A markov decision process with convex reward and its associated stopping game

Masami Yasuda
Department of Mathematics & Informatics
Chiba University
Inage-ku, Chiba 263
Japan

ABSTRACT

This note concerns with a special case of Markov decision process, whose reward function is convex in the action variable and the system dynamics is linear. We show that the derivative of the optimal value in the Markov decision process reduces a game value for the two person zero-sum stopping game. The result is the discrete version of so-called the singular stochastic control. By this discussion, a slight extension of the previous results is obtained. It will be applied to the smoothing problem such as the inventory or the cash management.

1. INTRODUCTION

Stopping game problem is motivated from the two-person zero sum game version of the optimal stopping problem. There are many papers which discuss existence of the optimal strategy and the solution of optimality equation. See Dynkin (1969), Neveu (1975). On the other hand the singular stochastic control by Harrison, Sellke and Taylar (1983) is known as the descendant from a sequential decision or a stochastic control problem. In the theory of Brownian motion, the connection between stopping game problem and the singular stochastic control was pointed out by Karatzas and Shereve (1984, 1985). These two problems have the same feature of the model structure. That means the monotonicity of optimal policies as in sequential decision processes (Serfozo (1976), Heyman and Sobel (1982)). In this paper Markov decision processes with special structure are formulated as the optimization problem and we clarify its connection with the stopping game problem. By this argument a slight extension for results is obtained and the fundamental condition will be seen that the linearity of dynamic system and the convexity of

Journal of Information & Optimization Sciences
Vol. 18 (1997), No. 3, pp. 01-09
© Analytic Publishing Co. 0252-2667/97 $2.00+.25

reward. It will be applied to the smoothing problem such as the inventory or the cash management of Schäl (1976) and Beckman (1961).

2. FORMULATION OF MARKOV DECISION PROCESSES

In this section we will formulate Markov decision processes, whose optimal value and optimal policy are explicitly derived.

2.1. *Formulation*

The Markov decision processes consist of four components (S, A, r, P^a): Let a state space S and an action A be $(-\infty, \infty)$ in the real Euclid space respectively. A reward function is denoted by $r = r(x, a), x \in S, a \in A$ such that

$$r(x, a) := \begin{cases} r_+(x, a), & \text{if } a > 0 \\ 0, & \text{if } a = 0 \\ r_-(x, a), & \text{if } a < 0 \end{cases} \tag{2.1}$$

where $r_\pm(x, a)$ are continuous in x, convex in a, and it tends to ∞ as $a \to \pm\infty$ respectively. For the dominance of two functions $r_\pm$,

$$r_+(x, a) - r_-(x, a) \begin{cases} > 0, & \text{if } a > 0 \\ = 0, & \text{if } a = 0 \\ < 0, & \text{if } a < 0. \end{cases} \tag{2.2}$$

Furthermore the reward function is assumed that each of $r_\pm$ decomposed as the sum of $r_1^\pm$ and $r_2^\pm$ respectively

$$r_\pm(x, a) := r_1^\pm(x) + r_2^\pm(x + a). \tag{2.3}$$

For a transition probability $P^a = \{P^a(x, y); x, y \in S\}, a \in A$, assume that there exists a random variable ξ with its density p_ξ such that

$$\begin{aligned} P^a f(x) &:= \int_S f(y) P^a(x, dy) := E[f(x + a + \xi)] \\ &:= \int_{-\infty}^{\infty} f(x + a + t) p_\xi(t) dt \end{aligned} \tag{2.4}$$

for any integrable function f on S. The most essential is that the system dynamics is a linear case and its random disturbance ξ has a density with respect to Lebesgue measure on S. So $P^a f(x) = P^0 f(x+a)$ holds and it is smooth in x. Especially $P^0 f(x) = E[f(x+\xi)]$ will be used in the following Stephan problem.

The finite horizon Markov decision processes are defined by the policy, the mapping from the state space to the action space, which is denoted by $\{f_k\}_{k=1,2,\dots}$, $x \in S, f_k(x) \in A$ and the discount factor $0 < \beta < 1$. The aim of Markov decision process is to minimize the total discounted expected cost $v_n(x)$:

$$v_n(x) := \inf_{\{f_k\}} E\left[\sum_{k=1}^{n} \beta^{k-1} r(x_k, f_k(x_k)) \,|\, x_1 = x\right] \tag{2.5}$$

with the initial state x, where $\{x_k\}$ is the induced Markov chain with the non-stationary transition probability $\{P^{f_k}\}$.

In case of the infinite horizon, the total expected reward $v(x)$:

$$v(x) := \inf_{\{f_k\}} E\left[\sum_{k=1}^{\infty} \beta^{k-1} r(x_k, f_k(x_k)) \,|\, x_1 = x\right] \tag{2.6}$$

is minimized with respect to policies $\{f_k\}$.

Under these formulation, the optimality equation for the finite horizon case is given as follows:

$$\begin{aligned} v_0(x) &= \min_{-\infty<a<\infty} r(x, a), \\ v_i(x) &= \min_{-\infty<a<\infty} \{r(x, a) + \beta P^a v_{i-1}(x)\}, \quad 1 \le i \le n. \end{aligned} \tag{2.7}$$

Similarly it is obtained the infinite horizon case as

$$v(x) = \min_{-\infty<a<\infty} \{r(x, a) + \beta P^a v(x)\}. \qquad (2.8)$$

These cases of the Markov decision processes are bounded from below and so it is a positive case. Then we can prove the next lemma by Bertsekas (1973, 1976).

LEMMA 2.1 *If the infinite Markov decision process* (S, A, r, P^a) *satisfies the previous assumption, there exists a stationary policy and the solution of the optimality equation equals the optimal value.*

2.2. *Explicit form of the optimal policy*

The optimal policy for the Markov decision process (S, A, r, P) is described explicitly. Firstly we consider the finite case problem and then the infinite case by induction.

THEOREM 2.1. *For the finite Markov decision process (S, A, r, P^a), there exist two thresholds L_i and U_i for each $i = 1, 2, ..., n$ and the optimal policy f_i^* is given by*

$$f_i^*(x) = \begin{cases} L_i - x . & \text{if } x \le L_i \,; \\ 0, & \text{if } L_i \le x \le U_i \,; \\ U_i - x, & \text{if } U_i \le x \,; \end{cases} \tag{2.9}$$

its associated optimal value v_i equals

$$v_{i+1}(x) = \begin{cases} r_+(x, L_i - x) + \beta P^0 v_i(L_i), & \text{if } x \le L_i \,; \\ \beta P^0 v_i(x), & \text{if } L_i \le x \le U_i \,; \\ r_-(x, U_i - x) + \beta P^0 v_i(U_i), & \text{if } U_i \le x \,. \end{cases} \tag{2.10}$$

PROOF. First we note that $v_0(x)$ is convex in x and $v_0(x) \to \infty$ as $|x| \to \infty$. So

$$w_i(x, a) := r(x, a) + \beta P^a v_i(x) = r_1(x) + r_2(x + a) + \beta P^0 v_i(x + a) \tag{2.11}$$

is also convex in a and $w_i(x, a) \to \infty$ as $a \to \infty$ for fixed x from (2.1)–(2.3). Since the feasible action space is $a \in A = (-\infty, \infty)$, these exists its minimum with respect to $y = x + a \in (-\infty, \infty)$ for all x respectively.

$$\begin{aligned} L_i &:= \operatorname*{argmin}_{-\infty<y<\infty} \{r_2^-(y) + \beta P^0 v_i(y)\} \\ U_i &:= \operatorname*{argmin}_{-\infty<y<\infty} \{r_2^+(y) + \beta P^0 v_i(y)\} \end{aligned} \tag{2.12}$$

and satisfy, because (2.2),

$$L_i < U_i \tag{2.13}$$

for each i.

To consider the minimization of $w_i(x, a)$ with respect to a for each x, we divide the feasible action space into three: $a \ge 0, a = 0, a \le 0$.

That is,

$$\min_{-\infty<a<\infty} \{w_i(x, a)\} = \min \begin{Bmatrix} \min\limits_{a\geq 0}\{r_+(x, a) + \beta P^a v_{i-1}(x)\} \\ \\ \beta P^0 v_{i-1}(x) \\ \\ \min\limits_{a\leq 0}\{r_-(x, a) + \beta P^a v_{i-1}(x)\} \end{Bmatrix}$$

from (2.1). In case of $a \geq 0$, by (2.3) and (2.4)

$$\begin{aligned} &\min_{a\geq 0} \{r_+(x, a) + \beta P^a v_{i-1}(x)\} \\ &= \min_{a\geq 0} \{r_1^+(x) + r_2^+(x+a) + \beta P^0 v_{i-1}(x+a)\} \\ &= \min_{y\geq x} \{r_2^+(y) + \beta P^0 v_{i-1}(y)\} + r_1^+(x) \end{aligned}$$

with a fixed x. By (2.12), the minimization is achieved in y^* as follows:

$$y^* := \begin{cases} L_i, & \text{if } L_i \geq x\,; \\ x, & \text{if } L_i \leq x\,. \end{cases}$$

That is,

$$a^* := \begin{cases} L_i - x, & \text{if } L_i \geq x\,; \\ 0, & \text{if } L_i \leq x\,. \end{cases} \tag{2.14}$$

Similarly, the case of $a \leq 0$ could be shown as

$$a^* := y^* - x = \begin{cases} U_i - x, & \text{if } U_i \geq x; \\ 0, & \text{if } U_i \leq x. \end{cases} \tag{2.15}$$

Combined the case of $a = 0$ with (2.14) and (2.15), the optimal policy f_i^* is explicitly determined. Its associated value for the optimal policy is immediate. □

As a remark, this optimal policy is same as that of the inventory problem. Schäl discussed it with a setup cost but no cost in this situation. If we impose the cost the optimal policy reduces "both side (s, S)-policy". We must find also the threshold value, L, U as free boundaries in the stochastic control problem (Bensoussan and Lions (1982)).

3. INFINITE HORIZON PROBLEM

In the previous section we show the explicit form of the optimal policy. Imposing a discount factor for the reward (cost) function in the infinite horizon problem are considered here. In order to avoid the trivial case, we assume that the finiteness of the reward for each policy and let define

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

Then the problem is to find L, U and $v(x)$ such that

$$v(x) = \begin{cases} r_+(x, L-x) + \beta P^0 v(L), & \text{if } x \le L, \\ \beta P^0 v(x), & \text{if } L \le x \le U, \\ r_-(x, U-x) + \beta P^0 v(U), & \text{if } U \le x, \end{cases} \tag{3.2}$$

because of the finite horizon in (2.10). Since $v_n(x)$ is convex for each n, $v(x)$ is $v(x)$. Also we have assumed that the random variable ξ has a density and is absolutely continuous, so $P^0 v(x) = E[v(x+\xi)]$ is differentiable in x for each point. Therefore

$$f(x) := \frac{dv(x)}{dx} \tag{3.3}$$

is well-defined. Then differentiating (3.2), it leads to the next relation.

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

Further, by letting

$$\begin{cases} \phi(x,a) := \dfrac{\partial r_+}{\partial x}(x,a) - \dfrac{\partial r_+}{\partial a}(x,a), \\ \psi(x,a) := \dfrac{\partial r_-}{\partial x}(x,a) - \dfrac{\partial r_-}{\partial a}(x,a) \end{cases} \tag{3.5}$$

the above (3.3) and (3.4) imply the Stephan problem as stated in the next theorem.

THEOREM 3.1. *The infinite horizon Markov decision problem* (S, A, r, P^a) *could be reduced to the following problem, so called as two phase Stephan problem: For given data* ϕ, ψ *and* P^0 *by* (3.5), (2.4), *find* $f(x)$, L, U *such that*

$$f(x) = \begin{cases} \phi(x, L - x), & x \le L, \\ \beta P^0 f(x), & L \le x \le U, \\ \psi(x, U - x), & U \le x. \end{cases} \tag{3.6}$$

PROOF. The proof is immediate from the discussion of (3.4) and (3.5) □

It is seen that the Stephan problem (3.6) is equivalent to the variational equation

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

associated with

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

Harrison, Sellke and Taylar (1983) had solved the variational inequality with a constant number $\chi(x, a) = K$ if $a > 0, = 0$ if $a = 0, = H$ if $a < 0$ in the Brownian motion system. Their coefficient of the system is constant so $I - \beta P^0$ corresponds to the differential operator with a constant coefficient.

4. STOPPING GAME AND ITS DERIVATIVE

The zero-sum two-person stopping game for Markov chain $\{X_t\}_{t=0,1,2,\ldots}$ is to find a stopping time τ, σ so as to minimize/maximize payoff for each player. The payoff of each player are defined by two kind function of $\rho_1(x)$ and $\rho_2(x)$ as follows:

$$\underline{g}(x) := \sup_{\tau\geq 0} \inf_{\sigma\geq 0} E[R(\tau, \sigma) \,|\, X_0 = x],$$
$$\bar{g}(x) := \inf_{\sigma\geq 0}\sup_{\tau\geq 0} E[R(\tau, \sigma) \,|\, X_0 = x] \tag{4.1}$$

where $R(\tau, \sigma) = \beta^{\tau}\rho_1(X_\tau)I_{\{\tau\leq\sigma\}} + \beta^{\sigma}\rho_2(X_\sigma)I_{\{\tau>\sigma\}}$, respectively.

The discussion, as in the usual game theory, leads to seek an equilibrium point:

$$g(x) := \bar{g}(x) = \underline{g}(x). \tag{4.2}$$

If we impose an appropriate condition on payoff function such as

$$\rho_1(x) > \rho_2(x) \tag{4.3}$$

with $\rho_1(x) \to -\infty\ (x \to -\infty)$ and $\rho_2(x) \to \infty\ (x \to \infty)$. The equilibrium point exists in the class of the ordinary stopping times (Yasuda (1985)), that is, a class of pure strategies and the game value $g(x)$ by (4.2) satisfies the optimality equation:

$$g(x) = \begin{cases} \max\{\rho_1(x), \beta Pg(x)\} \\ \beta Pg(x) \\ \min\{\rho_2(x), \beta Pg(x)\} \end{cases} \tag{4.4}$$

where P is a transition probability for Markov chain $\{X_t\}$. If we write a free boundary l and u such that $\{\rho_1(x) \geq \beta Pg(x)\} = \{x \leq l\}$ and $\{\rho_2(x) \leq \beta Pg(x)\} = \{x \geq u\}$. Then (4.4) could be rewritten as

$$g(x) = \begin{cases} \rho_1(x) & \text{if } x \leq l, \\ \beta Pg(x) & \text{if } l \leq x \leq u, \\ \rho_2(x) & \text{if } u \leq x. \end{cases} \tag{4.5}$$

This Stephan problem is the same form of that of (3.6) in the Markov decision process with convex reward.

Karatzas and Shereve (1984, 1985) have derived this connection between optimal stopping and stochastic control in the diffusion processes.

To make in concord with the previous result, let us assume that the reward $\phi(x, a)$, $\psi(x, a)$ are simply dependent on x. So let $\rho_1(x) = \psi(x, a)$, $\rho_2(x) = \psi(x, a)$ and $P^0 f(x) = Pf(x)$ for every $f(x)$, then we could obtain the next theorem. Thus we have a slight extension of the previous result.

THEOREM 4.1. *The derivative (3.3) of the optimal value in the Markov decision processes with convex reward is consistent with the game value (4.1) of the zero-sum stopping game.*

REFERENCES

1. M. J. Beckmann (1961), Product Smoothing and Inventory Control, *Oper. Res.*, Vol. 9, p. 456-567.
2. A. Bensoussan and J. L. Lions (1982), *Applications of Variational Inequalities in Stochastic Control*, North-Holland, Amsterdam.
3. D. P. Bertsekas (1973), Stochastic Optimization Problems with Non-differentiable Cost Functionals, *J. Opti. Theor. Appl.*, Vol. 12, pp. 218-231.
4. D. P. Bertsekas (1976), *Dynamic Programming and Stochastic Control*, John Wiley & Sons, New York.
5. E. B. Dynkin (1969), Game Variant of a Problem on Optimal Stopping, *Soviet Math. Dokl.*, Vol. 10, pp. 270-274.
6. J. M. Harrison, T. M. Sellke and A. J. Taylar (1983), Impulse Control of Brownian Motion, *Math. Oper. Res.*, Vol. 8, pp. 454-466.
7. D. P. Heyman and M. Sobel (1982), *Stochastic Models in Operations research, II: Stochastic Optimization*, McGraw-Hill.
8. I. Karatzas and S. E. Shereve (1984, 1985), Connection between Optimal Stopping and Stochastic Control, I: Monotone Follower Problems, II: Reflected Follower Problems, *SIAM J. Control Optim.*, Vols. 22, 23, pp. 856-877, 433-451.
9. J. Neveu (1975), *Discrete-Parameter Martingales*, North-Holland, Amsterdam.
10. M. Schäl (1976), On the Optimality of (s, S)-policies in Dynamic Inventory Models with finite Horizon, *SIAM J. Appl. Math.*, Vol. 30, pp. 528-537.
11. R. Serfozo (1976), Monotone Optimal Policies for Markov Decision Processes, *Math. Prog. Study*, Vol. 6, pp. 202-215.
12. M. Yasuda (1985), On a Randomized Strategy in Neveu's Stopping Problem, *Stoch. Proc. Appli.*, Vol. 21, pp. 159-166.

Received March, 1997