数理解析研究所講究録
第 726 巻 1990 年 44-56

Characterization of optimal strategies using their expected payoff

長岡高専　涌田和芳 (Kazuyoshi Wakuta)

§1. Introduction

Stochastic game における optimal strategies の特徴付けについては、Groenewegen [3] が、かなり一般的なモデルで議論している．そこでは、saddle conserving , saddling など様々な概念が saddle function を用いて導入され、optimal strategies が特徴付けられている．ここでは、２人零和 game について、saddle function のかわりに player I への expected payoff を用いて optimal strategies を特徴付けることを考える．このような特徴付けでは、Borel 可測性の枠の中で問題を議論することが可能となる．特別な結果として、stationary strategies が optimal であるための必要十分条件が得られ、更に、Shapley [5] により最初に証明された、stationary strategies が optimal であるための良く知られた十分条件が得られる．また、我々の特徴付けは Dynamic programming に

おける Blackwell [1]の定理「policyは、その rewardが optimality equationを満たすとき、またそのときに限り optimalである」の game 的拡張にもなっている.

§2. The semi-Markov game

2人零和 semi-Markov game は次のもので定義される.

S, A, B： nonempty Borel sets, Sは state space, Aは player Iの action space, Bは player IIの action space

$A(\cdot), B(\cdot)$： Sから A (or B) への multifunction で、各 $s \in S$ に対して nonempty permissible set of actions $A(s)$ (or $B(s)$) を与える. $K = \{(s, a, b) \mid s \in S, a \in A(s), b \in B(s)\}$ は Borel set と仮定する. また $f: S \to A$ s.t. $f(s) \in A(s)$, $s \in S$ なる Borel 可測写像の全体を $\mathcal{F}$, $g: S \to B$ s.t. $g(s) \in B(s)$, $s \in S$ なる Borel 可測写像の全体を $\mathcal{G}$ とし、$\mathcal{F} \neq \phi$, $\mathcal{G} \neq \phi$ と仮定する.

$p \in Q(S \mid S \times A \times B)$： the law of motion of state. ただし、$Q(Y \mid X)$はXからYへの transition probabilities の全体を表わす.

$q \in Q(R_+ \mid S \times A \times B \times S)$： the distribution of the sojourn time. ただし、$R_+ = [0, \infty)$.

r : $S \times A \times B$ 上の Borel 可測関数で、player I への payoff function.

$\alpha > 0$: discount factor.

Π は player I の strategies の全体, Γ は player II の strategies の全体を表わすとする. player I に対する expected total discounted payoff を

$$v(\pi, \gamma)(s_1) = E_{(\pi, \gamma)}\left[\sum_{n=1}^{\infty} e^{-\alpha T_{n-1}} r(s_n, a_n, b_n) \middle| s_1 \right], \quad \pi \in \Pi, \gamma \in \Gamma, s_1 \in S$$

で定義する. ここで、s_n, a_n, b_n はそれぞれ、n-th state, n-th action for player I, n-th action for player II, T_n (ただし $T_0 = 0$) は $(n+1)$-st decision epoch を表わす.

$$v(\pi', \gamma)(s_1) \leqq v(\pi, \gamma)(s_1) \leqq v(\pi, \gamma')(s_1), \quad \pi' \in \Pi, \gamma' \in \Gamma, s_1 \in S$$

を満たす strategies (π, γ) があれば、π, γ はそれぞれ player I, II に対して optimal であるという. 次の条件が成り立つものと仮定する.

Condition (1) (cf. Nunen and Wessels [8]). 次の条件を満たす S 上の Borel 可測関数 $w \geqq 1$, ρ $(0<\rho<1)$, $M>0$ が存在する.

$$|r(s,a,b)| \leqq M\, w(s),$$

$$\int_S \beta(s,a,b,s')\, w(s')\, dp(s' \mid s,a,b) \leqq \rho\, w(s), \quad s \in S,\ a \in A,\ b \in B.$$

ここで, $\beta(s,a,b,s') = \int_0^\infty e^{-\alpha t} dq(t \mid s,a,b,s')$.

$H_1 = S$, $H_{n+1} = K \times R_+ \times H_n$ $(n \geqq 1)$ とする. H_nは n-th decision epoch までのシステムの可能な歴史の全体を表わす. H_n上の Borel 可測関数 v に対して

$$\|v\|_n = \sup_{h_n \in H_n} |v(h_n)|\, w(s_n)^{-1}$$

とおき、$\|v\|_n < \infty$ なるすべての v からなる Banach space を H_n^w で表わす. 特に、$n=1$ のとき、H_1^w を S^w, $\|\cdot\|_1$ を $\|\cdot\|$ とかく. Condition (1) により次の結果が得られる (cf. Wakuta [7]).

Proposition 2.1. 任意の strategies (π, γ) に対して、$\angle = \sup_{n \geqq 0} T_n$ は a.s. で infinite である.

Proposition 2.2. 任意の strategies (π,γ) に対して、

$$\| v(\pi,\gamma) \| \leqq M/(1-\rho).$$

§3. The characterization of optimal strategies

$v(\pi,\gamma)(s_1)$, $s_1 \in S$, は次のようにかくことができる.

$$v(\pi,\gamma)(s_1) = E_{(\pi,\gamma)}\left[\sum_{n=1}^{\infty} e^{-\alpha T_{n-1}} r(\pi_n,\gamma_n)(h_n) \middle| s_1\right], \quad s_1 \in S.$$

ここで,

$$r(\pi_n,\gamma_n)(h_n) = \iint_{A\times B} r(s_n,a_n,b_n)\, d\pi_n(a_n|h_n)\, d\gamma_n(b_n|h_n), \quad h_n \in H_n.$$

$v \in S^w$ に対して、

$$P(\pi_n,\gamma_n)v(h_n) = \iiint_{S\times A\times B} \beta(s_n,a_n,b_n,s_{n+1})\, v(s_{n+1})\, dp(s_{n+1}|s_n,a_n,b_n) \times d\pi_n(a_n|h_n)\, d\gamma_n(b_n|h_n),$$

$$L(\pi_n,\gamma_n)v(h_n) = r(\pi_n,\gamma_n)(h_n) + P(\pi_n,\gamma_n)v(h_n)$$

とおく. L は S^w から H_n^w への operator である. また、

$$v(\pi,\gamma)(h_n) = E_{(\pi,\gamma)}\left[\sum_{k=n}^{\infty} e^{-\alpha(T_{k-1}-T_{n-1})} r(\pi_k,\gamma_k)(h_k) \,\Big|\, h_n\right],$$

$$v(\pi,\gamma)(s_n) = R(\pi,\gamma)(s_n)$$

とおく．ただし、

$$R(\pi,\gamma)(s) = E_{(\pi,\gamma)}\left[\sum_{n=1}^{\infty} e^{-\alpha T_{n-1}} r(\pi_n,\gamma_n)(h_n) \,\Big|\, s_1 = s\right].$$

最初に、strategies (π,γ) が optimal であるための必要条件を考える．次の定理の証明方法は、Groenewegen and Wessels [2] が Markov strategies に対して用いたものを一般の strategies に拡張したものである．

Theorem 3.1. strategies (π,γ) が optimal ならば、任意の $(f,g)\in F\times G$ と $s_1\in S$ に対して

$$L(f,\gamma_n)v(\pi,\gamma)(h_n) \leqq v(\pi,\gamma)(s_n) \leqq L(\pi_n,g)v(\pi,\gamma)(h_n),$$
$$P_{(\pi,\gamma),s_1}\text{-a.s. } h_n,\ n\geqq 1,$$

が成り立つ．

Proof. (1st step) $\pi(n)=\{\pi_1,\cdots,\pi_n,\pi_1,\pi_2,\cdots\}$, $\gamma(n)=\{\gamma_1,\cdots,\gamma_n,\gamma_1,\gamma_2,\cdots\}$ も optimal strategies であることを示す．任意の $\gamma'\in\Gamma$ に対して、

$$v(\pi(n),\gamma')(s_1)=E_{(\pi(n),\gamma')}\left[\sum_{k=1}^{n}e^{-\alpha T_{k-1}}r(\pi_k,\gamma_k')(h_k)+e^{-\alpha T_n}v(\pi(n),\gamma')(h_{n+1})\,\middle|\,s_1\right].$$

ここで、$v(\pi(n),\gamma')(h_{n+1})=v(\pi,\gamma^*)(s_{n+1})$

$$\begin{aligned}&\geqq v(\pi,\gamma)(s_{n+1})\\&\geqq v(\pi^*,\gamma)(s_{n+1})\\&=v(\pi,\{\gamma_1',\cdots,\gamma_n',\gamma\})(h_{n+1}),\ p_{(\pi(n),\gamma'),s_1}\text{-a.s. } h_{n+1}.\end{aligned}$$

ただし、$\pi_m^*(\cdot\mid h_m')=\pi_{n+m}(\cdot\mid h_{n+1},h_m')$, $\gamma_m^*(\cdot\mid h_m')=\gamma_{n+m}(\cdot\mid h_{n+1},h_m')$.

$(\pi(n),\gamma')$ と $(\pi,\{\gamma_1',\cdots,\gamma_n',\gamma\})$ は、n-th step までは同じ strategies なので、

$$v(\pi(n),\gamma')(s_1)\geqq v(\pi,\{\gamma_1',\cdots,\gamma_n',\gamma\})(s_1)\geqq v(\pi,\gamma)(s_1),\ s_1\in S.$$

同様に、

$$v(\pi',\gamma(n))(s_1)\leqq v(\pi,\gamma)(s_1),\ s_1\in S.$$

故に、$v(\pi(n),\gamma(n))(s_1)=v(\pi,\gamma)(s_1)$, $s_1\in S$ が成り立ち、$(\pi(n),\gamma(n))$ も optimal である.

(2nd step) $\pi^0=\{\pi_1,\cdots,\pi_{n-1},\pi'\}$, $\gamma^0=\{\gamma_1,\cdots,\gamma_{n-1},\gamma'\}$, $\pi'\in\Pi$, $\gamma'\in\Gamma$ に対して、

$$v(\pi^0,\gamma)(h_n)\leqq v(\pi,\gamma)(h_n)\leqq v(\pi,\gamma^0)(h_n),$$

$$v(\pi,\gamma)(h_n)=v(\pi,\gamma)(s_n),\quad p_{(\pi,\gamma),s_1}\text{-a.s. } h_n$$

が成り立つことを示す．最初の不等式は明らか．特に、

$\pi^0 = \pi(n-1)$, $\gamma^0 = \gamma(n-1)$ とおけば，

$$v(\pi,\gamma)(h_n) \geqq v(\pi(n-1),\gamma)(h_n) = v(\pi,\gamma')(s_n) \geqq v(\pi,\gamma)(s_n)$$

ただし、$\gamma'_m(\cdot \mid h'_m) = \gamma_{n+m-1}(\cdot \mid h_n, h'_m)$．

同様に、

$$v(\pi,\gamma)(h_n) \leqq v(\pi,\gamma(n-1))(h_n) = v(\pi',\gamma)(s_n) \leqq v(\pi,\gamma)(s_n).$$

故に等式が得られる．

以上のことから、次のことが成り立つ．

$$\begin{aligned} v(\pi,\gamma)(s_n) &= v(\pi(n),\gamma(n))(s_n) \\ &= v(\pi(n),\gamma(n))(h_n) \\ &\leqq v(\pi(n),\{\gamma_1,\cdots,\gamma_{n-1},g,\gamma\})(h_n) \\ &= L(\pi_n,g)\,v(\pi,\gamma)(h_n),\quad p_{(\pi,\gamma),s_1}\text{-a.s. } h_n. \end{aligned}$$

同様にして、

$$v(\pi,\gamma)(s_n) \geqq L(f,\gamma_n)\,v(\pi,\gamma)(h_n),\quad p_{(\pi,\gamma),s_1}\text{-a.s. } h_n.$$

故に定理が証明された．

次に、strategies (π,γ) が optimal であるための十分条件を考える．

Lemma 3.1. (cf. Wakuta [7, Proposition 3.1])．任意の $\pi,\pi' \in \Pi$, $\gamma,\gamma' \in \Gamma$ に対して、

$$\lim_{n\to\infty} E_{(\pi',\gamma')}\left[e^{-\alpha T_{n-1}} v(\pi,\gamma)(s_n) \,\middle|\, s_1 \right] = 0$$

が成り立つ.

次の定理の証明方法は、Wakuta [6] [7] が (semi-) Markov decision process に対して用いたものを game 的に拡張したものである.

Theorem 3.2. 任意の $(f, g) \in F \times G$ と $s_1 \in S$ に対して、

$$L(f,\gamma_n)v(\pi,\gamma)(h_n) \leqq v(\pi,\gamma)(s_n) \leqq L(\pi_n,g)v(\pi,\gamma)(h_n),\ h_n \in H_n, n \geqq 1,$$

が成り立てば、strategies (π,γ) は optimal である.

Proof. $v(s_n) = v(\pi,\gamma)(s_n)$ とおく. 任意の $\gamma' \in \Gamma$ に対して、

$$E_{(\pi,\gamma')}\left[\sum_{n=1}^{N}\left\{ e^{-\alpha T_n} v(s_{n+1}) - E_{(\pi,\gamma')}\left[e^{-\alpha T_n} v(s_{n+1}) \,\middle|\, h_n \right]\right\} \,\middle|\, s_1 \right] = 0.$$

ここで、

$$\begin{aligned}
&E_{(\pi,\gamma')}\left[e^{-\alpha T_n} v(s_{n+1}) \,\middle|\, h_n \right] \\
&= e^{-\alpha T_{n-1}} P(\pi_n, \gamma_n') v(h_n) \\
&= e^{-\alpha T_{n-1}} L(\pi_n, \gamma_n') v(h_n) - e^{-\alpha T_{n-1}} r(\pi_n, \gamma_n')(h_n) \\
&\geqq e^{-\alpha T_{n-1}} v(s_n) - e^{-\alpha T_{n-1}} r(\pi_n, \gamma_n')(h_n),\ P_{(\pi,\gamma'),s_1}\text{-a.s. } h_n.
\end{aligned}$$

これを上の等式に代入して、Lemma 3.1を適用すれば、

$$v(\pi,\gamma)(s_1) = v(s_1) \leqq v(\pi,\gamma')(s_1),\quad \gamma' \in \Gamma,\ s_1 \in S$$

が成り立つ. 同様に、

$$v(\pi,\gamma)(s_1) \geqq v(\pi',\gamma)(s_1),\quad \pi' \in \Pi,\ s_1 \in S$$

が成り立つので、strategies (π,γ) はoptimalである. 故に定理が証明された.

任意の $(f,g) \in F \times G$ に対して、

$$L(f,g)v(s) = r(f,g)(s) + \iiint_{S\times A\times B} \beta(s,a,b,s')\, v(s')\, dp(s'|s,a,b) \times df(a|s)\, dg(b|s).$$

とおくと、$L(f,g)$ は S^w 上のoperatorである. ただし、

$$r(f,g)(s) = \iint_{A\times B} r(s,a,b)\, df(a|s)\, dg(b|s).$$

Theorems 3.1と3.2から次の結果を得る. これは、optimal stationary strategiesを特徴づける.

Corollary 3.1. stationary strategies (f^∞, g^∞) は、

$$L(f',g)v(f^\infty,g^\infty)(s_1) \leqq v(f^\infty,g^\infty)(s_1) \leqq L(f,g')v(f^\infty,g^\infty)(s_1),$$
$$(f',g') \in F\times G,\ s_1 \in S$$

が成り立つときに限りoptimalである.

このcorollaryから次の結果を得る. これはstochastic game に対して Shapley [5]が最初に証明したものである.

Corollary 3.2. ある $v \in S^\omega$ に対して

$$L(f', g)v(s_1) \leqq v(s_1) \leqq L(f, g')v(s_1), \quad (f', g') \in F \times G, \ s_1 \in S$$

が成り立てば、stationary strategies (f^∞, g^∞) はoptimalである.

Proof. $v(s_1) = L(f, g)v(s_1)$, $s_1 \in S$ が成り立つ. $L(f, g)$ は S^ω上の縮小写像で、一意の不動点 $v(f^\infty, g^\infty)(s_1)$, $s_1 \in S$ をもつ. したがって、$v(s_1) = v(f^\infty, g^\infty)(s_1)$, $s_1 \in S$. Corollary 3.1 より、(f^∞, g^∞) はoptimalで、v は game の値である.

Remark 3.1. Theorem 3.2 の十分条件は、次のように少し弱くできる. 任意の $(f, g) \in F \times G$ と $s_1 \in S$ に対して、Theorem 3.2 の右辺の不等式が $p_{(\pi, \gamma'), s_1}$ − a.s. h_n, $\gamma' \in \Gamma$ に対して成り立ち、左辺の不等式が $p_{(\pi', \gamma), s_1}$ − a.s. h_n, $\pi' \in \Pi$ に対して成り立つならば、(π, γ) は optimal である.

Remark 3.2. player II が stationary strategy g^∞ しか選択できず、$g(\{1\} | s, a, b, s') = 1$, $s, s' \in S$, $a \in A$, $b \in B$ であるとき,

semi-Markov game は Markov decision process になり、policy π は、strategies (π, g^∞) が optimal であるとき、またそのときに限り optimal である. $I(\pi)(s_1) = v(\pi, g^\infty)(s_1)$, $p_{\pi, s_1} = p_{(\pi, g^\infty), s_1}$ とおく. π が optimal ならば、Theorem 3.1 より任意の $f \in F$ と $s_1 \in S$ に対して、

$$I(f, \pi)(h_n) \leqq I(\pi)(h_n), \quad h_n \in H_n$$
$$I(\pi)(h_n) \leqq I(\pi_n, \pi)(h_n), \quad p_{\pi, s_1} - \text{a.s. } h_n$$

が成り立つ. π の optimality より、最初の不等式はすべての $h_n \in H_n$ に対して成り立つ. 一方, Theorem 3.2 と Remark 3.1 より、これらの条件は π が optimal であるための十分条件にもなっている. 更に、これらの条件は

$$I(\pi)(s_1) = \sup_{a \in A(s_1)} \left\{ r(s_1, a) + \int_S I(\pi)(s') \, dp(s' | s_1, a) \right\}, \quad s_1 \in S$$

と同値である. ここで、

$$r(s, a) = \int_B r(s, a, b) \, dg(b | s)$$
$$p(ds' | s, a) = \int_B p(ds' | s, a, b) \, dg(b | s)$$

この方程式は、policy π が optimal であるための必要十分条件であることが Blackwell [1] により証明されている.

References

[1] D. Blackwell, *Discounted dynamic programming*, Ann. Math. statist. 36 (1965), 226–235.

[2] L. P. J. Groenwegen & J. Wessels, *On the relation between optimality and saddle-conserving in Markov games*, Bonner Mathematische Schriften 98 (1977).

[3] L. P. J. Groenewegen, *Characterization of Optimal Strategies in Dynamic Games*, Mathematisch Centrum, Amsterdam, 1981.

[4] J. van Nunen & J. Wessels, *A note on dynamic programming with unbounded rewards*, Management Sci. 24 (1978), 576–580.

[5] L. Shapley, *Stochastic games*, Proc. Nat. Acad. Sci. 39 (1953), 1095–1100.

[6] K. Wakuta, *The Bellman's principle of optimality in the discounted dynamic programming*, J. Math. Anal. Appl. 125 (1987), 213–217.

[7] ________, *Arbitrary state semi-Markov decision processes with unbounded rewards*, Optimization 18 (1987), 447–454.