

Characterization of optimal strategies using their expected payoff

長岡高専 滝田和芳 (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。ここでは、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 $f: S \rightarrow A$ s.t. $f(s) \in A(s), s \in S$ なる Borel 可測写像の全体を F , $g: S \rightarrow B$ s.t. $g(s) \in B(s), s \in S$ なる Borel 可測写像の全体を G とし、 $F \neq \emptyset, G \neq \emptyset$ 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。

$g \in Q(R_+ \mid S \times A \times B \times S)$: 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

$$v(\pi, \gamma)(\rho_1) = E_{(\pi, \gamma)} \left[\sum_{n=1}^{\infty} e^{-\alpha T_{n-1}} r(\rho_n, a_n, b_n) \mid \rho_1 \right], \pi \in \Pi, \gamma \in \Gamma, \rho_1 \in S$$

で定義する. ここで、 ρ_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.

$$v(\pi', \gamma)(\rho_1) \leq v(\pi, \gamma)(\rho_1) \leq v(\pi, \gamma')(\rho_1), \pi' \in \Pi, \gamma' \in \Gamma, \rho_1 \in S$$

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 $w \geq 1$, p ($0 < p < 1$), $M > 0$ が存在する.

$$|t(s, a, b)| \leq M w(s),$$

$$\int_S \beta(s, a, b, s') w(s') d\phi(s' | s, a, b) \leq p w(s), s \in S, a \in A, b \in B.$$

$$\text{ここで, } \beta(s, a, b, s') = \int_0^\infty e^{-\lambda t} d\phi(t | s, a, b, s').$$

$H_1 = S$, $H_{n+1} = K \times R_+ \times H_n$ ($n \geq 1$) とする. H_n は n -th decision epoch までのシステムの可能な履歴の全体を表わす. 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[1]).

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

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

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

§ 3. The characterization of optimal strategies

$v(\pi, \gamma)(\rho_1)$, $\rho_1 \in S$, は次のようにかくことができる.

$$v(\pi, \gamma)(\rho_1) = E_{(\pi, \gamma)} \left[\sum_{n=1}^{\infty} e^{-\alpha T_{n-1}} r(\pi_n, \gamma_n)(h_n) \mid \rho_1 \right], \rho_1 \in S.$$

ここで、

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

$v \in S^W$ に対して、

$$P(\pi_n, \gamma_n)v(h_n) = \iiint_{S \times A \times B} \beta(\rho_n, a_n, b_n, \rho_{n+1}) v(\rho_{n+1}) d\beta(\rho_{n+1} \mid \rho_n, a_n, b_n) \\ \times d\pi_n(a_n \mid h_n) d\gamma_n(b_n \mid h_n),$$

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

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

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

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

とおく. たゞし,

$$R(\pi, \gamma)(\Delta) = E_{(\pi, \gamma)} \left[\sum_{n=1}^{\infty} e^{-\alpha T_{n-1}} r(\pi_n, \gamma_n)(h_n) \mid \Delta_1 = \Delta \right].$$

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 $\Delta_1 \in S$ に対して

$$\angle(f, \gamma_n) v(\pi, \gamma)(h_n) \leq v(\pi, \gamma)(\Delta_n) \leq \angle(\pi_n, g) v(\pi, \gamma)(h_n),$$

$$\phi(\pi, \gamma), \Delta_1 - \text{a.s. } h_n, n \geq 1,$$

が成り立つ.

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

$$v(\pi(n), \gamma')(\rho_1) = E_{(\pi(n), \gamma')} \left[\sum_{k=1}^n e^{-\alpha T_{k-1}} r(\pi_k, \gamma'_k)(h_k) + e^{-\alpha T_n} v(\pi(n), \gamma')(h_{n+1}) \mid \rho_1 \right].$$

$$\text{ここ } \tau, v(\pi(n), \gamma')(h_{n+1}) = v(\pi, \gamma^*)(\rho_{n+1})$$

$$\geq v(\pi, \gamma)(\rho_{n+1})$$

$$\geq v(\pi^*, \gamma)(\rho_{n+1})$$

$$= v(\pi, \{\gamma'_1, \dots, \gamma'_n, \gamma\})(h_{n+1}), \quad \mathcal{P}_{(\pi(n), \gamma'), \rho_1} - \text{a.s. } h_{n+1}.$$

$$\text{また } \tau, \pi_m^*(\cdot \mid h'_m) = \pi_{n+m}(\cdot \mid h_{n+1}, h'_m), \quad \gamma_m^*(\cdot \mid h'_m) = \gamma_{n+m}(\cdot \mid h_{n+1}, h'_m).$$

$(\pi(n), \gamma') \in (\pi, \{\gamma'_1, \dots, \gamma'_n, \gamma\})$ は、 n -th step まで同じ strategies なるので、

$$v(\pi(n), \gamma')(\rho_1) \geq v(\pi, \{\gamma'_1, \dots, \gamma'_n, \gamma\})(\rho_1) \geq v(\pi, \gamma)(\rho_1), \quad \rho_1 \in S.$$

同様に、

$$v(\pi', \gamma(n))(\rho_1) \leq v(\pi, \gamma)(\rho_1), \quad \rho_1 \in S.$$

故に、 $v(\pi(n), \gamma(n))(\rho_1) = v(\pi, \gamma)(\rho_1), \quad \rho_1 \in S$ が成り立ち、

$(\pi(n), \gamma(n))$ は optimal である。

$$(2\text{nd step}) \quad \pi^0 = \{\pi_1, \dots, \pi_{n-1}, \pi'\}, \quad \gamma^0 = \{\gamma_1, \dots, \gamma_{n-1}, \gamma'\},$$

$$\pi' \in \Pi, \quad \gamma' \in \Gamma \quad \text{に対して、}$$

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

$$v(\pi, \gamma)(h_n) = v(\pi, \gamma)(\rho_n), \quad \mathcal{P}_{(\pi, \gamma), \rho_1} - \text{a.s. } h_n$$

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、

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

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

同様に、

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

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。

$$\begin{aligned} v(\pi, \gamma)(h_n) &= v(\pi(n), \gamma(n))(h_n) \\ &= v(\pi(n), \gamma(n))(h_n) \\ &\leq v(\pi(n), \{\gamma_1, \dots, \gamma_{n-1}, g, \gamma\})(h_n) \\ &= L(\pi_n, g) v(\pi, \gamma)(h_n), \quad \phi(\pi, \gamma), P_1 - \text{a.s. } h_n. \end{aligned}$$

同様にして、

$$v(\pi, \gamma)(h_n) \geq L(\pi, \gamma_n) v(\pi, \gamma)(h_n), \quad \phi(\pi, \gamma), P_1 - \text{a.s. } h_n.$$

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

$$\lim_{n \rightarrow \infty} E_{(\pi', \delta')} [e^{-\alpha T_{n-1}} v(\pi, \delta)(\Delta_n) | \Delta_1] = 0$$

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 $\Delta_1 \in S$ に対して,

$$L(f, \delta_n) v(\pi, \delta)(h_n) \leq v(\pi, \delta)(\Delta_n) \leq L(\pi_n, g) v(\pi, \delta)(h_n), h_n \in H_n, n \geq 1,$$

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

Proof. $v(\Delta_n) = v(\pi, \delta)(\Delta_n)$ とおく. 任意の $\delta' \in P$ に対して,

$$E_{(\pi, \delta')} \left[\sum_{n=1}^{\infty} \left\{ e^{-\alpha T_n} v(\Delta_{n+1}) - E_{(\pi, \delta')} [e^{-\alpha T_n} v(\Delta_{n+1}) | h_n] \right\} \middle| \Delta_1 \right] = 0.$$

ここで,

$$\begin{aligned} & E_{(\pi, \delta')} [e^{-\alpha T_n} v(\Delta_{n+1}) | h_n] \\ &= e^{-\alpha T_{n-1}} P(\pi_n, \delta'_n) v(h_n) \\ &= e^{-\alpha T_{n-1}} L(\pi_n, \delta'_n) v(h_n) - e^{-\alpha T_{n-1}} r(\pi_n, \delta'_n)(h_n) \\ &\geq e^{-\alpha T_{n-1}} v(\Delta_n) - e^{-\alpha T_{n-1}} r(\pi_n, \delta'_n)(h_n), \phi(\pi, \delta'), \Delta_1 - \text{a.s. } h_n. \end{aligned}$$

これを上の等式に代入して、Lemma 3.1 を適用すれば、

$$v(\pi, \delta)(\lambda_1) = v(\lambda_1) \leq v(\pi, \delta')(\lambda_1), \quad \delta' \in T, \lambda_1 \in S$$

が成り立つ。同様に、

$$v(\pi, \delta)(\lambda_1) \geq v(\pi', \delta)(\lambda_1), \quad \pi' \in \Pi, \lambda_1 \in S$$

が成り立つので、strategies (π, δ) 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、

$$L(f, g)v(\lambda) = t(f, g)(\lambda) + \iiint_{S \times A \times B} \beta(\lambda, a, b, \lambda') v(\lambda') dp(\lambda' | \lambda, a, b) \\ \times df(a|\lambda) dg(b|\lambda).$$

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

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

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、

$$L(f', g)v(f^\infty, g^\infty)(\lambda_1) \leq v(f^\infty, g^\infty)(\lambda_1) \leq L(f, g')(v(f^\infty, g^\infty)(\lambda_1), \\ (f', g') \in F \times G, \lambda_1 \in S$$

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 $v \in S^W$ に対して

$$L(f', g)v(\rho_1) \leq v(\rho_1) \leq L(f, g')v(\rho_1), \quad (f', g') \in F \times G, \rho_1 \in S$$

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

Proof. $v(\rho_1) = L(f, g)v(\rho_1)$, $\rho_1 \in S$ が成り立つ。 $L(f, g)$ は S^W 上の縮小写像で、一意の不動点、 $v(f^*, g^*)(\rho_1)$, $\rho_1 \in S$ をもつ。したがって、 $v(\rho_1) = v(f^*, g^*)(\rho_1)$, $\rho_1 \in S$ 。 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 $\rho_1 \in S$ に対して、Theorem 3.2 の右辺の不等式が $\phi(\pi, \delta'), \rho_1 - a.s. h_n$, $\delta' \in \Gamma$ に対して成り立ち、左辺の不等式が $\phi(\pi', \delta), \rho_1 - a.s. h_n$, $\pi' \in \Pi$ に対して成り立つならば、 (π, δ) は optimal である。

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

semi-Markov game は Markov decision process にたり、
 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。 $I(\pi)(\Delta_1) = v(\pi, g^\infty)(\Delta_1)$,
 $\phi_{\pi, \Delta_1} = \phi(\pi, g^\infty), \Delta_1$ とおく。 π が optimal ならば、Theorem 3.1
 より任意の $f \in F$ と $\Delta_1 \in S$ に対して、

$$I(f, \pi)(\Delta_n) \leq I(\pi)(\Delta_n), \quad \Delta_n \in H_n$$

$$I(\pi)(\Delta_n) \leq I(\pi_n, \pi)(\Delta_n), \quad \phi_{\pi, \Delta_1} = a, a \in H_n$$

が成り立つ。 π の optimality より、最初の不等式はすべての
 $\Delta_n \in H_n$ 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

$$I(\pi)(\Delta_1) = \sup_{a \in A(\Delta_1)} \left\{ r(\Delta_1, a) + \int_S I(\pi)(\Delta') d\phi(\Delta' | \Delta_1, a) \right\}, \quad \Delta_1 \in S$$

と等値である。 ここで、

$$r(\Delta, a) = \int_B r(\Delta, a, b) dg(b | \Delta)$$

$$\phi(d\Delta' | \Delta, a) = \int_B \phi(d\Delta' | \Delta, a, b) dg(b | \Delta)$$

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. Groenewegen & 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 Nuen & 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. Wakata, 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.