The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

強化学習プレイヤー同士の繰り返し Battle of the Sexes Game におけるコーディネーションの実現

Emergence of Coordination in Iterated Battle of the Sexes Game with Reinforcement Learning Players

真隅暁 *1*2 Akira Masumi　日髙昇平 *2 Shohei Hidaka

*1沖縄工業高等専門学校メディア情報工学科
National Institute of Technology, Okinawa College, Department of Information Engineering

*2北陸先端科学技術大学院大学知識科学研究科
Japan Advanced Institute of Science and Technology, School of Knowledge Science

A coordination between individuals with conflict of interests is a crucial issue for achieving a joint cooperation. The Iterated Battle of the Sexes Game (IBoS) is a game-theoretic model which formulates a difficulty of the coordination between individuals in repeated interaction, in which an existence of multiple Nash equilibria pose a problem of equilibrium selection. Regarding this problem, recent empirical studies have shown that human game players with bounded rationality can achieve a coordination based on experiences gained through repeated interaction. In this study, we introduced a reinforcement learning (RL) model as a player of the IBoS and analyzed an equilibrium structure of the game. As a result of theoretical and numerical analysis, we found a cooperative turn-taking is supported as a unique Nash equilibrium of the game, in which each of individual satisfy their interests alternatively. This results provide a possible mechanism to achieve a feasible cooperative turn-taking.

1. 背景

私たちは，一人では達成できないタスクを達成するためにグループを形成する．例えば身近な例として，1 人では運ぶことができない大きな荷物を運ぼうとしている場合などが挙げられる．こういった共同的な行為を実現するには，各人が互いの行動を合わせるコーディネーションを必要とする．しかし，例えば荷物が複数あって，運ぶ際の優先順位が各人で異なっているなどの利害対立がある場合には，コーディネーションの実現は容易ではなくなる．

一方で，各人が同じ状況に繰り返し直面する場合には，より協力的なコーディネーションの仕方があり得る．それは，互いにとって望ましい状態を交互に実現するような，協力的なターンテイキングである．実際，このような振る舞いは，村落共同体における希少資源の共同利用や [Ostrom 90]，人被験者を対象とした実験 [Helbing 05] においても観察されている．だがこの場合には，各主体に対し，交替を拒んで利益を独占するインセンティブが働くため，これが実現されるメカニズムは自明でない．では，現実に観察される協力的なターンテイキングはいかにして実現されているのだろうか？

利害対立下のコーディネーションの問題を定式化したゲームとして，Battle of the Sexes Game (BoS) およびその繰り返しゲームである Iterated BoS (IBoS) がある (詳細な説明は 2 章を参照)．IBoS では，コーディネーションは，各プレイヤーが互いに行動を変える動機を持たない状態であるナッシュ均衡となる．しかし IBoS には複数のナッシュ均衡が存在するため，“ある均衡を他より優先する十分な根拠がない状況で，いずれの均衡を選択すべきか” という均衡選択が問題となる．

先行研究では，IBoS においてターンテイキングがサブゲーム完全均衡となることが示されているが [Lau 12]，均衡の存在が示されるにとどまり，均衡選択の問題が解決されていない．また，現実の人間の振る舞いを理解するうえでは，人の振る舞いが，認知能力の限界や情報の不完全性に起因して，限定合理的であることを考慮する必要がある [Kahneman 03]．

これを踏まえて本研究では，IBoS のプレイヤーとして，自らの行動に対する報酬をもとに最適な行動を獲得する強化学習 [Sutton 98] を導入し，このプレイヤー同士のゲームによって IBoS におけるコーディネーションが実現可能かどうかを分析する．このアプローチは，近年，人被験者による行動実験によって，繰り返しゲームをプレイする経験を通じて人がコーディネーションを実現できるという知見 [Erev 98] からも支持される．しかし，強化学習の導入により IBoS は学習パラメータを戦略とするゲームへと変換されるため，このゲームの均衡の構造は自明でない．従って本研究では，強化学習プレイヤー同士の IBoS は唯一のナッシュ均衡を持つか，およびその振る舞いとしてターンテイキングが生じるかどうかを分析する．

2. モデル

2.1 Iterated Battle of the Sexes Game (IBoS)

Battle of the Sexes Game (BoS) は 2×2 ゲームで，プレイヤー $i \in \{1, 2\}$ は純粋戦略 $a_i \in \{X, Y\}$ のうちいずれか一つを選択する．BoS では，もし 2 人のプレイヤーが同じ戦略を選ぶことができれば高い利得が得られるが，互いに好ましい戦略が異なっている．それぞれ，自分の好む戦略で互いの選択を一致させることができればもっとも高い利得を得られるが，両者の選択が異なる場合には低い利得しか得られない．ここで，両者の選択が一致することをコーディネーションの実現と呼び，そうでない場合を失敗と呼ぶ．BoS には，戦略ペア (X, X) と (Y, Y) という，コーディネーションの実現に対応した 2 つの純粋戦略ナッシュ均衡があり，各均衡では互いのプレイヤーが各自の選好を反映した利得を得る (表 1 を参照)．

繰り返しの BoS (Iterated BoS, IBoS) は，BoS をステージゲームとしてこれを繰り返しプレイするゲームである．IBoS において生じ得る協力的ターンテイキングは，BoS の 2 つのナッシュ均衡の間を移り変わる振る舞いとして捉えられる．本

連絡先: 真隅暁，石川県能美市旭台 1-1，a-masumi@jaist.ac.jp

1\2	X	Y
X	$(1, q)$	$(0, 0)$
Y	$(0, 0)$	$(q, 1)$

表 1: 本研究で用いる BoS の利得行列．$0 < q < 1$.

研究では，表 1 で表される BoS を用い，これを無限回繰り返しプレイする IBoS を分析する．

2.2 強化学習プレイヤー同士の IBoS (RLIBoS)

本研究では，強化学習プレイヤー同士の IBoS を分析する．強化学習プレイヤーは各戦略に対する価値関数の条件付き確率に従って意思決定を行う．本研究では，プレイヤー i は，各時刻 t で以下の条件付き確率に従って戦略 $a_i(t) \in \{X, Y\}$ を選択するものとする；

$$P_{\alpha_i,\beta_i}\left(a_i(t) = X \mid Q_{i,X}(t), Q_{i,Y}(t)\right) = \frac{e^{\beta_i Q_{i,X}(t)}}{e^{\beta_i Q_{i,X}(t)} + e^{\beta_i Q_{i,Y}(t)}}, \tag{1}$$

ここで $\alpha_i \in [0, 1]$ は学習率，$\beta_i \in [0, \infty)$ は鋭敏性を表すパラメータで，$Q_{i,a}(t) \in \mathbb{R}$，$i = 1, 2$, $a \in A$ は価値関数である．$\beta_i = 0$ のとき，プレイヤー i は戦略 X か Y を一様ランダムに選択する．他方，$\beta_i \to \infty$ のとき，プレイヤー i が戦略 a を選ぶ条件付き確率 $P_{\alpha_i,\beta_i}(a \mid Q_{i,X}(t), Q_{i,Y}(t)) \to 1$ である．

価値関数 $(Q_{i,X}(t), Q_{i,Y}(t))$ は，以下に従って更新される；

$$(Q_X, Q_Y) := \begin{cases} ((1-\alpha)Q_X + \alpha r, Q_Y), & a = X \\ (Q_X, (1-\alpha)Q_Y + \alpha r), & \text{otherwise.} \end{cases} \tag{2}$$

パラメータ α_i は学習率で，得られた利得に対し $Q_{i,a}(t)$ をどの程度更新するかを定める．$\alpha_i = 0$ のとき，価値関数は時刻 $t = 0$ から更新されず，$Q_{i,a}(t) = Q_{i,a}(0)$ となる．$\alpha_i = 1$ のときは，$Q_{i,a}(t) = r_{i,a}(t-1)$ であり，戦略 a の価値関数は 1 時刻前に得た利得の値となる．ここで $r_{i,a}(t-1)$ は，プレイヤー i が時刻 $t-1$ で戦略 a を選んだ際に得た利得を表す．

強化学習プレイヤー同士の IBoS におけるプレイヤー i の期待利得は，各プレイヤーの学習パラメータの組 (α_i, β_i) から定まる．従ってこの場合には，学習率 α_i と鋭敏性 β_i がゲームの実質的な戦略であると考えることができる．本研究では，この α_i と β_i を戦略としたゲームを Reinforcement Learner's IBoS (RLIBoS) と呼び，分析の対象とする．

2.3 マルコフ過程としての定式化

RLIBoS では，各プレイヤーは式 (1) に従って確率的な意思決定を行う．従って RLIBoS の振る舞いは，学習パラメータの組 $\theta := (\alpha_1, \alpha_2, \beta_1, \beta_2)$，価値関数の初期値 $Q_0 := (Q_{1,X}(0), Q_{1,Y}(0), Q_{2,X}(0), Q_{2,Y}(0))$，および利得行列 ω の 3 つ組 (θ, Q_0, ω) が与えられると，価値関数の組 $Q := (Q_{1,X}, Q_{1,Y}, Q_{2,X}, Q_{2,Y})$ を状態とするマルコフ過程として定式化できる．

つまり，上記の 3 つ組 (θ, Q_0, ω) で定まる RLIBoS は，状態空間 $S := \{Q = (Q_{1,X}, Q_{1,Y}, Q_{2,X}, Q_{2,Y}) \in \mathbb{R}^4\}$ のうえのマルコフ過程で，状態 Q から Q' への遷移確率は

$$P(Q' \mid Q) := \prod_{i=1,2} P_{\alpha_1,\beta_1}(a_i \mid Q_{i,X}, Q_{i,Y}),$$

で与えられる．ここで Q' は式 (2) に従って更新された価値関数を表す．本研究では，状態空間 S のうち，定常分布において確率が 0 でないものを指して状態と呼ぶこととする．

3. マルコフ過程の数理解析

3.1 学習率 α に基づくマルコフ過程の分類

前章で定式化したマルコフ過程は，各プレイヤーの学習率 α_i が 0 か 1 をとる場合，著しく状態数が減り分析が容易になる．この特殊ケースも含め，上記のマルコフ過程は，プレイヤー 1，2 の学習率の組 $\alpha = (\alpha_1, \alpha_2)$ に対して主に 4 つの場合に分類できる．以下ではこの 4 つのケースについて説明する．

3.1.1 BoS by initial values: $\alpha = (0, 0)$

この場合，価値関数 Q は初期値 Q_0 から更新されないため，マルコフ過程の状態は Q_0 のみとなる．従ってこのとき，プレイヤーの実質的な戦略は $(Q_{i,X}(0), Q_{i,Y}(0)) \in \mathbb{R}^2$ となり，$t = 0$ 以降は Q_0 から定まる条件付き確率に従って確率的な意思決定を行う．これは，混合戦略を許す 1 ステージの BoS と本質的に同じゲームであるため，本研究での分析からは除外する．

3.1.2 Quick RLIBoS: $\alpha = (1, 1)$

この場合には，各プレイヤーの価値関数 $Q_{i,a}(t)$ は，式 (2) から，1 時刻前に得た利得の値となる．従ってこの RLIBoS は，直前に得た利得を価値関数とする "学習の速い" プレイヤー同士のゲームと考えることができる．

3.1.3 Slow-and-fast RLIBoS: $\alpha = (0, 1)$ (or $(1, 0)$)

この場合には，プレイヤー 1 は価値関数を更新せず，一方でプレイヤー 2 は 1 時刻前に得た利得を価値関数とする．言い換えると，この場合の RLIBoS は，学習の遅いプレイヤーと速いプレイヤーの間のゲームとなる．このケースは，学習率が非対称な場合の特殊ケースだと考えることができる．

3.1.4 General RLIBoS: $0 < \alpha_1, \alpha_2 < 1$

これは上記 3 つのケースと比べ最も一般的なケースで，有限個の状態を持つマルコフ過程へと帰着することはできない．

3.2 特殊ケースの分析

3.2.1 Slow-and-fast RLIBoS: $\alpha = (0, 1)$

ここでは $\alpha = (0, 1)$ のケース，すなわち slow-and-fast RLIBoS を分析する．上述の通りこの場合には，プレイヤー 1 は価値関数を更新せず，一方でプレイヤー 2 は 1 時刻前に得た利得を価値関数とする．そして利得行列 (表 1) から，$Q_{2,X}(t) \in \{0, q\}$，$Q_{2,Y}(t) \in \{0, 1\}$ であるため，slow-and-fast RLIBoS の状態空間 $S_{0,1}(Q_{1,X}, Q_{1,Y})$ は，以下の 4 つの状態のみからなることがわかる；

$$S_{0,1}(Q_{1,X}, Q_{1,Y}) =$$
$$\{(Q_{1,X}(0), Q_{1,Y}(0), Q_{2,X}(t), Q_{2,Y}(t)) | Q_{2,X} \in \{0, q\}, Q_{2,Y} \in \{0, 1\}\}$$

slow-and-fast RLIBoS では，プレイヤー 1 は価値関数の初期値 $Q_{1,a}(0)$ を戦略とし，他方でプレイヤー 2 は鋭敏性 $\beta_2 \in [0, \infty)$ を戦略とする．いま，プレイヤー 1 の選択においては戦略 X と Y の価値関数の差のみが問題となるため，$Q_{1,Y} = 0$ とすると，プレイヤー 1 の価値関数の組 $(Q_{1,X}, 0)$ と，プレイヤー 1 が戦略 X を選ぶ確率 $P(X) \in [0, 1]$ が一対一に対応する．プレイヤー 1 は $Q_{1,X}(0)$ の選択によって $P(X)$ を任意の値にとることができるため，表 1 で与えられた利得行列のもとで自身の期待利得を最大化するように，$P(X) = 1$ を与える $Q_{1,X}(0)$ を選ぶ．すなわち，$Q_{1,X} \to \infty$ とする．他方プレイヤー 2 はプレイヤー 1 の決定に従うほかなく，この条件のもとで自身の期待利得を最大化するために $\beta_2 \to \infty$ とする．従って，slow-and-fast RLIBoS においては，$(Q_{1,X}, Q_{1,Y}) = (\infty, 0)$ および $\beta_2 \to \infty$ が唯一のナッシュ均衡となる．

3.2.2 Slow-and-fast RLIBoS: $\alpha = (0, \alpha_2)$, $0 < \alpha_2 < 1$

上記の議論は $\alpha = (0, \alpha_2)$, $0 < \alpha_2 < 1$ の場合についても成り立つ．この場合には，$Q_{2,a}$ の値に応じた無数の状態が存在するが，プレイヤー 1 が確率 1 で戦略 X を選ぶため，プレイヤー 2 の期待利得を最大化する価値関数は $\alpha_2 = 0$ の場合と変わらない．このことから，$\alpha = (0, \alpha_2)$, $0 < \alpha_2 < 1$ のケースは，$\alpha = (0, 1)$ と同じ帰結をもたらすことがわかる．

さらにこの分析から，$\alpha_1 = 0$ のもとでは，α_2 の値が小さいほどプレイヤー 2 の期待利得は大きくなることがわかる．利得行列および式 (2) の α_1，α_2 に対する対称性，さらに期待利得が α_i について連続であることを考慮すると，プレイヤー i の期待利得は，$\alpha_i = 0$ の近傍 $\alpha \in [0, \epsilon_1] \times [0, \epsilon_2]$ において最大値をとると考えられる (ϵ_1，ϵ_2 は十分小さい定数)．

3.2.3 Quick RLIBoS

ここでは $\alpha = (1, 1)$ のケースである quick RLIBoS を分析する．この場合には，両プレイヤーの価値関数 $Q_{i,a}(t)$ は 1 時刻前に得た利得であるため，その値は $Q_{1,X}, Q_{2,Y} \in \{0, 1\}$, $Q_{1,Y}, Q_{2,X} \in \{0, q\}$ となる．従ってこのマルコフ過程には，両プレイヤーの価値関数のとり得る値の組み合わせ ($2^4 = 16$) から，実際には実現され得ない 2 通りを除いた 14 の状態があることがわかる．これらのうち，$(1, 0, q, 0)$ と $(0, q, 0, 1)$ の 2 つは自己遷移の確率が最大で，従ってこれら 2 つの状態での振る舞いが quick RLIBoS の振る舞いを決定づける．これらはそれぞれ戦略ペア (X, X) と (Y, Y) に対応し，本研究ではこの 2 つの状態を persistent state と呼ぶ．

以下では，persistent state において，自己遷移が連続して発生する確率を求めることにより，quick RLIBoS の振る舞いを分析する．Quick RLIBoS における状態を，両プレイヤーの戦略ペア (a_1, a_2) および価値関数の組 Q の関数として $s(a_1, a_2, Q)$ と表す．このとき，状態 $s(a_1, a_2, Q)$ が m 回連続して生じる条件付き確率は，両プレイヤーがそれぞれ戦略 a_1，a_2 を連続して m 回選ぶ同時確率として

$$P(m \mid s(a_1, a_2, Q)) = \prod_{i=1,2} \left(1 + e^{\beta_i (Q_{i,\neg a_i} - Q_{i,a_i})}\right)^{-m} \quad (3)$$

によって与えられる．ここで $\neg a$ は $A \setminus \{a\}$ に含まれる a 以外の戦略を表す．式 (3) は，persistent state の持続時間が m の指数分布に従うことを示している．ここで $r = \prod_{i=1,2} \left(1 + e^{\beta_i (Q_{i,\neg a_i} - Q_{i,a_i})}\right)^{-1}$ である．さらに，$\beta(Q_{i,a_i} - Q_{i,\neg a_i}) \ll 0$ を仮定した場合，以下の近似が成り立つ；

$$P(m \mid s(a_1, a_2, Q)) \approx c e^{-m \sum_{i=1,2} \gamma\left(\beta_i, Q_{i,a_i} - Q_{i,\neg a_i}\right)}. \quad (4)$$

ここで $c = \left(1 - e^{-\sum_{i=1,2} \gamma\left(\beta_i, Q_{i,a_i} - Q_{i,\neg a_i}\right)}\right)^{-1}$ は正規化のための定数で，また，$\gamma(\beta_i, \Delta Q) := e^{-\beta \Delta Q}$ である．式 (4) は，式 (3) で表される分布の指数自体が指数分布に従うことを意味している．これらの結果は，quick RLIBoS では，交替までの平均周期 (待ち時間) が分布の指数の逆数で与えられるようなターンテイキングが生じることを示している．

4. シミュレーションによる分析

本章では，まず，計算機シミュレーションによって前章で行った数理解析の結果を検証する．次に，RLIBoS における最も一般的なケースである general RLIBoS の均衡構造を計算機シミュレーションによって分析する．

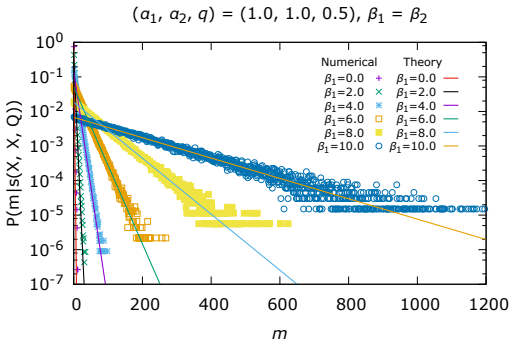

図 1: 戦略ペア (X, X) の持続時間の確率分布．$(\alpha_1, \alpha_2, q) = (1.0, 1.0, 0.5)$．$\beta_1 = \beta_2 \in \{0.0, 2.0, \cdots, 10.0\}$．実線は数理解析，点はシミュレーションによる計算結果を表す．

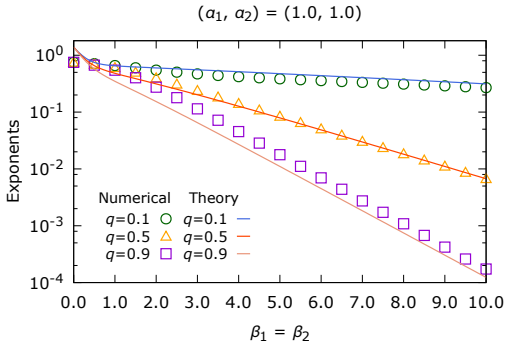

図 2: 戦略ペア (X, X) の持続時間の確率分布の指数 (傾き)．$(\alpha_1, \alpha_2) = (1.0, 1.0)$．$q \in \{0.1, 0.5, 0.9\}$．実線は数理解析，点はシミュレーションによる計算結果を表す．

4.1 Persistent state の自己遷移確率

ここでは，persistent state の持続時間の分布に対する数理解析の結果を計算機シミュレーションによって検証する．図 1 に，戦略ペア (X, X) について，与えられた (α_1, α_2, q) と $\beta_1 = \beta_2$ のもとでいくつかの β_i の値について式 (3) の確率分布を計算した結果を示す．これを見ると，戦略ペア (X, X) の持続時間の確率分布は指数分布に従い，数理解析の結果とシミュレーション結果が良く合うことがわかる．さらに，その指数 (傾き) は β_i の増加に伴って 0 に近づいていき，$\beta_i \to \infty$ においては持続時間 m の期待値は無限大になることが示唆される．$\alpha_i \neq 1$ の場合においてもシミュレーションを行い，持続時間は指数分布に従うことが確認された．

図 2 に，戦略ペア (X, X) に対し，与えられた (α_1, α_2, q) のもとで持続時間分布の指数を式 (4) およびシミュレーションによって計算した結果を示す．数理解析とシミュレーションの結果が良く合い，分布の指数が指数分布に従うことがわかる．

4.2 General RLIBoS

次に，general RLIBoS について計算機シミュレーションによって分析した結果について述べる．3 章で行った slow-and-fast RLIBoS の数理解析の結果は以下を示唆する；(1) $(Q_{1,X}, Q_{1,Y}) = (\infty, 0)$，$\beta_2 \to \infty$ が，slow-and-fast RLIBoS における唯一のナッシュ均衡である，(2) $\alpha_1 = 0$ のもとでは，α_2 の値が小さいほどプレイヤー 2 の期待利得は大きくなる．これをふまえて，ここではこの結果をより一般的なケースへと

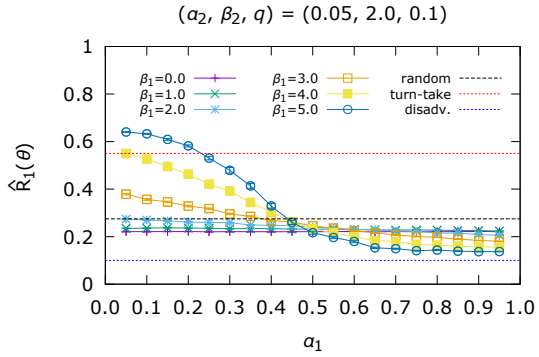

図 3: プレイヤー 1 の期待利得．$(\alpha_2, \beta_2, q) = (0.05, 0.05, 0.1)$．$\beta_1 \in \{0.0, 1, 0, \cdots, 5.0\}$．破線は，それぞれ，一様ランダムな選択 (黒)，コーディネーションの失敗のない理想的なターンテイキング (赤)，コーディネーション実現時の不利な立場 (青) において得られる期待利得を表す．

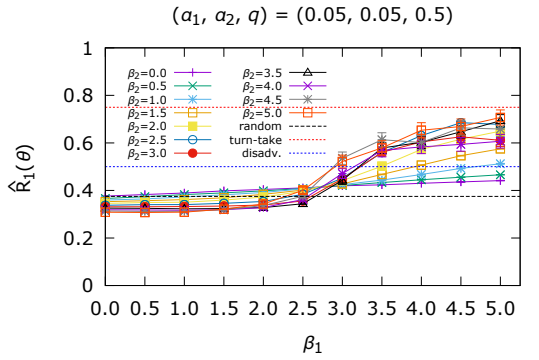

図 4: プレイヤー 1 の期待利得．$(\alpha_2, \beta_2, q) = (0.05, 0.05, 0.1)$．$\beta_1 \in \{0.0, 0, 5, \cdots, 5.0\}$．破線の意味は図 3 と同じ．

拡張することを目指し，general RLIBoS の振る舞いを計算機シミュレーションによって分析する．本研究では，シミュレーションの計算時間を $T = 1.0 \times 10^4$ とし，価値関数の初期値は $Q_0 = (0, 0, 0, 0)$ とした．さらに各パラメータに対して 50 種類の疑似乱数生成のためのシードを用いて計算を行った．

図 3 は，与えられた (α_2, β_2, q) のもとでのプレイヤー 1 の期待利得 $\hat{R}(\theta)$ を，学習率 α_1 の関数としてプロットしたグラフである．これを見ると，プレイヤー 1 (学習の速いプレイヤー) の期待利得は，ある β_1 のもとで，α_1 が 0 に近いほど大きいことが見て取れる．さらに，β_1 が大きいほど，期待利得 $\hat{R}(\theta)$ の最大値 $(\max_{\alpha_1 \in \{0.05, 0.1, \cdots, 0.95\}} \hat{R}(\theta))$ も大きくなることがわかる．図 3 で示した β_2，q の値以外でもシミュレーションを行い，同様の傾向が確認された．これらの結果は前章で得た数理解析の結果と一致し，学習率が低いほど，すなわち学習が遅いほど高い期待利得が得られることを示唆している．

図 4 は，与えられた (α_1, α_2, q) のもとでのプレイヤー 1 の期待利得 $\hat{R}(\theta)$ を，鋭敏性 β_1 の関数としてプロットしたグラフである．β_1 が比較的大きい領域では有限サンプルの影響によるゆらぎが見られるが，β の値が大きいほど期待利得が高くなる傾向が見て取れる．q の値を変えたシミュレーションの結果からも同様の傾向が確認された．この結果もまた数理解析の結果と一致しており，鋭敏性が大きく，価値関数の差異に敏感である方が高い期待利得を得られることを示唆している．

以上の結果から，RLIBoS では，学習率を 0 に近づけ，鋭敏性をできる限り大きくすることが期待利得を最大化する戦略であり，戦略ペア $(\alpha_1, \beta_1, \alpha_2, \beta_2) = (0, \infty, 0, \infty)$ がゲームの唯一のナッシュ均衡であることが示唆される．

5. 結論

本研究では，利害対立下でのコーディネーションの実現メカニズムを明らかにするために，強化学習プレイヤー同士の繰り返し BoS (RLIBoS) を導入し，その振る舞いを分析した．数理解析およびシミュレーションによる分析から，RLIBoS では，両プレイヤーともに学習が遅く，かつ鋭敏性が高い戦略の組が唯一のナッシュ均衡となることが示唆された．この結果は，強化学習プレイヤーの導入により，もとは複数均衡を有するゲームが唯一の均衡を持つゲームへと変換することで均衡選択問題を解決できる可能性を示唆している．

他方，本研究の結果は，ナッシュ均衡での振る舞いが “交替の待ち時間が極めて長いターンテイキング” であることも示唆している．従って，本研究の結果をもとに，現実に観察される利害対立下でのターンテイキングを即座に説明することは難しい．この問題を解決するためには，ゲームの繰り返し回数に対する制約を明示的に導入し，期待利得を，この時間的制約の関数として導出する必要があると考えられる．この場合，極めて長い待ち時間をもたらす戦略 (小さな学習率) では低い期待利得しか得られないため，この制約のもとでは，十分大きな学習率の組が RLIBoS の唯一のナッシュ均衡となる可能性がある．実際，生物である人間には “寿命” という根本的な時間的制約があり，現実に観察されるターンテイキングは，この種の時間制約のもとで最適な戦略をとった結果として生じている可能性がある．従って，この制約を明示的に考慮したうえでゲームの均衡構造を分析することは，ターンテイキングの発生メカニズムを明らかにするうえで重要だと考えられる．

参考文献

[Erev 98] Erev, I. and Roth, A. E.: Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria, *American economic review*, pp. 848–881 (1998)

[Helbing 05] Helbing, D., Schönhof, M., Stark, H.-U., and Hołyst, J. A.: How individuals learn to take turns: Emergence of alternating cooperation in a congestion game and the prisoner's dilemma, *Advances in Complex Systems*, Vol. 8, No. 01, pp. 87–116 (2005)

[Kahneman 03] Kahneman, D.: Maps of bounded rationality: Psychology for behavioral economics, *American economic review*, Vol. 93, No. 5, pp. 1449–1475 (2003)

[Lau 12] Lau, S.-H. P. and Mui, V.-L.: Using turn taking to achieve intertemporal cooperation and symmetry in infinitely repeated 2×2 games, *Theory and Decision*, Vol. 72, No. 2, pp. 167–188 (2012)

[Ostrom 90] Ostrom, E.: *Governing the commons: the evolution of institutions for collective action*, Cambridge, Cambridge University Press (1990)

[Sutton 98] Sutton, R. S. and Barto, A. G.: *Reinforcement learning: An introduction*, MIT press Cambridge (1998)