繰り返し局所影響関数を用いた行動選択

井越　一穂† 　三浦　孝夫†

† 法政大学工学研究科電気工学専攻
〒 184–8584 東京都小金井市梶野町 3–7–2
E-mail: †{igoshi.k.15@gs-eng, miurat@k}.hosei.ac.jp

あらまし　本研究では，マルチエージェント環境における行動選択手法を提案し，その有効性を実験で示す．行動選択基準としては，確率的ナッシュ均衡を前提とした強化学習方式を用いる．一般に，このような方式ではナッシュ均衡とパレート最適が一致せず，また繰り返し学習での収束性の保証も無いため，報酬を改善するための協調行動を得ることは容易ではない．ここでは，局所影響関数を用いた相関手法に基づいた混合ナッシュ戦略を提案し，いくつかの実験により，協調行動が獲得できることを示す．

キーワード　マルチエージェントシステム, ゲーム理論, 相関手法, ナッシュ均衡, 局所影響ゲーム

Selecting Actions on Repeated Local Effect Functions

Kazuho IGOSHI† and Takao MIURA†

† Dept. of Electrical Engineering, Graduate School of Engineering, HOSEI University,
3–7–2, KajinoCho, Koganei-City, Tokyo, 184–8584 JAPAN
E-mail: †{igoshi.k.15@gs-eng, miurat@k}.hosei.ac.jp

Abstract　In this study, we propose a framework for behavior selection under multi-agent environment, and show empirically how well our approach works. For this purpose we introduce a notion of probabilistic Nash equilibrium into reinforcement learning method. Unfortunately it is well-known that this method provides us with poor properties. For instance, Nash Equilibrium doesn't mean Pareto optimum and we can't always guarantee the convergence of learning. Here we take an approach of mixed Nash strategy based on Correlated technique in terms of Local Effect Method. We examine some experiment results to show some ideal properties for collaboration approach.

Key words　Multi-Agent System, Game theory, Correlated Technique, Nash equilibrium, Local Effect Games

1. 前　書　き

マルチエージェントシステムとは，1 つの計算空間に複数の自律的な行動主体が存在するシステムをいう．各行動主体をエージェントと呼び，行動の結果の評価値として報酬が与えられるとする．各エージェントは自律的に行動し，また他のエージェントとの敵対・協力を介してより多くの報酬の獲得を期待している．

本稿では，エージェントの行動選択に注目し，全てのエージェントの行動選択方式が個々のエージェントの行動選択方式とどのように関係するか，どのような性質を持つかを論じる．行動を選択するとき，互いのエージェントの行動により各報酬が変化するため，個別の評価を考え行動しても全体を改善することは難しい．

Hu らは，非協力状態でのマルチエージェント環境に対応させるため，ゲーム理論のナッシュ均衡を用いたナッシュQ 学習を行動選択手法に提案している [2]．ナッシュ Q 学習では，単一エージェントの学習手法である Q 学習にナッシュ均衡を利用して Q 値を更新する．しかし，競合するエージェントの振る舞いを Q 値で表現できるか疑問がある．実際，Shoham はナッシュ Q 学習を批判し，ナッシュ均衡の存在や一意性，さらに学習の収束性の保証が無いことを指摘している. [6]．そこで著者らは，ナッシュ均衡がただ 1 つ存在するよう報酬完全分配法を提案している [3].

本稿では繰り返し双方向局所影響ゲーム (R-B-LRG) を基にナッシュ Q 学習を行う，協力的なマルチエージェントシステムの枠組みを提案する．局所影響関数を用いることで利得行列を導出できることを示し，本稿で用いるナッシュ Q 学習をマルチエージェント環境に適応する．その実験結果により協力の有効性と収束性について示す.

第 2 章では，マルチエージェントシステムと行動選択を定義し，第 3 章では，相関手法における行動選択を示す．第 4 章

では，局所影響ゲームとこれに基づく行動選択，どのように局所影響ゲームにナッシュQ学習を適応するかを提案する．第 5 章では実験結果を示し，第 6 章で本稿をまとめる．

2. マルチエージェントシステムと行動選択

マルチエージェントシステムでは，同時協力・非協力ゲームでの自分の報酬の評価に，自分のみならず全エージェントの行動を評価する必要がある．これは，社会と呼ばれるような計算空間において，各エージェントの報酬がそれぞれ全エージェントにより決定するためである．そのため，個別及び全体的に改善された報酬を得るには，統一的な原理，行動選択が必要である．

本稿では，行動選択手法をゲーム理論に対応させ，複数の行動主体が選択する戦略を導入すし，毎回より多くの報酬を得るよう行動するものと仮定する．

この場合，エージェント間の関係としては協力と非協力の 2 つが存在する．互いに自分の報酬を最大化するため，交渉 (協力) すること無く戦略選択するゲームを非協力であると呼ぶ．

エージェントの行動の組が**ナッシュ均衡**であるとは，互いの行動が相手の行動に対して最適反応となるときをいう．最適反応とは，他のエージェントが行動を変えない限り，自分が行動を変える動機を持たない状態をいう．それぞれの エージェント i は，自分の報酬を最大化するような合理的な行動を選択し，エージェント i 以外の エージェント $-i$ に対して，エージェント i の報酬が最大となるような最適な行動選択をする．

例えば，エージェント A, B という 2 つが存在し，非協力 (同時) 的な行動 I, II のいずれかを選択するときを考える．行動の組み合わせにより，各エージェントは 2×2 の表 (利得行列という) に従いそれぞれ報酬を得る．表 1. の $(1, 5)$ とは，エージェント A が行動 I ，エージェント B が行動 II を取ることで エージェント A が報酬 1 ，エージェント B が報酬 5 を得ることを表す.（表 1. のような報酬のゲームのことをチキンゲームと呼ぶ）

		エージェント B	
		I	II
エージェント A	I	$(4, 4)$	$(1, 5)$
	II	$(5, 1)$	$(0, 0)$

表 1　チキンゲーム

表 1. で，エージェント B が行動 I を選択した場合，エージェント A は，行動 II により大きな報酬 5 を得る．一方，エージェント B が行動 II を選択した場合には，エージェント A は，行動 I により大きな報酬 1 を得る．同様に，エージェント B の選択すべき行動も一致することから，ナッシュ均衡は (I, II) , (II, I) の 2 つとなる．このような手法を，**純粋戦略**と呼ぶ．しかし，ナッシュ均衡が必ず存在する訳ではなく，また 1 つよりも多くの均衡が存在することもある．

もう 1 つの方法として，**混合戦略**という，確率的に行動選択することを考える．混合戦略の場合には，ナッシュ均衡は必ず存在することが知られている．混合戦略で，表 1. を考えるとき，エージェント A が戦略 I を選択する確率を p ，エージェント B が戦略 I を選択する確率を q とすると，エージェント A の報酬の期待値 $f(p, q)$ 及び，エージェント B の報酬の期待値 $g(p, q)$ は次の式で表される．

$$
\begin{aligned}
f(p,q) &= 4pq + 1p(1-q) + 5(1-p)q + 0(1-p)(1-q) \\
&= p(1-2q) + 5q \\
g(p,q) &= 4pq + 5p(1-q) + 1(1-p)q + 0(1-p)(1-q) \\
&= q(1-2p) + 5p
\end{aligned}
$$

$f(p, q)$ を p に関する関数とみれば次のように書き換えることができる．

$$
\begin{cases}
1-2q > 0 \text{ のとき}, p = 1 \text{ で最大値 } 1+3q \\
1-2q < 0 \text{ のとき}, p = 0 \text{ で最大値 } 5q \\
1-2q = 0 \text{ のとき，最大値は} \dfrac{5}{2} \ (p \text{ は任意})
\end{cases}
$$

同様に エージェント B についても計算できる．q に関する関数として，$g(p, q)$ の最大値を考える．

$$
\begin{cases}
1-2p > 0 \text{ のとき}, q = 1 \text{ で最大値 } 1+3p \\
1-2p < 0 \text{ のとき}, q = 0 \text{ で最大値 } 5p \\
1-2p = 0 \text{ のとき，最大値は} \dfrac{5}{2} \ (q \text{ は任意})
\end{cases}
$$

各エージェントは，より多くの報酬を獲得するために報酬の期待値が最大となるように行動を選択する．全てのエージェントは，利得行列や確率を知っているものとし，図 1. の交点がナッシュ均衡となる．

この例では，$(I, II), (II, I), \left(\{\frac{1}{2}I + \frac{1}{2}II\}, \{\frac{1}{2}I + \frac{1}{2}II\}\right)$ の 3 つがナッシュ均衡となる．このようなナッシュ均衡を，混合戦略ナッシュ均衡と呼ぶ．

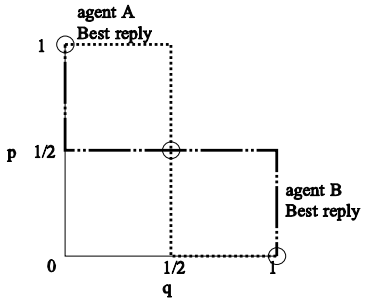

図 1　各エージェントの最適反応

ここまでのナッシュ均衡の議論では，エージェント間に通信路が無く，各々のエージェントがどの均衡（行動）を選ぶべきの選択手段が無い．また，ナッシュ均衡は総報酬がより改善されるパレート最適を意味せず，最多報酬を得られるとは限らない．そのため，ナッシュ均衡はマルチエージェントの行動選択を行う唯一の基準と考えることは難しい．

3. 相 関 手 法

本研究では，Aumann らにより提案された**相関手法**を導入する [1]．相関手法とは，ナッシュ均衡と同様に合理的行動選択を前提とし，エージェント間は通信路を仮定しそれらの相互作用を示す「共通の情報」を許容する．合理的な行動選択は，式 1. の様に定義する．

$$\sum_{s\in S_{-i}} u^i_{as} - u^i_{a's} \geqq 0 \tag{1}$$

エージェント数 n に対して，それぞれの エージェント i は行動の集合 S_i を持ち，S_i を除いた行動の集合を S_{-i} と表す．報酬 u^i_{as} は エージェント i の行動 $a_i \in S_i$ と エージェント i 以外の行動の組み合わせ $s = (s_1, ..., s_{i-1}, s_{i+1}, ..., s_n)$ で決定されるとする．

合理的な選択 (式 1.) では，エージェント i の行動 a は他の行動 a' を選択するよりも報酬が高く，各エージェントは自ら均衡を逸脱する動機を持たないことを意味する．

例えば，表 1. において，エージェント B が行動 I を選択する場合，エージェント A は，より大きな報酬が得られる行動 II を選択する（この場合は，報酬 5 が得る）．

一方，エージェント B が行動 II を選択する場合には，エージェント A は，行動 I を選択するほうが大きな報酬が得られる．従って，(I, II) と (II, I) の 2 つのナッシュ均衡が存在する．

ナッシュ均衡との違いは，共通の情報を介した協調する機構にある．このような情報を共有する機構を持つゲームを協力ゲームと呼ぶ．多くの種類のエージェント間の協力ゲームがあるが，ここでは，マルチエージェントシステムにおける協調を目的とし，共通（相関）情報に基づき均衡の導出について考える．

4. 局所影響ゲームモデル

4.1 混雑ゲーム

マルチエージェントシステムにおける相関手法の問題として，Rosenthal によって提案された，**混雑ゲーム** (Congestion Games) が知られている [5]．混雑ゲームとは，場が混雑するにつれてより負の報酬であるコストが増大するという，ネットワークや施設の混雑などをモデル化したものである．ここで，混雑ゲームを定式化する．混雑ゲーム G は，エージェント数 n ，コスト (負の報酬) r ，要因 k からなる．エージェント $i(i = 1...n)$ の報酬 $r_i(a_1, ..., a_n)$ は式 2. のように表される．

$$r_i(a_1, ..., a_n) = \sum_{k\in a_i} F_k(D_k(a_1, ..., a_n)) \tag{2}$$

各 $a_i(a_i = \{1, ..., s_i\})$ は選択可能な エージェント i の行動を表し，D_k は要因 k を選択したエージェント数，F_k は要因 k による エージェント i のコストを表す関数である．報酬 $r_i(a_1, ..., a_n)$ は，混雑ゲームは全 n エージェントの行動により決定され，混雑に応じて負の報酬が増えることを意味している．

しかし，全エージェントとの相互作用を算出する必要があるため，エージェント間の相互作用は複雑になる．エージェント数に応じてパラメタ数も増え，膨大な計算が必要となる．

4.2 局所影響ゲーム

本稿では，混雑ゲームに対して局所的（直接）に影響を与えるエージェントの相互作用のみを考える，Leyton-Brown らによって提案された**局所影響ゲーム** (Local Effect Games: LEG) をモデル化の基準とする [4]．局所影響ゲームは，全エージェントとの相互作用ではなく，局所的に影響するエージェントとの相互作用のみを考えるため，計算を減少させることができる．

ここで，エージェント数 n $(n > 1)$ の局所影響ゲーム G を定義する．G は，$\mathcal{A}$，F，n からなり，$\mathcal{A}$ は各エージェントが選択可能な行動集合を表す．局所影響ゲームでは，各エージェントが選択可能な行動が共通である．$D(x)$ は，行動 x を選択しているエージェント数を表す．$F_{a',a}(D(x))$ は，エージェント i が行動 a を選択しているとき，行動 x を取るエージェント $D(x)$ エージェントが与えるコストを表す関数である．この関数 F はを **局所影響関数** (Local Effect Function) と呼び，エージェント間の相互作用（共通情報）を示す．$F_{a',a}(0) = 0$ とする．

エージェント i が行動 $a \in \mathcal{A}$ を選択するとき，エージェント i は，a と エージェント i 以外の行動 $a' =< b_1, ..., b_{n-1} >\in \mathcal{A}^{n-1}$ と $\mathcal{B} = \{a, b_1, ..., b_{n-1}\}$ により決定される報酬 $r_i(a, a')$ を獲得する．この状況を以下のように表す．

$$r_i(a', a) = \sum_{y\in\mathcal{B}^{n-1}, x\in y} F_{y,a}(D(x)) \tag{3}$$

局所影響関数 F を与えることで，利得行列 M を定義できる．エージェント i の行動 a と エージェント i 以外の行動 a' の組により，報酬 $r_i(a', a)$ が決まる．行動 $< a_1, .., a_n >\in \mathcal{A}^n$ の行列の各項目 $(r_1, .., r_n)$ に対し，局所影響関数 F から報酬 $r_i(a', a)$ により r_i を見積もる．

$F_{I,I}(x)$	$-100x$
$F_{II,I}(x)$	$-50x$
$F_{I,II}(x)$	$-60x$
$F_{II,II}(x)$	$-120x$

表 2　局所影響関数

例えば，エージェント数が 2 であるときの局所影響関数を表 2. に示す．エージェント A が行動 I ，エージェント B が行動 I を取ったときの エージェント A の報酬は，

$$F_{I,I}(2) + F_{II,I}(0) = -200$$

エージェント A が行動 I ，エージェント B が行動 II を取ったのときの エージェント A の報酬は，

$$F_{I,I}(1) + F_{II,I}(1) = -150$$

となる．結果，得られる報酬の組み合わせによる利得行列を表 3. で示す．

	エージェント B		
エージェント A		I	II
	I	$(-200,-200)$	$(-150,-180)$
	II	$(-180,-150)$	$(-240,-240)$

表 3　局所影響ゲームの利得行列

4.3　双方向局所影響ゲーム

本実験では単純化のため，**双方向局所影響ゲーム** (Bidirectional Local Effect Games: B-LEG) を用いる [4]．B-LEG では，局所影響ゲームに制限を加え"相手に与える影響"と"自分が受け取る影響"が同じであり，エージェント数に関してその効果は線形関数であると仮定する．以下，本稿では B-LEG について扱う．B-LEG では純粋戦略ナッシュ均衡が 1 つ以上存在することが示されている [4] 一般の局所影響ゲームの場合は，純粋戦略ナッシュ均衡が存在しない場合がある．エージェント i が行動 $a \in \mathcal{A}$ を，他のエージェントが行動 $a' =< b_1, ..., b_{n-1} >\in \mathcal{A}^{n-1}$ を取ったとき，B-LEG における局所影響関数を以下の式で定義する．

$$F_{a',a} = F_{a''} \quad a'' \in \mathcal{B}^n, \text{ i.e., } a'' \text{ は } \mathcal{B} \text{ を置換したもの} \quad (4)$$

$F_{I,I}(x)$	$-100x$
$F_{II,I}(x) = F_{I,II}(x)$	$-50x$
$F_{II,II}(x)$	$-120x$

表 4　局所影響関数 (B-LEG)

例えば，表 4. で 2 エージェントの局所影響関数が与えられたとき，エージェント A が行動 I ，エージェント B が行動 I を取るとき，つまりエージェントの行動の組が (I, I) のとき，エージェント A の報酬は，-200 となる．

$$F_{I,I}(2) + F_{II,I}(0) = (-100) \times 2 + (-50) \times 0 = -200$$

同様に，エージェント A が行動 I ，エージェント B が行動 II を取るとき，つまり，一方のエージェントの行動の組が (I, II) で他方のエージェントの行動の組が (II, I) であるとき，エージェント A の報酬は，-150 となる．

$$F_{I,I}(1) + F_{II,I}(1) = (-100) \times 1 + (-50) \times 1 = -150$$

その結果，報酬を利得行列にまとめたものを表 5. に示す．

	エージェント B		
エージェント A		I	II
	I	$(-200,-200)$	$(-150,-170)$
	II	$(-170,-150)$	$(-240,-240)$

表 5　B-LEG の利得行列

本稿で考える協力ゲームでは，我々は共通情報として局所影響関数を共有し，共通情報を踏まえた利得行列に従い，混合戦略ナッシュ均衡により行動を選択すると仮定する．非協力であるときに比べ，局所影響関数により共通情報を持つためよい行動の選択に作用する．そのため，本手法はマルチエージェントシステムで自律的に協調行動の獲得が期待できる．例えば，表 5. では，報酬の期待値は $(-187.5, -187.5)$ となり非協力ゲームの場合よりも，協調行動を獲得しやすい．

5.　繰り返し双方向局所影響ゲーム

B-LEG など今まで考えてきたゲームでは，各エージェントが取る行動の報酬が既知として与えられた**1回限りのゲーム**を前提としている．しかし実際のマルチエージェントシステムでは，報酬が事前に判明していることは少ないであろう．多くの場合，「状況を観測し，行動することで，評価値として報酬を得る」という過程を繰り返し，そこから選択した行動と報酬の関係のルールを獲得しようとする．このような手法を**強化学習**という．本稿では，**繰り返しゲーム**にて強化学習の枠組みを用いる．

単一エージェント の学習手法として， Q **学習**が知られている．この手法では，状態空間と行動集合が与えられ，状態遷移関数 f を持つ．$f(s, a)$ は 状態 s で行動 a を取ったときの次の状態 s' を意味する．そのとき，$Q(s, a)$ で表す Q 値と呼ばれる将来に得られる期待報酬を見積もる．Q 学習では式 5 ．に従い Q 値を更新する．

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s', a') - Q(s,a)) \quad (5)$$

状態 s で行動 a を取るとき， 行動 a により得る報酬 r と次の状態 s' の Q 値により $Q(s, a)$ を更新する．

ここで，価の更新を調整するために α と γ の 2 つのパラメタを用いる．$\alpha (0 \leqq \alpha < 1)$ を **学習率** と呼び，行動 a により得られる報酬 r の優先度を表す．また，$\gamma (0 \leqq \gamma \leqq 1)$ を**割引率** と呼び，期待値 (次状態の Q 値: $Q(s', a')$) の優先度を表す．Q 学習は学習を無限回繰り返すことで収束することが知られている [7].

しかし，R-B-LEG では複数のエージェントが存在するため，行動を決定する指針が無く， Q 学習をそのまま適用することはできない．**ナッシュ** Q **学習**は，Q 値の更新と行動選択にナッシュ均衡を用い，マルチエージェント環境に対応させた Q 学習である. [2]．各エージェントの行動 a_i $(i = 1, ..., n)$ と状態 s が与えられたとき，状態遷移関数 f は $s' = f(s, a_1, ..., a_n)$ で定義され，式 6 により Q 値 $Q_i(s, a_1, ..., a_n), i = 1, ..., n$ を更新する．

$$\begin{aligned} Q_i(s, a_1, ..., a_n) \leftarrow\ & Q_i(s, a_1, ..., a_n) \\ & +\alpha\{r_i + \gamma NashQ_i(s') - Q_i(s, a_1, ..., a_n)\} \\ NashQ_i(s') =\ & \pi_1(s')...\pi_n(s')Q_i(s') \end{aligned} \quad (6)$$

π_i は，エージェント i の取る行動の確率を表し，混合戦略ナッシュ均衡により確率的に決定される行動を表す．ナッシュ Q 学習では，Q 値の更新を繰り返すことにより，期待報酬値に近づくことが期待され，各エージェントは学習した Q 値を基に行動を選択する．

本稿では，繰り返し双方向局所影響ゲーム (Repeated Local Effect Games: R-B-LEG) に基づく，ナッシュ Q 学習を考える．ここでは，繰り返し手法を B-LEG に適用する．すなわち，

局所影響関数に基づく利得行列を用い，この行列を Q 値計算の報酬定義と考える．ここで，Q 値は Q **行列**という行列の形で表すことができる．例えば，$n=2$ の場合，ある状態で 2×2 の行列を考える．行列の (I,II) における項目 (α,β) は，状態 s でにおける $\alpha=Q_1(s,I,II)$ と $\beta=Q_2(s,I,II)$ である．

		エージェント B	
		I	II
エージェント A	I	$(-200,-200)$	$(-150,-170)$
	II	$(-170,-150)$	$(-240,-240)$

表 6　Q 値 行 列

学習の結果, 表 6. の Q 行列にある Q 値が得られたとする．純粋戦略ナッシュ均衡を考えた場合には，ナッシュ均衡は (I,II) と (II,I) の 2 つが存在する．混合戦略でのナッシュ均衡を加えると，(I,II) と (II,I)，$(\{0.75I+0.25II\},\{0.75I+0.25II\})$ の 3 つである．混合戦略ナッシュ均衡が複数存在する場合は，期待値が最大の行動を選択する．もし純粋戦略ナッシュ均衡のみが存在する場合には，純粋戦略ナッシュ均衡を選択する(注1)．そのため，ここでは (I,II) , (II,I) , $(\{0.75I+0.25II\},\{0.75I+0.25II\})$ の均衡のうち，$(\{0.75I+0.25II\},\{0.75I+0.25II\})$ を選択する．

ナッシュ Q 学習では，特殊なケースを除いて収束は保証されない [6]．本稿では，B-LEG を基にしていることから，一般的なナッシュ Q 学習よりも学習の収束が期待できる．

6. 実　　験

6.1 準　　備

本実験では，R-B-LEG によるナッシュ Q 値の収束，及び協調行動の獲得について確認する．ここでは，比較対象として一様分布する利得行列に基づく**ランダムゲーム** *(RandG)* を用いる．

本実験では，エージェント数を 2 , 状態数を 1 とし，選択可能な行動数を $\mathcal{A}=\{I,II,III\}$ の 3 通り とする．

ゲームは，2 エージェントが同時に行動を選択する．エージェントは利得行列から行動を選択し，選択に従い報酬を得る．それぞれのエージェントがゲームを 1 回プレイすることを 1 ゲーム , Q 値の学習のために, 5000 ゲーム 繰り返すことを 1 エピソード とする．各パラメータに対し，1 エピソード を 10 回 繰り返し，このことを，1 サンプル という．本実験では 50 サンプル 用意する．従って，合計 50[サンプル]$\times$10[エピソード]$\times$5,000[ゲーム] $=$ 2,500,000 回ゲームを行う．

R-B-LEG では局所影響関数 F を，RandG では利得行列のパラメタを，共に 50 サンプル用意する．

本実験の R-B-LEG では，0 ～ 10 の間で一様分布に従い 50 セット の局所影響関数を生成する．RandG の利得行列は，一様分布に従い，0 ～ 10 の間で 50 セット の利得行列を生成し，実験に用いる．

(注1)：混合戦略ナッシュ均衡が存在しない場合は，純粋戦略ナッシュ均衡が唯一の支配解となっている場合であり他の行動を選択する余地は無い．

ナッシュ Q 学習においては．2 エージェントは互いに同一の行動決定手法に基づく．

評価のため，各ゲームの報酬結果を用い，学習が十分収束している時点での報酬を比較する．本実験では，サンプル毎において最大繰り返し 4501 − 5000 回で獲得した報酬の平均と，繰り返しが 5000 回 未満との差を比較する．報酬が早くに収束している場合には，2 つの差は小さくなることが期待できる．

本稿では，協調行動が機能していることを，報酬の偏りが少なくどのエージェントも同程度の報酬を獲得すること，勝敗数に大差が無いまたは引き分けが多いこと，とする．ここでは，一方のエージェントよりも報酬が多いエージェントを”勝ち”，互いのエージェントが共に同程度の報酬の場合を”引き分け”と定義する．本実験において報酬を無作為に生成し報酬自体の大小は意味を持たないため，勝ち，負け，引き分けの数を調べる．

6.2 実 験 結 果

繰り返し回数	951 − 1000 回		4501 − 5000 回	
	平均	分散	平均	分散
R-B-LEG	15.94	0.05	15.95	0.00
RandG	5.04	0.02	5.05	0.00

表 7　報酬の平均と分散

R-B-LEG 及び RandG について，繰り返し回数 951−1000 回と 4951 − 5000 回 の報酬の平均と分散を表 7. に示す．表 7. では，1000 回 まで 50 回 の平均と分散と 5000 回 まで 500 回 の平均と分散の，50 サンプル平均である．

R-B-LEG，RandG とも，繰り返し増えることで分散が小さくなっている．

実際，R-B-LEG と RandG の各回数の報酬と比較するために，式 7. を用いる．

$$\text{差 }(reward)[\%]=\frac{reward-Reward}{Reward}\times100 \qquad (7)$$

ここで，繰り返し回数が 5000 回 (4501 − 5000 平均報酬) 時点の報酬を Q 値が十分に収束していると仮定する．$reward$ は各回の報酬，$Reward$ は 5000 回 (4501 − 5000 平均) の報酬として，安定度の差を考える．何回差が 5% 以上をカウントした結果を，表 8. に示す．50 サンプル 中で R-B-LEG では，500 回 以降 5% 以上の差が無い．しかし RandG では，不安定な状況のものが残っている．

5000 回 までの差の比較結果を表 9. に示す．表 7. の 5000 回 までの結果を用い，式 7. の結果を ±4% で 0.5% ずつ 18 分割した結果が，表 9. である．

両者の分布が異なることを確かめるため，RandG を理論値，R-B-LEG を標本値として χ 2 乗検定を行う．有意水準を 5% とし適合度を比較すると，3 カ所（ 500 回, 1000 回, 4000 回 ）を除き適合性は棄却され独立な（別個）の分布であると判断してよい．3 カ所のうち，初めの 2 カ所は学習の途中あるため,. 4000 回 では RandG が 5000 回 に近い値をとっているため，適合度が高くなっている．しかし，前後の 3500 回，4500 回 での適合度が低いため，偶然に生じたものであろう．

繰り返し回数	50	100	150	200	250	300	350	400	450	500
R-B-LEG	5	0	0	0	0	1	0	0	0	1
RandG	2	3	1	4	2	1	2	1	1	0
繰り返し回数	550	600	650	700	750	800	850	900	950	1000
R-B-LEG	0	0	0	0	0	0	0	0	0	0
RandG	2	1	1	2	0	2	1	2	2	1

表 8　5000 回 (4501-5000 平均) との比較

繰り返し回数		500	1000	1500	2000	2500	3000	3500	4000	4500
理論値	4.0%	0	0	0	0	0	0	0	0	0
RandG	3.5%	0	0	0	1	0	0	0	0	0
	3.0%	1	0	0	1	1	1	0	2	0
	2.5%	0	2	0	1	1	0	2	2	0
	2.0%	0	0	0	3	3	3	1	0	1
	1.5%	1	2	2	1	1	2	2	2	1
	1.0%	5	3	2	6	2	6	2	1	5
	0.5%	6	7	8	7	7	3	9	5	11
	0.0%	8	11	10	7	9	5	12	11	5
	-0.5%	13	10	6	11	9	13	11	10	9
	-1.0%	7	7	8	6	7	7	3	6	7
	-1.5%	5	6	8	4	6	5	5	6	4
	-2.0%	3	0	5	1	3	2	2	2	3
	-2.5%	0	1	0	1	0	2	0	1	3
	-3.0%	1	0	0	0	1	0	1	1	0
	-3.5%	0	1	1	0	0	0	0	1	1
	-4.0%	0	0	0	0	0	1	0	0	0
	-4.0%未満	0	0	0	0	0	0	0	0	0
標本値	4.0%	0	0	0	0	0	0	0	0	0
R-B-LEG	3.5%	0	0	0	0	0	0	0	0	0
	3.0%	0	0	0	0	0	0	0	0	0
	2.5%	0	0	0	0	0	0	0	0	0
	2.0%	0	0	0	1	0	0	0	0	0
	1.5%	0	0	1	0	0	0	0	1	0
	1.0%	3	2	1	0	2	1	2	1	1
	0.5%	4	6	5	5	5	6	7	10	10
	0.0%	14	21	17	20	20	12	12	15	14
	-0.5%	18	15	18	20	19	24	22	17	15
	-1.0%	5	4	6	2	3	4	7	5	4
	-1.5%	6	1	1	1	1	2	0	1	5
	-2.0%	0	0	1	1	0	1	0	0	1
	-2.5%	0	1	0	0	0	0	0	0	0
	-3.0%	0	0	0	0	0	0	0	0	0
	-3.5%	0	0	0	0	0	0	0	0	0
	-4.0%	0	0	0	0	0	0	0	0	0
	-4.0%未満	0	0	0	0	0	0	0	0	0
χ 2乗検定 適合度		61.99%	16.56%	0.07%	0.01%	0.08%	0.19%	2.80%	9.06%	1.36%

表 9　5000 回 (4501-5000 平均) との比較した場合の χ 2乗検定

繰り返し回数	500	1000	1500	2000	2500	3000	3500	4000	4500	5000	平均
勝	52953	52348	52155	52548	52372	52175	52622	52737	52271	52238	
	21.18%	20.94%	20.86%	21.02%	20.95%	20.87%	21.05%	21.09%	20.91%	20.90%	20.98%
引き分け	144314	145234	145111	145062	145106	145533	144920	144987	145416	145555	
	57.73%	58.09%	58.04%	58.02%	58.04%	58.21%	57.97%	57.99%	58.17%	58.22%	58.05%
負	52733	52418	52734	52390	52522	52292	52458	52276	52313	52207	
	21.09%	20.97%	21.09%	20.96%	21.01%	20.92%	20.98%	20.91%	20.93%	20.88%	20.97%

表 10　エージェント A の勝ち負け数 (R-B-LEG)

繰り返し回数	500	1000	1500	2000	2500	3000	3500	4000	4500	5000	平均
勝	118072	117900	117962	118568	117815	118157	118165	118224	117615	118115	
	47.23%	47.16%	47.18%	47.43%	47.13%	47.26%	47.27%	47.29%	47.05%	47.25%	47.22%
引き分け	13184	13112	13162	13253	13132	13145	13147	12993	13282	13108	
	5.27%	5.24%	5.26%	5.30%	5.25%	5.26%	5.26%	5.20%	5.31%	5.24%	5.26%
負	118744	118988	118876	118179	119053	118698	118688	118783	119103	118777	
	47.50%	47.60%	47.55%	47.27%	47.62%	47.48%	47.48%	47.51%	47.64%	47.51%	47.52%

表 11　エージェント A の勝ち負け数 (RandG)

6.3 考　　察

各ゲームにおけるの勝ち負け数について考える．前述の通り，協調行動の確認のために，勝ち，負け，引き分けの数を用いる．ここでは，エージェントの報酬が他方の 5% 以上であるとき”勝ち”とし，報酬の差が 5% 以内のときを”引き分け”とする．

R-B-LEG 及び RandG における エージェント A の勝ち負け数を，表 10.(R-B-LEG) 及び表 11.(RandG) に示す．R-B-LEG と RandG において，勝ち・負けの数は両者ともほぼ同じである．しかし，R-B-LEG が引き分けが非常に多く（ 58.05% ）を占め，RandG は非常に少なく（ 5.26% ）なっている．従って，R-B-LEG は協調行動を獲得している．

	$951-1{,}000$ 回 平均	$4{,}951-5{,}000$ 回 平均
50 回より大	24 (48.0%)	25 (50.0%)
2.0%上昇	16 (32.0%)	17 (34.0%)
5.0%上昇	4 (8.0%)	4 (8.0%)

表 12　繰り返し 1 ～ 50 回との比較

ナッシュQ 学習の効果を確認するため，報酬の上昇傾向を表 12. に示す．この表は繰り返し回数が 1−50 回 の報酬の平均と比較し，どれだけ多くのゲーム（951 − 1,000 回，4,951 − 5,000 回の平均）で報酬が上昇しているかを示している．表より，例えば 951 − 1,000 回 の平均では，24 ゲーム では 1 − 50 回 よりも報酬が上昇している（48%）．24 ゲーム の中でも 2% 以上報酬が上昇したものは，16 ゲーム である．

一方，報酬が 5% 以上上昇したものは，8% しかない．これは エージェント B が得る報酬が エージェント A の得る報酬より少ないため，エージェント間の報酬差が減少するよう作用したことに起因する.. これは，協調行動の 1 つの欠点であると推測できる，

収束について考えると，表 9. では，R-B-LEG の繰り返し回数が 2500 回 以上で差が ±2% に収束しており，χ 2 乗検定により，RandG に比べ収束する傾向が強い．

7. 結　　論

本稿では，繰り返し双方向局所影響ゲーム（R-B-LEG）に基づくナッシュ Q 学習を用い，協調的なマルチエージェントシステムの枠組みを提案した．本手法では，ナッシュ均衡が必ず存在し，混合戦略により期待値最大の均衡を 1 つ選ぶことができる点で興味深い．

ここでは局所影響関数を導入したため，利得行列に対応できる．また，マルチエージェント環境でのナッシュ Q 学習を適用した．一般的にナッシュ Q 学習は多くの問題があるが，本手法はそれらの問題を克服している．協調の有効性と実験における収束の結果から，本手法が有効であることを示した．

文　　献

[1] R.J. Aumann: Subjectivity and correlation in randomized strategies, Journal of Mathematical Economics 1:pp 67-96, (1974).

[2] J. Hu and M.P. Wellman: Nash Q-Learning for General-Sum Stochastic Games, Journal of Machine Learning Research 4, pp.1039-1069, (2003).

[3] K. Igoshi and T. Miura: Strategic Knowledge By Nash-Q Learning for Reward Distribution, First IEEE International Conference on the Applications of Digital Information and Web Technologies (ICADIWT), (2008).

[4] K. Leyton-brown and M. Tennenholtz: Local-Effect Games, International Joint Conference on Artificial Intelligence (IJCAI), (2003).

[5] R.W. Rosenthal: A class of games possessing pure-strategy Nash equilibria. International Journal of Game Theory, Volume 2, Number 1, pp.65-67, (1973).

[6] Y. Shoham, R. Powers and T. Grenager: Multi-Agent Reinforcement Learning - A Critical Survey, Technical Report, (2003).

[7] C.J.C.H. Watkins and P. Dayan: Technical note: Q-Learning, Machine Learning, Volume 8, pp.279-292, (1992).