff を表している．上の方程式 (4.1) には，前の (3.6) と比較してみると，関数の値のなかには未知定数が入らないことに注意する．SGP の目的関数は，2 つの $r_1(x)$, $r_2(x)$ が与えられたとき，各プレーヤの stopping time τ, σ についての利得最大化／最小化で

$$\begin{aligned} g(x) &\stackrel{\text{def}}{=} \sup_{\tau\ge 0}\inf_{\sigma\ge 0} E[\beta^\tau r_1(x_\tau) I_{\{\tau\le\sigma\}} + \beta^\sigma r_2(x_\sigma) I_{\{\tau>\sigma\}} | x_0 = x] \\ &= \inf_{\sigma\ge 0}\sup_{\tau\ge 0} E[\beta^\tau r_1(x_\tau) I_{\{\tau\le\sigma\}} + \beta^\sigma r_2(x_\sigma) I_{\{\tau>\sigma\}} | x_0 = x] \end{aligned} \tag{4.2}$$

とした上の等式 (4.2) が成り立ち，これが均衡解を与えるゲーム問題とみなしている.

ここで，action は停止時刻だけでシステムにはかかわらない．したがって $\{x_n; n = 0,1,\cdots\}$ は

$$x_{n+1} \stackrel{\text{def}}{=} x_n - \xi_n \tag{4.3}$$

で $\xi_n, n = 0, 1, \cdots, \xi$ は独立同一分布にしたがうとする．さらに $r_1(x) < r_2(x)$ が重要な仮定として必要である．

このような *SGP* の最適方程式 (4.1) と §3 での *MDP* にたいする最適方程式 (3.2) を微分した式 (3.6) を比較すると，*MDP* の場合では未知定数が関数の中に入ってくる．特別な場合，すなわち，定数係数では action の デルタ関数として，2つの最適方程式は簡単に一致させることができる．簡潔にいえば，2つの問題 *SGP* と *SSC*(convex reward をもつ *MDP*) の関連として *MDP* の最適値の微分は，*SGP* の最適値であるということができよう．これは一般的には成り立たないが，2つを一致させる条件として，たとえば式 (3.5) で定義した $\varphi(x,a), \psi(x,a)$ が a によらなければ，$L = l, U = u$ で2つの関数は一致することが分る．これをまとめるとつぎの定理を得る．

定理 4.1. 関数 $\varphi(x,a), \psi(x,a)$ が a によらないとき，$r_1(x) = \varphi(x,a), r_2(x) = \psi(x,a)$ とおいて, convex reward をもつ *MDP* 最適利得関数の微分 $f(x)$ は， non-constant な係数利得の *SGP* の最適値 $g(x)$ に帰着させることができ，$f(x) = g(x)$ が成り立つ．

参考文献

[1] Beckmann, M. J.; *Product Smoothing and Inventory Control*, Oper. Res. **9** (1961) 456-567.

[2] Benes, V. E., Shepp, L. A. and Witsenhausen, H. S., *Some solvable stochastic control problems*, Stochastics, **4** (1980) 39-83.

[3] Bensoussan, A. and Lions, J. L., *Nouvelles Methodes en Control Impulsionnel*, Appl. Math. Optim. **1** (1975) 289-312.

[4] Bertsekas, D. P.; *Stochastic Optimization Problems with Non-differentiable Cost functionals*, J. Opt. Th. Appl. **12** (1973) 218-231.

[5] Bertsekas, D. P.; *Dynamic Programming and Stochastic Control*, Academic Press, New York 1976.

[6] Harrison, J. M., *Brownian motion and stochastic flow systems*, John Wiley, New York, 1985.

[7] Harrison, J. M., Selleke, T. M. and Taylor, A. J.; *Impulse Conrtol of Brownian Motion*, Math. Oper. Res. **8** (1983) 454-466.

[8] Heyman, D. P. and Sobel, M. *Stochastic Models in Operations research,* II: *Stochastic Optimization*, McGraw-Hill, 1982.

[9] Karatzas, I. and Shreve, S. E., *Connection between optimal stopping and stochastic control* , I: *Monotone follower problems* , SIAM J. Control Optim. **22** (1984) 856-877.

[10] Karatzas, I. and Shreve, S.E., *Connection between optimal stopping and stochastic control*, II: *Reflected follower problems* , SIAM J. Control Optim. **23** (1985) 433-451.

[11] Neave, E. H.;*The stochastic cash balance problem with fixed costs for increases and decreases*, Manag. Sci. **16**(1970) 472-490.

[12] Schäl, M. *On the optimality of (s,S)-policies in dynamic inventory models with finite horizon*, SIAM J. Appl. Math., **30**(1976) 528-537.

[13] Serfozo, R. *Monotone optimal policies for Markov decision processes*, Math. Prog. Study, **6** (1976) 202-215.

[14] Topkis, D.M. *Minimizing a subadditive function on a lattice*, Oper. Res. , **26** (1978) 305-321.