Characterization of optimal strategies
using their expected payoff

長岡高専 浦田和茂 (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、二人零和 game において、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られる。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 であるための良し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 optimaity 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

$A(\cdot), B(\cdot): S \rightarrow A$ (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。$K = \{(a, a, b) \mid a \in S, a \in A(s), b \in B(s)\}$ は Borel set と仮定する。また$

\phi: S \rightarrow A$. あたって $\phi(s) \in A(s), a \in S$ なる Borel 可測写像の全体を $\Phi$, $g: S \rightarrow B$. あたって $g(s) \in B(s), a \in S$ なる Borel 可測写像の全体を $G$ とし、$\Phi \cap G$ と仮定する。

$\mathcal{F} \in \mathcal{Q}(S \times S \times A \times B)$: the law of motion of state. ただし、

$\mathcal{Q}(Y | X)$ は $X$ から $Y$ への transition probabilities の全体を表す。

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

1
\( r : S \times A \times B \) 上の Borel 可測関数で、player I への payoff function．
\( \lambda > 0 \) : discount factor．

\( \Pi \) は player I の strategies の全体、\( \Gamma \) 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

\[ v(\pi, \gamma)(a_i) = E_{(\pi, \gamma)} \left[ \sum_{n=0}^{\infty} e^{-\lambda T_n} \, r(a_n, an, bn) \, | a_i \right], \pi \in \Pi, \gamma \in \Gamma, a_i \in S \]

で定義する．ここで、\( a_n, an, bn \) 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．

\[ v(\pi, \gamma)(a_i) \leq v(\pi, \gamma')(a_i) \leq v(\pi', \gamma)(a_i), \pi \in \Pi, \gamma \in \Gamma, a_i \in S \]

を満たす strategies \((\pi, \gamma)\) があれば、\( \pi, \gamma \) はそれぞれ player I, II に対して optimal であるという．次の条件がなり立つものと仮定する．
Condition (1) (cf. Nunen and Wessels [6]). 次の条件を満たす $S$ 上の Borel 可測関数 $w \equiv 1, \rho (0 < \rho < 1), M > 0$ が存在する。

$$\| \rho (a, b, b') \| \leq M w(a),$$

$$\int_S \rho (a, b, b') w(x) dp(a', b, b) \leq \rho w(a), a \in S, a \in A, b \in B.$$ ここで，$$\rho (a, b, b') = \int_0^\infty e^{-ax} dp(a', b, b').$$

$H_n = S, H_{nt} = K \times R^+ \times H_n (n \geq 1)$ とする。$H_n$ は $n$-th decision epoch までのシステムの可能な選択の全体を表す。$H_n$ 上の Borel 可測関数 $v$ に対して

$$\| v \|_{H_n} = \sup_{h_{nt} \in H_n} | v(h_{nt}) | w(a_n)^{-1}$$

ここで，$\| v \|_{H_n} \leq \infty$ なるすべての $v$ からなる Banach space を $H^w_n$ で表す。特に，$n=1$ のとき，$H^w_1 = S^w_1, \| \|_1$ と $\| \|$ とがく。Condition (1) により次の結果が得られる（cf. Waku[7]).

Proposition 2.1. 任意の strategies $(\pi, s)$ に対して，

$$\zeta = \sup_{n \geq 0} T_n$$ は $a.s.$ で infinite である。
Proposition 2.2. 任意のstrategies \((\pi, \xi)\)に対して、

\[ \| \nu(\pi, \xi) \| \leq M / (1 - p). \]

§ 3. The characterization of optimal strategies

\(\nu(\pi, \xi)(\alpha_i), \alpha_i \in S \), は次のようにかくことができる。

\[ \nu(\pi, \xi)(\alpha_i) = E(\pi, \xi) \left[ \sum_{n=1}^{\infty} e^{-\Delta n-1} r(\pi_n, \xi_n)(\alpha_n) \mid \alpha_i \right], \alpha_i \in S. \]

ここで、

\[ r(\pi_n, \xi_n)(\alpha_n) = \iiint_{A \times B} r(\alpha_n, a_n, b_n) \, d\pi_n(a_n \mid \alpha_n) \, d\tau_n(b_n \mid \alpha_n), \quad n \in \mathbb{N}. \]

\(\nu \in S^w\)に対して、

\[ P(\pi_n, \xi_n) \nu(\alpha_n) = \iiint_{S \times A \times B} \beta(\alpha_n, a_n, b_n, \alpha_n) \nu(\alpha_n) \, d\beta(\alpha_n \mid a_n, b_n, \alpha_n) \]

\[ \times d\pi_n(a_n \mid \alpha_n) \, d\tau_n(b_n \mid \alpha_n), \]

\[ L(\pi_n, \xi_n) \nu(\alpha_n) = r(\pi_n, \xi_n)(\alpha_n) + P(\pi_n, \xi_n) \nu(\alpha_n) \]

とおく。\(\nu\)は\(S^w\)から\(H^w\)へのoperatorである。また、
$$v(\pi, \gamma)(\lambda_n) = E_{(\pi, \gamma)} \left[ \sum_{k=n}^{\infty} e^{-\lambda(T_k-1-T_{k-1})} r(\pi_k, \gamma_k)(f_{\lambda_k}) | \lambda_n \right],$$

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

ここで、ただし、

$$R(\pi, \gamma)(\lambda) = E_{(\pi, \gamma)} \left[ \sum_{n=1}^{\infty} e^{-\lambda T_{n-1}} r(\pi_n, \gamma_n)(f_{\lambda_n}) \bigg| \lambda_1 = \lambda \right].$$

最初に、strategies $$(\pi, \gamma)$$ 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 対して用いたものと一般の strategies へ拡張したものである。

**Theorem 3.1.** strategies $$(\pi, \gamma)$$ が optimal ならば、任意の $$(f, g) \in F \times G$$ と $\lambda \in S$$ に対して

$$\angle (f, \gamma_n) v(\pi, \gamma)(\lambda_n) \leq v(\pi, \gamma)(\lambda_n) \leq \angle (f, g) v(\pi, \gamma)(\lambda_n),$$

$$\Phi(\pi, \gamma), \lambda \sim \alpha_1, \lambda_n, n \geq 1,$$

が成り立つ。

**Proof.** (1st step) $$(\pi(n) = \{ \pi_1, \cdots, \pi_n, \pi_1, \pi_2, \cdots \}, \gamma(n) = \{ \gamma_1, \cdots, \gamma_n, \gamma_1, \gamma_2, \cdots \}$$ は optimal strategies であることを示す。任意の $$(f, g)$$ に対して、
\[ \nu(\pi(n), \gamma')(\omega_i) = E_{(\pi(n), \gamma')} \left[ \sum_{k=1}^{n} e^{-\alpha T_k} \nu_{(\pi_k, \gamma_k')}((h_k)) + e^{-\alpha T_n} \nu_{(\pi(n), \gamma')(h_n)} | A_i \right]. \]

ここで、\[ \nu(\pi(n), \gamma')(h_{n+1}) = \nu(\pi, \gamma')(A_{n+1}) \]
\[ \leq \nu(\pi, \gamma)(A_{n+1}) \]
\[ \leq \nu(\pi', \gamma)(A_{n+1}) \]
\[ = \nu(\pi, \{ \gamma'_1, \ldots, \gamma'_n, \gamma \})(h_{n+1}), \quad \hat{\theta}((\pi(n), \gamma'), A_i \sim A, h_{n+1}). \]

ただし、\[ \pi_n^* (\cdot | h_n') = \pi_{n+1}^* (\cdot | h_{n+1}, h_n'), \quad \gamma_n^* (\cdot | h_n') = \gamma_{n+1}^* (\cdot | h_{n+1}, h_n'). \]

\((\pi(n), \gamma') \in (\pi, \{ \gamma_1', \ldots, \gamma_n', \gamma \}) \) は、n-th step まで "は同じ strategies" 立ち、
\[ \nu(\pi(n), \gamma')(\omega_i) \leq \nu(\pi, \{ \gamma_1', \ldots, \gamma_n', \gamma \})(A_i) \leq \nu(\pi, \gamma)(A_i), \quad A_i \in S. \]

同様に、
\[ \nu(\pi', \gamma(n))(A_i) \leq \nu(\pi, \gamma)(A_i), \quad A_i \in S. \]
故に、\[ \nu(\pi(n), \gamma(n))(A_i) = \nu(\pi, \gamma)(A_i), \quad A_i \in S \] かつ " optimum " である。

\((2nd \ step) \quad \pi^o = \{ \pi_1, \ldots, \pi_{n-1}, \pi' \}, \quad \gamma^o = \{ \gamma_1, \ldots, \gamma_{n-1}, \gamma' \}, \quad \pi' \in \Pi, \quad \gamma' \in \Gamma \) に対し、
\[ \nu(\pi^o, \gamma)(h_n) \leq \nu(\pi, \gamma)(h_n) \leq \nu(\pi, \gamma^o)(h_n), \]
\[ \nu(\pi, \gamma)(h_n) = \nu(\pi, \gamma)(A_n), \quad \hat{\theta}((\pi, \gamma), A_i \sim A, h_n). \]
が放り立つことを示す。最初の不等式は明らかに、特に、
\[ \pi^0 = \pi(n-1), \quad \gamma^0 = \gamma(n-1) \] とおけば、
\[ \nu(\pi, \gamma)(fn) \leq \nu(\pi(n-1), \gamma)(fn) = \nu(\pi, \gamma')(fn) \leq \nu(\pi, \gamma)(fn) \]
ただし、\[ \gamma'(\cdot | fn) = \gamma_{n+m-1}(\cdot | fn, fn') \]
同様に、
\[ \nu(\pi, \gamma)(fn) \leq \nu(\pi, \gamma(n-1))(fn) = \nu(\pi', \gamma)(fn) \leq \nu(\pi, \gamma)(fn) \]
故に等式が得られる。
以上のことから、次のことが放り立つ。
\[ \nu(\pi, \gamma)(fn) = \nu(\pi(n), \gamma(n))(fn) \]
\[ = \nu(\pi(n), \gamma(n))(fn) \]
\[ \leq \nu(\pi(n), \{ \gamma_1, \ldots, \gamma_{n-1}, \gamma, \gamma' \})(fn) \]
\[ = \nu(\pi(n), \gamma)(fn), \quad \nu(\pi, \gamma)(fn), \quad \gamma \in \gamma, \quad \gamma \in \gamma \]
同様にして、
\[ \nu(\pi, \gamma)(fn) \leq \nu(\pi, \gamma)(fn), \quad \nu(\pi, \gamma)(fn), \quad \gamma \in \gamma, \quad \gamma \in \gamma \]
故に定理が証明された。
次に、strategies \( (\pi, \gamma) \) 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、
\[
\lim_{n \to \infty} E_{(\pi', \delta')} \left[ e^{-\alpha T_n-1} v(\pi, \delta) \mid \delta_i \right] = 0
\]
がなり立つ。

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と 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 \(\delta_i \in S\) に対して、

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

が成り立つ。strategies \((\pi, \delta)\) は optimal である。

Proof. \(v(\delta_n) = v(\pi, \delta)(\delta_n)\) とおく。任意の \(\delta \text{ と } \pi'\) に対して、

\[
E_{(\pi, \delta)} \left[ \sum_{n=1}^{\infty} \left\{ e^{-\alpha T_n} v(\delta_{n+1}) - E_{(\pi, \delta')} \left[ e^{-\alpha T_n} v(\delta_{n+1}) \mid h_n \right] \right\} \mid \delta_i \right] = 0.
\]

ここで、

\[
E_{(\pi, \delta')} \left[ e^{-\alpha T_n} v(\delta_{n+1}) \mid h_n \right]
\]

\[
= e^{-\alpha T_n-1} P(\pi_n, \delta') v(h_n)
\]

\[
= e^{-\alpha T_n-1} L(\pi_n, \delta') v(h_n) - e^{-\alpha T_n-1} v(\pi_n, \delta')(h_n)
\]

\[
= e^{-\alpha T_n-1} v(\delta_n) - e^{-\alpha T_n-1} v(\pi_n, \delta')(h_n), \varphi(\pi, \delta), \delta_i - \alpha . \delta . h_n.
\]
これを上の等式に代入して、Lemma 3.1 を適用すれば、

\[ \nu(\pi, \gamma)(\lambda) = \nu(\lambda) \leq \nu(\pi, \gamma')(\lambda), \quad \gamma' \in \Gamma, \quad \lambda \in S \]

が成り立つ。同様に、

\[ \nu(\pi, \gamma)(\lambda) \geq \nu(\pi', \gamma)(\lambda), \quad \pi', \gamma \in \Pi, \quad \lambda \in S \]

が成り立つので、strategies (\pi, \gamma) 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、

\[
L(f, g)(\lambda) = r(f, g)(\lambda) + \iint_{S \times A \times B} \beta(p, a, b, p') \nu(\delta) d\beta(p'|a, \lambda, b) \\
\times df(a|\lambda) \, dq(b|\lambda).
\]

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

\[
r(f, g)(\lambda) = \iint_{A \times B} r(\lambda, a, b) \, df(a|\lambda) \, dq(b|\lambda).
\]

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、

\[
L(f', g')(f^*, g^*)(\lambda) \leq \nu(f^*, g^*)(\lambda) \leq L(f, g')(f^*, g^*)(\lambda),
\]

\((f', g') \in F \times G, \lambda \in S\)

が成り立つときに限り optimal である。
この corollary ら次節の結果を得る。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

\[ L(f', g) v(\alpha) \leq v(\alpha) \leq L(f, g') v(\alpha), \quad (f', g') \in F \times G, \alpha \in S \]

が成り立つば、stationary strategies \( (f^\circ, g^\circ) \) は optimal である。

Proof. \( v(\alpha) = L(f, g) v(\alpha), \alpha \in S \) が成り立つ。\( L(f, g) \) は

\( S^w \) 上の縮小写像で、一意の不動点 \( v(f^\circ, g^\circ)(\alpha), \alpha \in S \) をもつ。
したがって、\( v(\alpha) = v(f^\circ, g^\circ)(\alpha), \alpha \in S \)。Corollary 3.1 により、

\( (f^\circ, g^\circ) \) は optimal で、\( v \) は game の値である。

Remark 3.1. Theorem 3.2 の十分条件は、次のようになり

\( (f, \alpha) \) と \( F \times G \) に対して、Theorem 3.2 の右辺の不等式が \( \phi(\pi, \alpha) \) 一意で \( \alpha \in A \) に対して成り立ち、左辺の不等式が \( \phi(\pi, \alpha') \) 一意で \( \alpha' \in A \) に対して成り立ちなければ、(\( \pi, \alpha \)) は optimal である。

Remark 3.2. player II が stationary strategy \( g^\circ \) しか選択できない、\( g \mid \{ l\} | a, a', b, b' = 1, a, a' \in S, a \in A, b \in B \) であるとき、
semi-Markov gameはMarkov decision processとかわり、policyπは、strategies（π, q̃*）がoptimalであるとき、またそのときに限りoptimalである。I(π)(α) = v(π, q̃*)(α)、
φπ,α = φ(π, q̃*), α, i でおく。πがoptimalならば、Theorem3.1より任意の非空o,s,tスに対しても、
I(t, π) (α) ≤ I(π)(αn), 仮のH、
I(π)(αn) ≤ I(π)(α, x) (αn), 仮のa1 = a, 仮のa
が成り立つ。πのoptimalityより、最初の不等式はすべての仮のH、仮のに対しても成り立つ。一方、Theorem3.2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なっている。更に、これらの条件は

I(π)(α) = sup_{α ∈ A(β)} \left\{ t(α, a) + \int_S I(π)(α')dφ(α'|α, a) \right\}, α ∈ S

と同値である。ここで、

φ(α, a) = \int_B φ(α, a, b) dg(b|α)

φ(α, a) = \int_B φ(α, a, b) dg(b|α)

少し方程式は、policyπ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。

12
References


