マルチエージェント環境におけるエージェントの行動選択に関する研究

井越, 一穂 / IGOSHI, Kazuho

(発行年 / Year)
2010-03-24
(学位授与年月日 / Date of Granted)
2010-03-24
(学位名 / Degree Name)
修士(工学)
(学位授与機関 / Degree Grantor)
法政大学 (Hosei University)

2009年度　修 士 論 文

マルチエージェント環境における
エージェントの行動選択に関する研究

STUDIES ON ACTION SELECTION OF AGENTS ON MULTI-AGENT ENVIRONMENT

指導教授　三浦 孝夫 教授

法政大学大学院工学研究科
電気工学専攻修士課程

08R3105　井越（イゴシ）　一穂（カズホ）
Kazuho IGOSHI

目 次

第1章　序論

1.1　研究の背景

人間や自動車は個々が自律的に行動している．また，近年，パーソナルコンピュータや携帯電話に代表されるように計算機の小型化や高性能化が進み，かつ安価に入手できるようになったことで，多くの人々に爆発的に普及してきている．その結果，高性能な情報機器が様々なものに搭載され，電車に搭載された自動列車運転装置（Automatic Train Operation : ATO）など自律的に行動する機器も増えてきている．そのため，自律的な行動主体（エージェント）がますます増加している．

例えば，自動車と渋滞の関係を考える．日本における渋滞による経済的損失は年間約 12 兆円 [20] といわれており，社会の一般的な問題の 1 つである．もし自動車が 1 台のみ走っている場合には，目的地まで最短経路を走ればよい．しかし，多数の自動車が走り渋滞が発生している場合はどうだろうか．多くの自動車は，早く目的地に着くために渋滞を避けようと考えるだろう．その中で，自身の自動車はどのような経路（行動）を選択すべきかという問題がある．

上記のように，複数の行動主体（エージェント）が相互に影響を与え合っている構造を，**マルチエージェントシステム**（Multi-Agent System : MAS）という．マルチエージェントシステムでは，個々のエージェントが協調して行動することで，全体の評価を向上させることを目的としている．従って，エージェントが増加している近年，マルチエージェントシステムの解決が必要不可欠である．

1.2　行動選択手法

マルチエージェントシステムで最も重要な点が，各エージェントが協調行動を選択することである．単一エージェントの場合は，先ほどの例の通り，単純に自身の評価を最大化するよう行動を選択すればよいだろう．しかし，マルチエージェント環境では他のエージェントの行動により環境が変化する．

例えば，図 1.1. のように，2 台の車 A・B が狭路を通過しようとしている．しかし，狭いため 1 台ずつしか道を通過できない．2 台の車の選択肢は,「進む」,「止まる」の 2 択とした場合，どちらの行動を選択すべきだろうか．2 台とも「止まる」を選択した場合は状況が変化せず，2 台とも「進む」を選択した場合は衝突してしまう.「止まる」

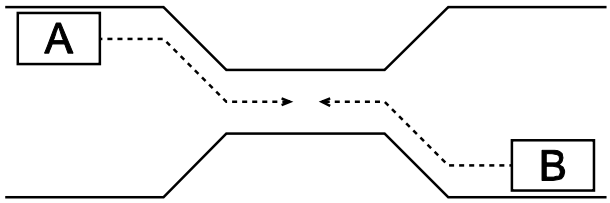

図 1.1: 狭路通過問題

「進む」を五分五分で選んだとしても，$\frac{1}{2}$ で「状況が変化しない」か「衝突」となる．従って，行動選択はマルチエージェントシステムにおいて重要な点である．

1.3 エージェント間の関係，１回限り・繰り返し

行動選択にて問題となるのが，エージェント間の関係である．エージェント間に通信路が存在する場合と，しない場合では協調行動のとりやすさに大きな差があり，その振る舞いが大きく異なると考えられるためである．

図 1.1. の例では，通信路が無い場合は，各エージェントが持っている情報のみで，「進む」「止まる」を選択する必要がある．通信路がある場合は，狭路に信号がある場合である．A・B が信号に従う限り衝突が起きることは無く，狭路を通過することができる．このように，エージェント間の関係で状況が大きく異なる．

また，行動選択が１回限りの試行か，繰り返し試行するかでも大きな違いがあるだろう．前者の場合は，現状でわかっている情報のみで行動選択をする必要がある．しかし，後者の場合は繰り返し試行をすることで，以前の行動選択の結果をフィードバックし，次の行動に反映させることができる．

図 1.1. の例では，狭路を１回限り通るか，何回も通るかである．何回も通る場合は，A が「進む」が多いことフィードバックすることで B は「止まる」を選択することができるようになる．そのため，１回限りの試行か，繰り返し試行するかでは大きな違いがある．

従って，マルチエージェントシステムにおける行動選択では,「エージェント間の通信路」と「繰り返し」に関する 2 点について考えなければならない．

1.4 扱う問題

本研究では,

- エージェント間に通信路の有無
- 行動選択が１回限りの試行か繰り返し試行するのか

の 2 点が，マルチエージェントシステムにおける行動選択に対し大きな影響を与えると考え中心に議論する．

まずはじめに，非協力ゲームにおける協調行動の選択について議論し，協調手法を提案する．次に，協力ゲームにおける 1 回限りの試行および繰り返しの試行の協調行動の獲得について論じる．

1.4.1 非協力ゲームにおける行動選択

単一エージェントの行動選択の手法としては環境に影響されず教師データを必要としない強化学習がよいとされ，学習の収束性が保証されている Q 学習がよく知られている．マルチエージェント環境に Q 学習をそのまま適用することは問題を抱えている．

そこで，マルチエージェント環境に対応させるため，Q 学習にゲーム理論の 1 つナッシュ均衡の導入したナッシュ Q 学習が提案されている [3]．本研究では，複雑なマルチエージェント環境にてナッシュ Q 学習を適用し，エージェントが協調行動を獲得できるかを確認する [4]．

1.4.2 非協力ゲームにおける協調行動

上で述べた，ナッシュ Q 学習では，ナッシュ均衡が一意解とは限らず，学習の収束性が保証されないという問題が指摘されている [17]．

本研究では，エージェントに与える報酬に制約を加えることでナッシュ均衡が一意に存在するよう報酬分配手法を提案し，本手法にて協調行動が獲得できることを示す [5]．

1.4.3 協力ゲームにおける行動選択

今までは非協力ゲームにおける行動選択手法について論じている．しかし，エージェントが共通の通信路を持っている場合を仮定した場合は，どのような振る舞いをするのだろうか．

ここでは，エージェント間に通信路を仮定し相互作用を表す**共通の情報**を許容する，**相関手法**を用いる．例えば「信号機」を共通の情報と考える場合，2 台の自動車が交差点にさしかかったとき，一方の信号は赤で停止をし，他方の信号は青で進むという行動選択をする場面が想定される．

本研究では，相関手法に対応した**局所影響ゲーム** を用いる．局所影響ゲームは局所影響関数を共通の情報として持つ．また，確率的に行動決定することを許容する混合戦略ナッシュ均衡を用いる．これらにより，相関手法が行動選択手法として有効であることを実験により示す [6]．

1.4.4 繰り返しゲームにおける行動選択

先ほど述べた手法では，1回限りのゲームにおける行動選択について論じている．これを繰り返しゲームに拡張する場合はどのような行動選択をするべきであろうか．

繰り返しゲームになると，以前の行動を経験として次の行動選択にフィードバックすることが一般的だろう．そこで，繰り返しゲームでは強化学習の枠組みで考える必要がある．マルチエージェント環境に対応した学習手法としては，ナッシュ Q 学習を用いることが考えられる．

本研究では，繰り返し非協力ゲームにて，相関手法に基づく混合ナッシュ戦略を提案し，実験により収束性と協調行動が獲得できることを示す [7] [8].

1.5 論文の構成

本研究では，以上の問題について以下の構成で論じる．第 2 章では，非協力ゲームにおける行動選択を実験にて検証し，第 3 章では，非零和ゲームにおける新たな協調手法を提案する．第 4 章では，協力ゲームにおける 1 回限りのゲームでの行動選択について論じ，第 5 章では，協力・繰り返しゲームにおける協調行動の獲得と学習の収束性を論じる．第 6 章で結論として本研究をまとめる．

1.6 発表論文

1.6.1 査読付き論文

1. 井越 一穂, 三浦 孝夫, 塩谷 勇: Distributing Rewards by Strategic Knowledge based on Nash-Q learning, First IEEE International Conference on the Applications of Digital Information and Web Technologies (*ICADIWT*), 2008.
   マルチエージェントシステムの機械学習手法の 1 つとして，Q 学習が用いられている．本稿では，Q 学習の学習過程においてある条件を満たしたときに，均衡状態をゲーム理論ナッシュ均衡に近似することにより，より高速な戦略的知識の獲得を目指す．そして，実験にて手法の有効性を確認する．

2. 井越 一穂, 三浦 孝夫: Selecting Behavior on Repeated Local Effect Functions to appear in ”2010 International Symposium on Collaborative Technologies and Systems (*CTS 2010*)”, 2010.
   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.

1.6.2 研究論文

1. 井越 一穂, 三浦 孝夫: Q 学習のナッシュ均衡による戦略的知識の獲得, 第 19 回データベースワークショップ (*DEWS*), 2008.
   マルチエージェントシステムの機械学習手法の 1 つとして，Q 学習が用いられている．本稿では，Q 学習の学習過程においてある条件を満たしたときに，均衡状態をゲーム理論ナッシュ均衡に近似することにより，より高速な戦略的知識の獲得を目指す．そして，実験にて手法の有効性を確認する．

2. 井越 一穂, 三浦 孝夫: Local Effect Game における相関均衡を用いた行動選択, 電子情報通信学会 データ工学研究会, 電子情報通信学会技術研究報告, Vol.109(153), pp.7-11, 2009.
   本研究では，マルチエージェント環境における戦略の導出アルゴリズムを提案する．ここでは，均衡状態として報酬総計が最大となるような均衡である Correlated equilibria（相関均衡）を用い，その適用状況として Local Effect Game を想定する．このときエージェントの生成する戦略により，互いに協調し可能な限り総報酬が最大になることを実験で示す.

3. 井越 一穂, 三浦 孝夫: 繰り返し局所影響関数を用いた行動選択, 第 2 回データ工学と情報マネジメントに関するフォーラム (*DEIM*), 2010.
   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.

1.6.3 口頭論文

1. 井越 一穂, 三浦 孝夫: Local Effect Game における最適相関関係の導出, 電子情報通信学会 2009 年総合大会 学生ポスターセッション, 2009.
   本研究では，マルチエージェント環境における戦略の導出アルゴリズムを提案する．ここでは，均衡状態として互いの報酬が最大となるような均衡である Correlated equilibria（相関均衡）を用いる.

第2章　Q学習のナッシュ均衡による戦略的知識の獲得

エージェントの学習手法の 1 つとして Q 学習が知られている．Q 学習はシングルエージェントを対象としたものであり，Q 学習をそのままマルチエージェントを適用した場合には，学習の収束性が保証されないなど問題がある．本稿では，Hu らにより提案された Q 学習の価値関数の更新にナッシュ均衡を用いるナッシュ Q 学習と Q 学習を比較し，有効な戦略的知識が獲得できることを実験により検証し，有効性を確認する．

2.1　前書き

マルチエージェントシステムとは，1 つの計算空間で 複数の行動主体が自律的に存在するシステムをいう．この行動主体をエージェントと呼ぶ．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利得を獲得することが期待される．多く利得の獲得するためには，エージェントは行動方法を学習する必要がある．学習方法として，環境に影響されずまた教師データを必要としない強化学習がよいとされる．特に，解の存在を保証した Q 学習が使われることが多い．Q 学習では，原則として期待報酬である Q 値を更新することで最適な行動を選択とする．しかし，Q 学習は単一のエージェントシステムを想定しており，そのままマルチエージェントシステムに適用した場合，学習の収束性が保証されない．

本稿では，ナッシュ Q 学習手法 [3] を人生ゲーム (Game of Life) に適用し，戦略的知識を獲得できることを示す．すなわち，人生ゲームを非ゼロ和の非協力ゲームとみなし，ゲーム理論で用いられるナッシュ均衡を各エージェントの行動価値関数の更新に用いることができることを論じる [14]．この結果，位置や状態に応じた高度な協調動作をモデル化でき，またある条件下で Q 値の収束することが証明されているため，マルチエージェント環境での学習方式として望ましいことを示す．

関連研究としては，ナッシュ Q 学習を格子ゲーム (Grid Game) に適用し，ナッシュ Q 学習と Q 学習とを学習性能の点で比較し評価したものがある [11]．

第 2 章では，Q 学習を要約し，第 3 章で，ゲーム理論のナッシュ均衡，ナッシュ Q 学習に関して述べる．第 4 章では，ナッシュ Q 学習による戦略決定を提案し，第 5 章では，実験結果を示し本手法の有効性を示す．

2.2 強化学習

人間が何も予備知識を持たずに周囲の環境から物事を学ぶとき，まず状況を観測し，何らかの変化を期待して行動をとり，その結果 (報酬) を判断するというステップを繰り返すであろう．この「経験を積む」過程を，機械学習分野では強化学習と呼ぶ．行動の正しさは報酬で表現され，その値 (評価) を最大化しようとする [18]．報酬がすぐに評価となるわけではないので，定常状態になるまで繰り返す必要がある．

Q 学習は強化学習の代表的な手法である．学習は期待効果値 $Q(s,a)$ を算出することに対応する．ここで状態 s にて行動 a をとるときの期待効果値 $Q(s,a)$ で表す．このとき状態 s で選択する行動 $A(s)$ は，効果値 $U(s)$ が最大となるものを選ぶであろう[1]．すなわち $U(s)$ は次のように表される．

$$\begin{aligned} A(s) &= \arg\max_a Q(s,a) \\ U(s) &= Q(s,A(s)) \end{aligned}$$

Q 学習では， Q 値の更新を繰り返すことにより真の効果値に収束する．すなわち，新しく状態は現在の状態によってのみ確率的に決定できるという仮定をマルコフ決定性 (MDP) と呼ぶが，この仮定の下では Q 値の収束が証明されている [13].

状態 i から j への遷移確率を σ_{ij}，状態 s で行動 a をとるときの報酬 $R(s,a)$ としたとき，Q 値は直接の報酬と今後の期待値の和として定義され，次式で表現される．

$$Q(s,A(s)) \leftarrow R(s,A(s)) + \sum \sigma_{sj} \cdot U(j)$$

ここで $\sum \sigma_{sj} \cdot U(j)$ は効果の期待値を表す.

しかし，マルコフ過程における極限値としての定常確率 σ_{ij} を求めることが容易ではないため，報酬と次状態 s' の効果値を基に Q 値を更新するブートストラップ方法を用いる．この状況は 2 つのパラメタ α, γ を用いて次式で表される.

$$Q(s,A(s)) \leftarrow (1-\alpha)Q(s,A(s)) + \alpha(R(s,A(s)) + \gamma Q(s',A(s'))$$

ここで，$\alpha (0 \leq \alpha < 1)$ を学習率といい，当該行動により得られる報酬の優先度を表す．大きい学習率は報酬を重視した更新を表す．また，$\gamma (0 \leq \gamma \leq 1)$ を割り引き率といい，期待値 (次状態の効果値) の優先度を表す．大きい割り引き率は期待値を重視することを表す.

従って，選択の無い Q 値は更新されず学習範囲が小さくなる．実際，Q 値が局所解となるおそれがあり，この状況を回避するため，確率的な行動選択のための Softmax 手法 (ボルツマン分布による確率場を用いる) などが知られる [18].

Q 学習においては単一エージェントのみを仮定していることに注意したい．マルチエージェント環境では，他エージェントの行動により環境が変化し，各エージェント

[1] このような貪欲な (greedy) 行動選択では，動作は単純なものに限られるため，実際には分散させる手法を選ぶ.

は互いの観測をしないため，Q 値が定常化しない．この結果 Q 値に基づく行動選択が困難となる．いったいマルチエージェント環境に対応した行動選択の手法とな何なのであろうか？ナッシュ均衡原理は 1 つの解となりえる戦略であり，本稿の狙いもここにある．

2.3 ナッシュ均衡とナッシュ Q 学習

2.3.1 ナッシュ均衡

ゲーム理論とは，複数の行動主体が選択する戦略の決定手法であり，互いに自分の報酬を最大化することを目的としている [14]．ゲーム理論において，行動主体をプレイヤと呼ぶ．本稿では 2 名のプレイヤを仮定する．プレイヤが協調せず (互いに交渉すること無く) 同時に戦略選択するゲームは非協力同時的であると呼ばれる．非協力同時ゲームで戦略 A, B のいずれかを選択するとき，各プレイヤがそれぞれの戦略で報酬を得る．報酬は (競合することから) 組み合わせによって異なるため，これを 2×2 の表 (利得行列という) で表す．各項目 (α, β) はそれぞれプレイヤ 1,2 が戦略 A, B を選択したときに得る報酬を表す．

プレイヤの戦略がナッシュ均衡 (Nash Equilibrium) であるとは，互いの戦略が相手の戦略に対して最適な反応となるときをいう．本稿でいう”反応”は報酬を意味し多いほど良い．他のプレイヤが戦略を変えない限り，自分が戦略を変える動機を持たない状態を最適であるという．

		プレイヤ 2	
		戦略 A	戦略 B
プレイヤ 1	戦略 A	$(3, 3)$	$(5, 2)$
	戦略 B	$(2, 5)$	$(4, 4)$

表 2.1: ナッシュ均衡 – 双方に支配戦略がある場合

表 2.1 で，プレイヤ 2 が戦略 A を選択した場合，プレイヤ 1 が 戦略 A を選択するほうが大きな報酬 3 を得る．またプレイヤ 2 が戦略 B を選択した場合でもプレイヤ 1 が戦略 A を選択するほうが大きな報酬 5 を得る．これを，プレイヤ 1 の戦略 A は 戦略 B の対して支配的 (dominant) であるといい，この戦略 A をプレイヤ 1 の支配戦略という．同様にプレイヤ 2 の支配戦略は戦略 A である．つまりプレイヤ 1 ，プレイヤ 2 にとって (A, A) は互いに支配的でありただ 1 つのナッシュ均衡戦略は (A, A) となり，このとき $(3, 3)$ の報酬を得る．互いが最適な行動を選択するため，プレイヤの得る報酬総和が最適となり，全体として協調動作を獲得することになればよい[2]．ナッシュ均

[2] しばしば”Win-Win” の関係にあるといわれるのはこの特性を意味する．

衡に基づく協調動作をとることにより，報酬総和の増加が期待でき，マルチエージェント環境の解を生成する可能性を有する．

	プレイヤ 2		
		戦略 A	戦略 B
プレイヤ 1	戦略 A	$(3,3)$	$(6,4)$
	戦略 B	$(4,6)$	$(2,2)$

表 2.2: ナッシュ均衡 – 支配戦略が無い場合

支配解は存在しないがナッシュ均衡解が存在することがある．表 2.2 では，プレイヤ 2 が戦略 A を選択した場合，プレイヤ 1 が 戦略 B を選択するならば，それ以外の方法より大きな報酬 4 を得る．しかしプレイヤ 2 が戦略 B を選択した場合，プレイヤ 1 は 戦略 A をとれば報酬 6 を得るので， B での報酬 2 より大きい．すなわち $(B,A),(A,B)$ が最適である．同様にプレイヤ 1 が 戦略 A を選択するならば，プレイヤ 2 は B においてより大きな報酬 6 を，プレイヤ 1 が B をとればプレイヤ 2 は A においてより大きな報酬 4 を得る．すなわち $(A,B),(B,A)$ が最適である．つまりプレイヤ 1 ，プレイヤ 2 双方にとって支配的な戦略は無いが最適である解 $(A,B),(B,A)$ はナッシュ均衡である．このとき報酬 (4,6), (6,4) を得る．

	プレイヤ 2		
		戦略 A	戦略 B
プレイヤ 1	戦略 A	$(3,3)$	$(4,2)$
	戦略 B	$(4,2)$	$(3,3)$

表 2.3: ナッシュ均衡が無い場合

ナッシュ均衡戦略が常に存在するとは限らない [14]．利得表 2.3 で，プレイヤ 1 にとって最適な解は $(A,B),(B,A)$ であるがプレイヤ 2 の最適解は $(A,A),(B,B)$ である．従って互いに最適となる共通解は存在しない．

支配解であっても必ずしも最大報酬を得るものではない[3]．実際，利得表 2.1 において (A,A) での報酬 (3,3) よりも (B,B) での報酬 (4,4) のほうが両プレイヤともに大きくなっている．どのプレイヤの報酬も低下すること無く利得を増やすことができない解をパレート最適 (Pareto optimum) という．ナッシュ均衡解は必ずしもパレート最適ではなく，またナッシュ均衡解が存在しない利得表 2.3 で $(A,A),(B,B)$ はパレート最適となる．つまりナッシュ均衡解は最悪解でないことを保証するに過ぎないことに注意したい．

[3] この問題は囚人のジレンマと呼ばれる．

2.3.2 ナッシュ Q 学習

ナッシュ均衡を Q 値の更新と行動選択に用いるように拡張した Q 学習をナッシュ Q (Nash Q) 学習と呼ぶ．マルチエージェント環境において，Q 値は更新を繰り返すことによりナッシュ均衡値 (期待報酬値) に近づき，各エージェントは (ナッシュ均衡戦略の意味で) 正しい行動をとる．また，一定の条件下で Q 値の収束が保証される [3]．ナッシュ Q 学習方式では，行動選択原理としてナッシュ均衡を用いるため，互いに最適な行動選択が可能になる．つまり最適な総和利得の獲得により，協調動作をとることが可能となる．

ナッシュ Q 学習において，各エージェントは全てのエージェントの Q 値を観測でき，この結果それぞれの行動と報酬が観測できると仮定する[4]．シングルエージェント Q 学習と同様に Q 値の更新を繰り返すが，割り引き対象となる次状態の Q 値の利用を，(期待値ではなくて) ナッシュ均衡値を用いる．エージェント i の Q 値を $Q^i(s, a^1, ..., a^n)$ で表す．エージェント $1, ..., n$ がとる行動 $a^1, ..., a^n$ に対し状態 s から s' に遷移するとき，Q 値の更新を次式で定義する．

$$\begin{aligned} Q^i(s, a^1, ..., a^n) &\leftarrow Q^i(s, a^1, ..., a^n) \\ &\quad +\alpha[r^i + \gamma NashQ^i(s') - Q^i(s, a^1, ..., a^n)] \end{aligned} \tag{2.1}$$

$$NashQ^i(s') = [\pi^1(s')...\pi^n(s')Q^i(s')] \tag{2.2}$$

ここで r^i はエージェント i が受けとる報酬，学習率 α，割り引き率 γ は Q 学習と同じである．全てのエージェントがナッシュ均衡戦略をとったときの最適報酬組を考え，エージェント j の状態 s' における戦略 $\pi^j(s')$ のもとでエージェント i が状態 s' で得られる Q 値の期待値を，$NashQ^i(s')$ で表す．

ナッシュ Q 学習の行動選択手法は，各エージェントが Q 値によるナッシュ均衡の組を選択すると仮定する．しかし，ナッシュ Q 学習においても行動が選択されない Q 値は更新されないため， Q 学習と同様に確率的な行動選択手法を用いる必要がある．Hu は ϵ 貪欲法に基づいて全エージェントがナッシュ均衡戦略に従うとし，確率 ϵ でナッシュ均衡戦略を，確率 $1-\epsilon$ でランダム戦略を選ぶとしている [3]．本稿も全てのエージェントがナッシュ均衡戦略に従うと仮定し，前述のボルツマン分布に基づく Softmax 手法を用いる．ナッシュQ 均衡は複数存在する場合があるため，Hu ははじめに発生したナッシュ均衡を用いるとき第 1 ナッシュ(first-Nash) 手法，複数の均衡が発生した場合の 2 番目のナッシュ均衡を用いるとき第 2 ナッシュ(second-Nash) 手法，最も多くの利得が期待できるナッシュ均衡を最適ナッシュ(best-expected-Nash) 手法と呼び，算出方法を提案しているが，本稿の計算手法も最適ナッシュ手法に従う．

[4] 人生ゲームでは，すごろく盤面や選択するカード内容，参加者の行動とその結果を全て知ることができる．

2.4　ナッシュQ 学習に基づく人生ゲーム戦略

この章では人生ゲームのゲーム理論への適用と，行動の選択原理をナッシュ均衡戦略に従うためのモデル化を示す．

2.4.1　人生ゲームのモデル化

本稿で論じる”人生ゲーム” (Game of Life) とは，2 人のプレイヤが N 個のマス目からなる 1 次元盤面上を，スタート位置からゴールまで同時にサイコロを振って進む，すごろくゲームの一種である．単純化のため，サイコロは $0, -1, +1$ の目だけを持つものとする．途中，マス目のいくつかにマークがあり，各プレイヤの得る報酬の選択や位置変化に従いながら，先にゴールへ到達してゲームが終了する．マークのあるマス目では，プレイヤはカードを引く．カードは，報酬額 V とその総額に関する 2×2 の利得表 M が含まれ，プレイヤは選択できる戦略を決定する．一度使ったカードは戻さない．当該マス目に 1 プレイヤしかいなければ報酬額 V を全額獲得できる．しかし，2 プレイヤが同時に同一マス目にいるときは，競合が発生したと考え報酬額 V を分け合う．ここでは，利得行列 M に従う戦略を選択してより大きな報酬を得るよう行動する．この人生ゲームでは，各プレイヤの位置と各プレイヤがどの指示を選択したかにより，得られる報酬が変化する．M は次の形式をとる．

M ＝

		プレイヤ 2	
		A	B
プレイヤ 1	A	(a, b)	(c, d)
	B	(e, f)	(g, h)

ここで，利得表は報酬額 V をプレイヤ間でどのように配分するかの割合を表す．例えば，プレイヤ 2 が戦略 A を選択したとき，プレイヤ 1 が A を選択すれば a, B を選択すれば e の報酬を得ることができる．ここで $a, .., h$ は V の分配割合を表し，$0.0 \leq a, ..., h \leq 1.0, a + b = c + d = e + f = g + h = 1.0$ となる．従って実際に得られる報酬は $V \times a, V \times b$ 等となる．盤面上の位置や引いたカード内容およびプレイヤの選択した戦略は他のプレイヤから参照できる．

ゲームでは，プレイヤ $i = 1, 2$ の位置を状態 $s, 1 \leq s \leq N$ とみなし，マス目での指示を行動として，マルチエージェントシステム環境に対応させる．マークのあるマス目で引く各カードには V, M が記され，その内容 V, M は他のプレイヤも知ることができる．ゲームの開始時は，累積報酬 H_1, H_2 および手持ち報酬額 T_1, T_2 がともに 0 であり，マス目 1 にいる状態でサイコロを同時に振る．サイコロの値に従って，プレイヤは移動するが，1 より以前にも N より後にも移動できない．N に達すればゴールとみなし，勝者 i の手持ち額だけをその累積報酬 H_i に加算する．一定回数繰り返して，勝敗および累積報酬の大小で評価する．

2.4.2 人生ゲームでのナッシュ Q 学習

本稿で論じる人生ゲームでは，報酬 V をプレイヤで完全に分け合うため，等式 $b = 1-a, d = 1-c, f = 1-e, h = 1-g$ が成り立つ．このため，(a, b) が支配解であるためには，$\{a, c, e, g\}$ において c, e がそれぞれ最大値，最小値であることが必要十分である[5]．同様に $(c, d), (g, h), (e, f)$ が支配解であるためには，a, c, e, g においてそれぞれ $a \cdot g$, $e \cdot c$, $g \cdot a$ が最大・最小であることが条件となる．また，支配解でないナッシュ解は存在せず[6]，ナッシュ解は必ずパレート最適となる[7]．ナッシュ解が存在しなければ，任意に項目を選択するとする．

本稿で論じる人生ゲームでは，ナッシュ Q 学習は，プレイヤ i $(i = 1, 2)$ が自らの行動 a^i $(a^i \in A(s)^i)$ と報酬　r^i (s, a^1, a^2) だけではなく，他のプレイヤの行動 a^{-i} $(a^{-i} \in A(s)^{-i})$ と報酬 $r^{-i}(s, a^1, ..., a^n)$ を観測する．ここで $A(s)^j$ とはプレイヤ j が状態 s が選択可能な全行動である．

そのため，プレイヤ i は マス目位置 s においてプレイヤ j が 5 とる戦略がナッシュ均衡であるとみなして自らも該当する戦略を選択し，Q 値:$Q^i(s, a^1, a^2)$ は 式 (2.1) より更新される．

2.5 実験

ここではナッシュQ 学習手法により人生ゲームにおけるエージェントが効果的に学習することを検証する．

2.5.1 準備

本実験では，”Happy Academic Life 2006: 研究者の人生ゲーム ”を利用（人工知能学会）を基にしたゲームを使用する[8][21]．スタートからゴールまでの 23 マスを 1 周，人生ゲーム 1 セットを 10000 周とし，人生ゲーム毎の収入のばらつきを抑えるため，人生ゲームを 50 セット行い, 収入合計の平均をとる．ここで，収入とはエージェントの 1 周の報酬の合計と定義する．また，総収入を 1 セット (10000 周) 分の収入合計とする．全てのマス目にマークがついている．従ってどのマス目でもカードを引き, そのカードに書かれた指示により報酬を受けとる [21]．(図 2.1).

[5]実際，定義より $a \geq e, c \geq g$ かつ $b \geq d, f \geq h$ であるから $1-a \geq 1-c, 1-e \geq 1-g$ が成り立ち，これを書き換えると $c \geq a, g \geq e$ となる．

[6]存在すれば，$a \geq e, c \leq g, b \geq d, f \leq h$ が同時に成り立つが，これらは矛盾する．

[7]支配解がパレート最適でなければ，$aleg, b \leq h$ から $a = g$ を得る．

[8]セルラーオートマトンで著名な Life Game ではない．

図 2.1: ゲームの実行画面

カードには，5 つの値 V, a, c, e, g が記されている．ここで $V, 0 \leq V \leq 100$ はこのカードの報酬, $0.0 \leq a, c, e, g \leq 1.0$ であり，これらは利得行列 M を規定する:

$M =$

		プレイヤ 2	
		A	B
プレイヤ 1	A	(a, b)	(c, d)
	B	(e, f)	(g, h)

ただし $b = 1 - a, d = 1 - c, f = 1 - e, h = 1 - g$ である．実験ではこれらは一様乱数としてその都度生成される．

本稿では 2 つの実験を行う．実験 1 では「エージェントが対等な場合」を検証する．対等とは，行動選択が同等・同程度の勝敗確率であることをいう．実験 2 では「エージェントが対等ではない場合」を検証する．ここでは『片方のエージェント (エージェント B) が確率 0.1 で行動選択手法とは逆の行動を選択する』という制限を加え，ゴール時に『一方のエージェントよりも 1 周で獲得した収入が多いエージェントの収入に $+1,500$』『互いの収入の差が ± 500 の場合は双方のエージェントにあまり差が無いと考え，双方の収入に $+1,500$』とする．

学習で用いる更新式の係数パラメータは，学習率 $\alpha = 0.01, 0.1, 0.5$，割り引き率 $\gamma = 0.5, 0.9, 0.99$，行動選択手法で用いる温度パラメータを $T = 1, 10, 100$ とする[9]．

実験結果の評価のため，収入，および各週での収入の差，およびそれらの合計を用いて比較する．ナッシュ Q 学習の収入差合計が小さい場合には，協調動作により利得が改善したと考えられる．

また，各エージェントの勝敗数を比較する．ここでは， 1 周毎に各エージェントの

[9]行動選択手法の softmax 手法ボルツマン分布の温度パラメータ T は，行動選択のばらつきを調整するもので，大きい場合には全行動が均一的に選択されることになり，小さい場合には最大の価値を持つルールをより多く選択するようになる．

収入の大小を比較し，収入が多いエージェントを勝者と定義する．協調行動をとっている場合は，両者の収入は拮抗し両者の勝ち回数は同等になる．

2.5.2 実験結果

エージェントが対等な場合

Q 学習の 10000 周までの総収入の合計の上位 10 項目は，表 2.4 の通りである．表 2.4 の上位 10 項目の平均学習率が 0.02 と低い値である．よって，学習率が低いものほど総収入の合計が大きいといえる．これは，相手が適当ではない行動をとった場合でも，学習率が低いために Q 値の更新に大きな影響が出なかったためであるといえる．

温度	学習率	割り引き率	総収入の合計
1	0.01	0.5	172080358.1
1	0.01	0.9	170830262
100	0.01	0.5	170356331.9
100	0.01	0.99	170159024.1
100	0.01	0.9	169711550.2
1	0.01	0.99	169666952.2
10	0.01	0.9	169263582.1
10	0.01	0.99	169055784.1
10	0.01	0.5	168857273.8
1	0.1	0.5	160137378.3

表 2.4: Q 学習の総収入の合計 上位 10 項目の一覧

ナッシュ Q 学習 10000 周までの総収入の合計の上位 10 項目は．表 2.5 の通りである．表 2.5 の上位 10 項目の平均学習率が 0.86 と低く，平均温度と 63.1 と高い．そのため，学習率が低く温度が高いほど総収入の合計が大きいといえる．学習率が低いため相手の不適当な行動の影響を受けにくく，温度が高いため行動選択のランダム性が高くなり，より適切な行動の学習ができたといえる．

表 2.4 より最も総収入の合計が高い Q 学習 (温度 1 学習率 0.01 割り引き率 0.5) と，表 2.5 より最も総収入の合計が高いナッシュ Q 学習 (温度 100 学習率 0.01 割り引き率 0.99) を選び比較すると表 2.6 になる．100 周における累計では，ナッシュ Q 学習より Q 学習が 736116 大きい．しかし，10000 周では Q 学習よりナッシュ Q 学習が 2825046 大きい．これは，ナッシュ Q 学習の学習速度が遅いことによる．

最も総収入の合計が高い Q 学習 (温度 1 学習率 0.01 割り引き率 0.5) と，最も総収入の合計が高いナッシュ Q 学習 (温度 100 学習率 0.01 割り引き率 0.99) について，協調行動をとっているかの結果は 2.7 の通りである．2.7 の平均で比較すると， Q 学習

温度	学習率	割り引き率	総収入の合計
100	0.01	0.99	174905403.9
10	0.01	0.99	174412427.8
100	0.01	0.9	171948664.2
10	0.01	0.5	171166999.9
100	0.1	0.99	170772763.7
10	0.01	0.9	169503103.9
100	0.1	0.9	169211573.7
1	0.01	0.5	165547050.1
100	0.1	0.5	162908275.9
100	0.5	0.5	162237957.8

表 2.5: ナッシュ Q 学習の総収入の合計 上位 10 項目の一覧

のエージェント A が”勝つ”割合が 42.34%，ナッシュ Q 学習のエージェント A が”勝つ”割合が 51.45%で，ナッシュ Q 学習が 50%に近い．従って，ナッシュ Q 学習では 2 エージェント間の勝つ回数の差が少ないことから，協調行動をとっているといえる．

エージェント A の”勝つ”回数をパラメータ別に一覧にすると表 2.8 のようになる．Q 学習で全 27 項目中の 19 項目，ナッシュ Q 学習で全 27 項目中の 23 項目でエージェント A の”勝つ”割合が 50%以下である．これは，選ばれたカードの特性によりエージェント A が負けやすい条件となっていることによる．本実験の Q 学習で平均 46.60%，ナッシュ Q 学習で 47.65%でともに 50%に近く，双方とも協調行動をとっているといえる．

従って，互いのエージェントが対等な場合において， Q 学習とナッシュ Q 学習の差が顕著ではない．

エージェントが対等ではない場合

エージェント A の”勝つ”回数をパラメータ別に一覧にすると表 2.9 の通りである．Q 学習，ナッシュ Q 学習ともにエージェント A が”勝つ”割合が 50%以下でありエージェント A にとって不利な条件であることがわかる．エージェント A の”勝つ”割合は， Q 学習が 20.94%，ナッシュ Q 学習が 27.41%で Q 学習よりも”勝つ”割合が 50%に近いものが多く，協調行動をとろうとしていると考えられる．

表 2.9 より， Q 学習とナッシュ Q 学習において最も”勝つ”割合が 50%に近い，Q 学習 (温度 1，学習率 0.01，学習率 0.01) とナッシュ Q 学習 (温度 100，学習率 0.1，割り引き率 0.5) を選び，収入の差を算出したのが表 2.10 である．1000 周までの時点で，ナッシュ Q 学習の収入の差が 204.04， Q 学習の収入の差が 64.24 でナッシュ Q 学習の収入の差のほうが大きいが，これはナッシュ Q 学習の学習速度遅いためである．10000 周でナッシュ Q 学習の収入の差が 32.32， Q 学習の収入の差が 1034.34 とナッシュ Q

	Q 学習		ナッシュ Q 学習	
周数	収入	累計	収入	累計
100	15726	1572564	8364	836448
1000	17244	17018932	17719	16427688
2000	17262	34265212	17748	34142888
3000	17238	51504728	17637	51852108
4000	17225	68735450	17603	69470580
5000	17227	85967616	17582	87068388
6000	17211	103184976	17565	104646702
7000	17217	120407362	17587	122208210
8000	17208	137628240	17600	139777440
9000	17245	154850628	17554	157333390
10000	17211	172080358	17556	174905404

表 2.6: 100 周毎の累計 (1000 周毎抜粋)

学習は差が小さく，協調行動によりエージェント同士の収入の差が小さくなっているいえる．

表 2.11 は， Q 学習 (温度 1，学習率 0.01，学習率 0.01) とナッシュ Q 学習 (温度 100，学習率 0.1，割り引き率 0.5) でエージェント A，B の”勝つ”回数を比較するものである．Q 学習では 10000 周の時点でエージェント B の”勝つ”割合が 71.58%，エージェント A の”勝つ”割合が 28.42%と，エージェント B が多く勝っている．ナッシュ Q 学習では 10000 周の時点でエージェント A が”勝つ”割合が 49.72%，エージェント B が”勝つ”割合が 50.28%とほぼ同数となっており協調行動をとっているものといえる．

2.5.3 考察

実験 1 では， Q 学習・ナッシュ Q 学習の総収入の合計に差はみられない．表 2.6 の実験結果より， Q 学習とナッシュ Q 学習の総収入の合計の比較を表 2.12 に示す．表 2.12 では，報酬の改善率は 1.64%にとどまっている．これは，実験で使用した人生ゲームにおいて適当である行動と，適当でない行動を選択した場合での報酬の差が小さいためである．そのため，大きな利得を得るためには互いの利得を尊重し協調する必要が少ない．また，行動の選択が 2 択と少ない．

表 2.13 は実験 1 における各エージェントの”勝ち”回数の合計である．表 2.8 および表 2.13 より， Q 学習とナッシュ Q 学習ではナッシュ Q 学習が協調行動をとっているといえる．しかし， Q 学習のエージェント A の”勝ち”回数の合計の割合が 46.40%で 50%と 3.6%の差しか無いため Q 学習でも協調行動を選択しているといえる．

実験 2 では，ナッシュ Q 学習がより協調行動を獲得していることが確認できる．表

	Q 学習		ナッシュ Q 学習	
周数	エージェント A	エージェント B	エージェント A	エージェント B
100	43.82%	56.18%	46.60%	53.40%
1000	42.00%	58.00%	52.76%	47.24%
2000	41.10%	58.90%	52.96%	47.04%
3000	41.64%	58.36%	50.82%	49.18%
4000	42.04%	57.96%	51.24%	48.76%
5000	42.94%	57.06%	51.00%	49.00%
6000	41.72%	58.28%	51.34%	48.66%
7000	42.26%	57.74%	51.74%	48.26%
8000	43.10%	56.90%	50.98%	49.02%
9000	42.08%	57.92%	51.14%	48.86%
10000	43.22%	56.78%	51.44%	48.56%
平均	42.34%	57.66%	51.45%	48.55%

表 2.7: 100 周毎の平均の”勝つ”回数の割合 (1000 周毎抜粋)

2.14 は実験 2 における各エージェントの”勝ち”回数の合計である．表 2.9 より， Q 学習よりも 6.47%協調状態に近いことがわかる．また表 2.11 より， Q 学習・ナッシュ Q 学習で最も協調行動を選択しているものを比較した場合には， Q 学習でエージェント A の勝つ頻度 3 割，ナッシュ Q 学習でエージェント A の勝つ頻度 5 割となり，より協調行動をとっていることが確認できる．これは，ナッシュ Q 学習の協調行動がうまく働いているためであるといえる．従って，互いのエージェントが対等ではない場合には，ナッシュ Q 学習は Q 学習よりも多く協調行動をとることが可能だといえる．

表 2.15 より，温度が 100 の場合には他の温度に比べ 10.63%以上協調行動をとっている状態に近い．よって，温度が高いの場合により協調行動を学習しているといえる．これは，温度が高いことで行動にある程度のランダム性が生まれよりよい行動を選択していることによる．このことから，適切な学習率・温度などのパラメータを選択する必要がある．

2.6 結論

本稿では，実験にて Q 学習とナッシュ Q 学習を比較することにより，ナッシュ Q 学習が戦略的知識として協調行動を学習することを確認した．また，ナッシュ Q 学習は学習率・温度（行動選択手法ボルツマン分布）により結果が変化することを確認した．

今回の実験では，ほぼ対等の場合において，総収入の合計や勝ち負けの回数に Q 学習との明確な有意差が確認できなかった．そのため，行動の選択肢が多い場合に対して有用性を確認する必要がある．

また，互いのエージェントが対等でない場合において， Q 学習との差が確認でき，ナッシュ Q 学習において競合状態が学習できることを示した．このことから，ナッシュ Q 学習は互いのエージェントが対等でない場合の協調行動が必要なときに特に有用であるといえる．

Q 学習				ナッシュ Q 学習			
温度	学習率	割り引き率	割合	温度	学習率	割り引き率	割合
10	0.5	0.99	58.75%	1	0.5	0.9	56.37%
1	0.1	0.5	52.26%	100	0.5	0.9	53.58%
10	0.5	0.5	52.00%	100	0.01	0.99	51.45%
100	0.5	0.99	51.96%	10	0.01	0.9	51.02%
1	0.01	0.99	51.56%	1	0.5	0.5	49.96%
1	0.1	0.99	51.48%	100	0.5	0.99	49.93%
100	0.5	0.9	50.61%	1	0.01	0.9	49.91%
10	0.01	0.99	50.25%	10	0.01	0.99	49.27%
10	0.5	0.9	49.43%	10	0.5	0.9	48.60%
1	0.01	0.9	48.75%	1	0.01	0.99	48.57%
100	0.1	0.9	48.06%	1	0.1	0.9	48.14%
100	0.5	0.5	47.22%	1	0.1	0.99	47.97%
10	0.1	0.99	47.02%	100	0.1	0.99	47.79%
100	0.01	0.5	46.14%	1	0.5	0.99	47.68%
10	0.1	0.5	46.07%	100	0.5	0.5	47.63%
1	0.1	0.9	45.72%	100	0.01	0.9	47.01%
1	0.5	0.5	44.40%	100	0.01	0.5	46.31%
100	0.01	0.9	43.96%	100	0.1	0.9	46.03%
1	0.5	0.9	43.64%	10	0.5	0.5	45.83%
1	0.5	0.99	43.61%	100	0.1	0.5	45.61%
10	0.01	0.5	42.36%	1	0.1	0.5	45.49%
1	0.01	0.5	42.34%	10	0.01	0.5	45.07%
100	0.1	0.99	42.22%	10	0.5	0.99	45.02%
100	0.01	0.99	39.97%	10	0.1	0.9	44.97%
10	0.1	0.9	38.91%	10	0.1	0.5	43.73%
10	0.01	0.9	38.22%	10	0.1	0.99	42.92%
100	0.1	0.5	36.00%	1	0.01	0.5	40.60%
平均			46.40%				47.65%

表 2.8: エージェント A の”勝つ”割合

今後に課題として，本実験では 2 体のエージェントが同時に毎回 1 マスずつ進む．そのため位置依存性への配慮をしていない点がある．

また，ナッシュ均衡が一意解とは限らない，純粋戦略のみを考えたときには解が存在しない場合がある，ナッシュ均衡解がパレート最適性を保証しない，というナッシュ均衡自体の問題がある．

Q 学習				ナッシュ Q 学習			
温度	学習率	割り引き率	割合	温度	学習率	割り引き率	割合
1	0.01	0.99	35.66%	100	0.1	0.5	48.66%
10	0.01	0.99	34.73%	100	0.5	0.5	43.71%
100	0.01	0.9	34.43%	100	0.01	0.5	42.88%
10	0.01	0.9	32.48%	100	0.01	0.9	42.50%
10	0.01	0.5	31.86%	100	0.01	0.99	36.19%
100	0.01	0.5	31.56%	100	0.1	0.99	34.25%
1	0.01	0.9	29.64%	10	0.01	0.5	34.14%
100	0.01	0.99	28.41%	100	0.1	0.9	31.63%
1	0.01	0.5	27.90%	10	0.5	0.5	31.45%
10	0.1	0.5	24.11%	10	0.01	0.9	30.44%
100	0.1	0.9	22.22%	10	0.01	0.99	29.72%
1	0.1	0.5	21.63%	10	0.5	0.9	27.17%
1	0.1	0.9	20.26%	1	0.5	0.5	26.49%
100	0.1	0.99	19.95%	100	0.5	0.9	23.68%
100	0.1	0.5	18.97%	1	0.01	0.99	23.05%
10	0.1	0.99	18.76%	1	0.5	0.9	22.40%
10	0.1	0.9	17.90%	1	0.5	0.99	22.11%
1	0.1	0.99	17.05%	100	0.5	0.99	21.52%
1	0.5	0.99	15.76%	10	0.1	0.99	21.42%
100	0.5	0.9	13.74%	1	0.1	0.9	20.22%
100	0.5	0.99	12.28%	10	0.1	0.5	19.13%
10	0.5	0.99	12.16%	1	0.1	0.99	18.86%
1	0.5	0.9	10.73%	1	0.1	0.5	18.64%
10	0.5	0.9	10.52%	10	0.1	0.9	18.59%
1	0.5	0.5	8.87%	1	0.01	0.9	18.37%
100	0.5	0.5	7.02%	10	0.5	0.99	17.22%
10	0.5	0.5	6.79%	1	0.01	0.5	15.51%
平均			20.94%				27.41%

表 2.9: エージェント A の”勝つ”割合

周数	Q 学習 温度 1 学習率 0.01 学習率 0.01	ナッシュ Q 学習 温度 100 学習率 0.1 割り引き率 0.5
100	108.8	179.74
1000	64.24	204.04
2000	183.12	109.56
3000	410.86	13.44
4000	677	0.04
5000	829.52	87.62
6000	898.6	67.14
7000	980.94	88.46
8000	1019.16	109.24
9000	1056.86	89.86
10000	1034.34	32.32

表 2.10: 100 周毎の平均の収入の差 (1000 周毎抜粋)

	Q 学習		ナッシュ Q 学習	
周数	エージェント A	エージェント B	エージェント A	エージェント B
100	46.18%	53.82%	43.38%	56.62%
1000	49.08%	50.92%	43.54%	56.46%
2000	46.52%	53.48%	45.36%	54.64%
3000	41.56%	58.44%	48.70%	51.30%
4000	35.80%	64.20%	48.70%	51.30%
5000	31.86%	68.14%	50.30%	49.70%
6000	29.62%	70.38%	49.94%	50.06%
7000	28.48%	71.52%	50.06%	49.94%
8000	28.84%	71.16%	51.24%	48.76%
9000	28.00%	72.00%	50.96%	49.04%
10000	28.42%	71.58%	49.72%	50.28%

表 2.11: 100 周毎の平均の”勝つ”回数の割合 (1000 周毎抜粋)

	総収入の合計	改善率
Q 学習	174905403.9	-
ナッシュ Q 学習	172080358.1	1.64%

表 2.12: 総収入の合計

	エージェント A	エージェント B
Q 学習	6264648(46.40%)	7235352(53.60%)
ナッシュ Q 学習	6432296(47.65%)	7067704(52.35%)

表 2.13: 実験 1 の各エージェントの”勝ち”回数の合計

	エージェント A	エージェント B
Q 学習	2826983(20.94%)	10673017(79.06%)
ナッシュ Q 学習	3699749(27.41%)	9800251(72.59%)

表 2.14: 実験 2 の各エージェントの”勝ち”回数の合計

温度	割合
100	36.11%
10	25.48%
1	20.63%

表 2.15: 実験 2 温度別エージェント A の”勝ち”回数の割合

第3章　ナッシュ Q 学習に基づく戦略知識による報酬分配

本稿では，位置と報酬の観点から，複数のゲームプレイヤが繰り返し非零和ゲームにおいて報酬分配による協調手法について論じる．今までも報酬分配に関する研究がされており，高度で高次元な知識が学習過程で得られるとされる強化学習が有効であると考えられている．強化学習の手法の 1 つとして Q 学習が知られているが，これは単一のエージェントを対象としたものである．Q 学習をそのままマルチエージェント環境に適用すれば，学習の収束性が保証されず，最適原理に基づいているか疑問である．本稿では，非零和確率ゲームとみてゲーム理論を用いた報酬分配問題を論じる．すなわち，Q 学習の価値関数の更新にナッシュ均衡を導入したナッシュ Q 学習に基づいて，有効な戦略的知識が獲得できることを複雑なゲーム (*game of life*) による実験により検証し，マルチエージェントシステムにおける機械学習手法の 1 つとして Q 学習が適用できることを示す.

3.1　前書き

マルチエージェントシステムとは， 1 つの計算空間で 複数の行動主体が自律的に存在するシステムをいう．この行動主体を**エージェント**と呼ぶ．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利得を獲得することが期待される．多く利得の獲得するためには，エージェントは行動方法を学習する必要がある．学習方法として，環境に影響されず，**教師データ** [16] を必要としない，また学習途中でも行動決定が可能な**強化学習**がよいとされる．特に， **マルコフ決定過程** (MDP) に基づき最適な期待報酬を保証した Q 学習が使われることが多い．Q 学習では，原則として期待報酬である Q **値** を更新することで最適な行動を選択とする．単一エージェントの下におけるマルコフ性により， Q 値が最適値に収束することが証明されいる．

しかしマルチエージェントに関してはこれらの議論を適用することは容易ではない，エージェントが全体としてどのような報酬を得ればよいのかの基準が自明ではないからである．ゼロ和ゲーム環境では一方の利得は他方のロスになるから単一エージェントとみなせる．しかし *game of life* のような非ゼロ和ゲーム環境では基準が無い．全く形式的に Q 学習をマルチエージェントシステムに適用しても，マルコフ性が欠けてお

り学習の収束性が保証されない．

本稿では，ゲーム理論に基づく繰り返し非零和ゲームを用い，どのように報酬を分配するかを論じる．ここでは，ナッシュ Q 学習手法 [3] を人生ゲーム (Game of Life) を繰り返し非零和ゲームとみなし，報酬分配のための戦略的知識を獲得できることを示す．ナッシュ均衡の意味での最適解を探し [17]，収束性を持ちつつ協調しエージェントが行動するよう対処する．

基本的な ナッシュQ 学習は，ゲーム理論のナッシュ均衡を基にした値の繰り返し学習アルゴリズムである．Littman [12] は零和ゲームでその概念を導入し，Hu [2] はナッシュ均衡に基づく新しいあるリズムを示しそれらを応用している [3]．Kitahara [11] はそのアルゴリズムを Grid games で実験している．

一方 Shoham [17] は NashQ 学習を批判し収束を保証せず一意解でないときの対応が明確ではないことを理由に基準として用いるのがふさわしくないと主張した．にもかかわらず NashQ はQ学習にて単純でありゲーム理論と強化学習を直感的に結びつける点で魅力的である．しかも，今回提案する報酬完全分配処理ではナッシュ解は高々 1 つであり存在しない場合でもパレート最適であることからこの問題は生じない．

第 2 章では， Q 学習を要約し，第 3 章で，ゲーム理論のナッシュ均衡，ナッシュ Q 学習に関して述べる．第 4 章では，ナッシュ Q 学習による戦略決定を提案し，第 5 章では，実験結果を示し本手法の有効性を示す．最後に第 6 章で結論とする．

3.2 Q 学習

一般的に，人間が何も予備知識を持たずに周囲の環境から物事を学ぶとき，まず状況を観測し，何らかの変化を期待して行動をとり，その結果 (報酬) を判断するというステップを繰り返すであろう．報酬がすぐに評価となるわけではないので，定常状態になるまで繰り返す必要がある．この「経験を積む」過程を，機械学習分野では**強化学習**と呼ぶ．行動の正しさは報酬で表現され，その値 (評価) を最大化しようとする [16].

Q 学習は強化学習の代表的な手法である．学習は期待効果値 $Q(s,a)$ を算出することに対応する．ここで状態 s で行動 a をとるときの期待効果値 $Q(s,a)$ で表す．このとき状態 s で選択する行動 $A(s)$ は，効果値 $U(s)$ が最大となるものを選ぶであろう．すなわち $U(s)$ は次のように表される．$A(s) = MaxArg_a Q(s,a), \quad U(s) = Q(s,A(s))$. このような戦略を **貪欲な** (greedy) 手法という．学習過程では， Q 値の更新を繰り返すことにより真の効果値に収束する．すなわち，新しく状態は現在の状態によってのみ確率的に決定できるという仮定をマルコフ決定性 (MDP) と呼ぶが，この仮定の下では Q 値の収束が証明されている [16].

状態 i から j への遷移確率を σ_{ij}，状態 s で行動 a をとるときの報酬 $R(s,a)$ としたとき，Q 値は直接の報酬と今後の期待値の和として定義され，次式で表現される[1].

$$Q(s,A(s)) \leftarrow R(s,A(s)) + \sum \sigma_{sj} \cdot U(j) \tag{3.1}$$

[1] これは *Bellman* 方程式と呼ばれるものである．

ここで $\sum \sigma_{sj} \cdot U(j)$ は，行動 $A(s)$ の期待値を表す．

しかし，マルコフ過程における極限値としての定常確率 σ_{ij} を求めることが容易ではないため，報酬と次状態 s' の効果値を基に Q 値を更新する**ブートストラップ** 手法や対話的探索を用いる必要がある．この標準解は 2 つのパラメタ**学習率** と **割り引き率** と呼ばれる α, γ を用いて次式で表される．

$$Q(s, A(s)) \leftarrow (1-\alpha)Q(s, A(s)) + \alpha(R(s, A(s)) + \gamma Q(s', A(s'))) \tag{3.2}$$

ここで，$\alpha(0 \leq \alpha < 1)$ を学習率といい，当該行動により得られる報酬の優先度を表す．大きい学習率は報酬を重視した更新を表す．また，$\gamma(0 \leq \gamma < 1)$ を割り引き率といい，期待値 (次状態の効果値) の優先度を表す．大きい割り引き率は期待値を重視することを表す．

従って，選択の無い Q 値は更新されず局所最適解となる可能性がある．この状況を回避するため，確率的な行動選択のための Softmax 手法 (ボルツマン分布による確率場を用いる) などが知られる [16]．ここでは，状態 s において 行動 a を選択する条件付き確率 $P(a|s)$ により行動を選択する．この定義は，以下の通りである．

$$P(a|s) = \frac{e^{\frac{Q(s,a)}{T}}}{\sum_{a_i \in A_s} e^{\frac{Q(s,a_i)}{T}}} \tag{3.3}$$

A_s は，状態 s においてとることができる行動 $\{a_1, a_2, ...\}$ の集合である．パラメタ T は**温度**と呼ばれる．温度パラメータ T は行動選択のばらつきを調整するもので，大きい場合には全行動が均一的に選択されることになり，小さい場合には最大の価値を持つルールをより多く選択するようになる．

Q 学習においては単一エージェントのみを仮定していることに注意したい．エージェントは，初期状態と初期 Q 値を与えられ，行動と報酬のみを観測する．マルチエージェント環境では，他エージェントの行動により環境が変化し，各エージェントは互いの観測をしないため，Q 値が収束しない．この結果 Q 値に基づく行動選択が困難となる．いったいマルチエージェント環境に対応した行動選択の手法とな何なのであろうか？**ナッシュ均衡手法**は 1 つの解となりえる戦略であり，本稿の狙いもここにある．

3.3 ナッシュ均衡と Q 学習

3.3.1 ナッシュ均衡

ゲーム理論とは，複数の行動主体が選択する戦略の決定手法であり，互いに自分の報酬（利得）を最大化することを目的としている [16]．ゲーム理論において，行動主体をプレイヤと呼ぶ．本稿では 2 名のプレイヤを仮定する．エージェントが協調せず (互いに交渉すること無く) 同時に戦略選択するゲームは**非協力ゲーム**と呼ばれる．各プレイヤはそれぞれの行動選択で報酬を得る．この状況は **利得行列** という 2×2 の表

で表される．利得行列の各項目 (α, β) at $(i, j), i, j = 1, 2$ において，プレイヤ 1 が行動 i をプレイヤ 2 が行動 j を選択したとき，プレイヤ 1 は報酬 α を得る．同様にプレイヤ 2 は報酬 β を得る．各ステップにおいて，報酬 (報酬の総和の期待値) を最大にするよう行動を決定するものとする．他のプレイヤが戦略を変えない限り，自分が戦略を変える動機を持たない行動選択を**最適反応**という．プレイヤの戦略が**ナッシュ均衡** (Nash equilibrium) であるとは，互いの戦略が相手の戦略に対して最適な反応となるときをいう．表 3.1 で，プレイヤ 2 が戦略 I を選択した場合，プレイヤ 1 が 戦略

	プレイヤ 2		
プレイヤ		I	II
	I	$(3, 3)$	$(5, 2)$
	II	$(2, 5)$	$(4, 4)$

表 3.1: ナッシュ均衡 – 支配戦略

I を選択するほうが大きな報酬 3 を得る．またプレイヤ 2 が戦略 II を選択した場合でもプレイヤ 1 が 戦略 I を選択するほうが大きな報酬 5 を得る．これを，プレイヤ 1 の戦略 I は 戦略 II の対して支配的 (dominant) であるといい，この戦略 I をプレイヤ 1 の支配戦略という．同様にプレイヤ 2 の支配戦略は戦略 I である．つまりプレイヤ 1 ，プレイヤ 2 にとって (I, I) は互いに支配的でありただ 1 つのナッシュ均衡戦略は (I, I) となり，このとき $(3, 3)$ の報酬を得る．この場合，互いが最適な行動を選択するため，プレイヤの得る報酬総和が最適となり，全体として協調動作を獲得することになればよい．ナッシュ均衡に基づく協調動作をとることにより，報酬総和の増加が期待でき，マルチエージェント環境の解を生成する可能性を有する．支配解が存

	プレイヤ 2		
プレイヤ 1		I	II
	I	$(3, 3)$	$(6, 4)$
	II	$(4, 6)$	$(2, 2)$

表 3.2: ナッシュ均衡 – 支配戦略無し

在しないがナッシュ均衡解が存在することがある．表 3.2 では，プレイヤ 2 が戦略 A を選択した場合，プレイヤ 1 が 戦略 II を選択するならば，それ以外の方法より大きな報酬 4 を得る．しかしプレイヤ 2 が戦略 II を選択した場合，プレイヤ 1 は 戦略 I をとれば報酬 6 を得るので， II での報酬 2 より大きい．同様にプレイヤ 1 が 戦略 A を選択するならば，プレイヤ 2 は II においてより大きな報酬 6 を，プレイヤ 1 が II をとればプレイヤ 2 は I においてより大きな報酬 4 を得る．すなわち $(I, II), (II, I)$ が最適である．つまりプレイヤ 1 ，プレイヤ 2 双方にとって支配的な戦略は無いが最適である解 $(I, II), (II, I)$ はナッシュ均衡である．このとき報酬 (4,6), (6,4) を得る．ナッシュ均衡戦略が常に存在するとは限らない．利得行列の表 3.3. で，プレイヤ 1 に

	プレイヤ 2		
		I	II
プレイヤ 1	I	$(3,3)$	$(4,2)$
	II	$(4,2)$	$(3,3)$

表 3.3: ナッシュ均衡無し

とって最適な解は $(I,II),(II,I)$ であるが，プレイヤ 2 の最適解は $(I,I),(II,II)$ である．従って互いに最適となる共通解は存在しない．

さらに複雑な状況として，**囚人のジレンマ** (Prisoner's dilemma) がある．この場合は，支配解であっても必ずしも最大報酬を得るものではない．実際，表 3.1 において (I,I) での報酬 (3,3) よりも (II,II) での報酬 (4,4) のほうが両プレイヤともに大きくなっている．どのプレイヤの報酬も低下すること無く利得を増やすことができない解を**パレート最適** (Pareto optimum) という．利得行列である表 3.3. のように，ナッシュ均衡解は必ずしもパレート最適ではない．また，パレート最適は常に存在する．つまりナッシュ均衡解は最悪解でないことを保証するに過ぎないことに注意したい．

3.3.2 ナッシュ Q 学習

ナッシュ均衡の意味でマルチエージェント環境に対応させた Q 学習を**ナッシュ Q 学習** (Nash-Q learning) と呼ぶ．ナッシュ Q 学習では，ナッシュ均衡を用い行動決定し，ナッシュ均衡により選択した行動の期待値を推測し Q 値の更新に用いる．各エージェントはナッシュ均衡に基づき行動する．また，一定の条件下で Q 値の収束が保証される [3]．ナッシュ Q 学習では，直接互いに協力しないが，ゲーム理論の意味で正しい行動を選択し協調動作をとることが可能となる．

ナッシュ Q 学習において，各エージェントは全てのエージェントの Q 値を観測でき，この結果それぞれの行動と報酬が観測できると仮定する本稿で用いる**人生ゲーム**では，すごろく盤面や選択するカード内容，参加者の行動とその結果を全て知ることができる．シングルエージェント Q 学習と同様に Q 値の更新を繰り返すが，割り引き対象となる次状態の Q 値の利用を，(期待値ではなくて) ナッシュ均衡値を用いる．

状態 s における エージェント i の Q 値を $Q_i(s,a_1,...,a_n)$ で表す．エージェント $1,...,n$ がとる行動 $a_1,...,a_n$ に対し状態 s から s' に遷移するとき，Q 値の更新を次式で定義する．

$$\begin{aligned} Q_i(s,a_1,...,a_n) &\leftarrow Q_i(s,a_1,...,a_n) \\ &\quad +\alpha\{r_i+\gamma NashQ_i(s')-Q_i(s,a_1,...,a_n)\} \\ NashQ_i(s') &= \pi_1(s')...\pi_n(s')Q_i(s') \end{aligned} \tag{3.4}$$

ここで r_i はエージェント i が受けとる報酬，α は学習率，γ は割り引き率である．全てのプレイヤはナッシュ均衡戦略に従い，$NashQ_i(s')$ は，ナッシュ均衡解に基づくエー

ジェント i の状態 s' における期待報酬である．

ナッシュ Q 学習の行動選択手法は，各エージェントが Q 値によるナッシュ均衡の組を選択すると仮定する．しかし，ナッシュ Q 学習においても行動が選択されない Q 値は更新されないため， Q 学習と同様に確率的な行動選択手法を用いる必要がある．Hu は ϵ 貪欲法に基づいて全エージェントがナッシュ均衡戦略に従うとし，確率 ϵ でナッシュ均衡戦略を，確率 $1-\epsilon$ でランダム戦略を選ぶとしている [3]. 本稿も全てのエージェントがナッシュ均衡戦略に従うと仮定し，前述のボルツマン分布に基づく Softmax 手法を用いる．

3.4 ナッシュQ 学習に基づく人生ゲーム戦略

この章では人生ゲームのゲーム理論への適用と，行動の選択原理をナッシュ均衡戦略に従うための枠組みを示す.

3.4.1 人生ゲームのモデル化

本稿で論じる”**人生ゲーム**” (Game of Life) では，2 人のプレイヤが N 個のマス目からなる 1 次元盤面上を進む，すごろくゲームの一種である．両方のプレイヤは，スタート位置の 1 マス目からゴールの N マス目に向かって同時にサイコロを振って進む．途中，マス目のいくつかにマークがあり，各プレイヤの得る報酬の選択や位置変化に従いながら，先にゴールへ到達してゲームが終了する．

マークのあるマス目では，報酬額 V とその総額に関する 2×2 の利得表 M が含まれ，プレイヤは選択を決定する．当該マス目に 1 プレイヤしかいなければ報酬額 V を何もできない．しかし，2 人のプレイヤが同時に同一マス目にいるときは，競合が発生したと考え報酬額 V を分け合う．ここでは，利得行列 M の選択した指示に従い，より大きな報酬を得るよう行動する．この人生ゲームでは，各プレイヤの位置と各プレイヤがどの指示を選択したかにより，得られる報酬が変化する．M は次の形式をとる．

M =

		プレイヤ 2	
		I	II
プレイヤ 1	I	(a,b)	(c,d)
	II	(e,f)	(g,h)

ここで，利得行列 M は報酬額 V をプレイヤ間の配分割合を表す．$a,..,h$ は V の分配割合を表し，$0.0 \leq a,...,h \leq 1.0, a+b=c+d=e+f=g+h=1.0$ である．例えば，プレイヤ 2 が戦略 I を選択したとき，プレイヤ 1 が I を選択すれば a を, II を選択すれば e の報酬を得ることができる．従って実際に得られる報酬は $V\times a$ となる．盤面上の位置やマス目上の指示内容およびプレイヤの選択は他のプレイヤから参照できる．

ゲームでは，プレイヤ $i = 1, 2$ の位置を状態 $s, 1 \leq s \leq N$ とみなし，マス目での指示を行動として，マルチエージェントシステム環境に対応させる．

ゲームの開始時，2 人のプレイヤはマス目 1 にいる状態でサイコロを同時に振る．サイコロの値に従って，プレイヤは移動するが，1 より前にも N より後にも移動できない．N に達すればゴールとみなしその周のゲームを終了する．学習のためこのゲームを繰り返す．学習の間，マス目での指示は変わらないことに注意したい．

本稿で論じる人生ゲームでは，プレイヤ i $(i = 1, 2)$ は，上の規則に従い行動 a_i を決定し (s, a_1, a_2) にて 報酬 r_i を得るだけではなく，もう一方のプレイヤの行動と報酬を観測する．それにより 定義に従い Q 値 $Q_i(s, a_1, a_2)$ を更新する．

3.4.2 人生ゲームでのナッシュ Q 学習のモデル化

本稿で論じる人生ゲームでは，報酬 V をプレイヤで完全に分け合うため，等式 $b = 1 - a, d = 1 - c, f = 1 - e, h = 1 - g$ が成り立つ．

このため，(a, b) が支配解であるためには，$\{a, c, e, g\}$ において c, e がそれぞれ最大値，最小値であることが必要十分である．実際，定義より $a \geq e, c \geq g$ かつ $b \geq d, f \geq h$ であるから $1-a \geq 1-c, 1-e \geq 1-g$ が成り立ち，これを書き換えると $c \geq \{a, g\} \geq e$ となる．同様に $(c, d), (g, h), (e, f)$ が支配解であるためには，a, c, e, g においてそれぞれ $a \cdot g$, $e \cdot c$ and $g \cdot a$ が最大・最小であることが条件となる．

また，支配解でないナッシュ解は存在せず[2]，ナッシュ解は必ずパレート最適となる[3]．

ナッシュ解が存在しなければ，どの項目もパレート最適解となる[4]ため，任意に項目を選択する．

3.1 章で述べたように，Shoham[17] はナッシュ Q 学習についていくつかの欠点を指摘している．しかし，今回の条件により大部分がナッシュ均衡解を持つ．また，ナッシュ均衡が存在しない場合には，パレート最適である (I, I) を選択する．

3.5 実験結果

この章では，実験に用いる人生ゲームについて述べ，ナッシュ Q 学習の動作の検証する．本実験では，人工知能学会 (JSAI) により提案された，*Happy Academic Life 2006:* **研究者の人生ゲーム** を基にしたゲーム [21] を使用する[5]．

[2]存在すれば，$a \geq e, c \leq g, b \geq d, f \leq h$ が同時に成り立つが，これらは矛盾する．

[3]支配解がパレート最適でなければ，$a \leq g, b \leq h$ から $a = g$ を得る．

[4]例えば，(a, b) が Pareto 最適でなく，$(a, b) \leq (c, d)$ とする．このとき $(a, 1-a) \leq (c, 1-c)$，つまり $a \leq c, (1-a) \leq (1-c)$ が成り立つから $a = c$，すなわち $(a, b) = (c, d)$ となって矛盾．

[5]セルラーオートマトンで著名な Life Game ではない．

3.5.1 準備

以下では実験的に作成するゲームの概要を示す.

人生ゲームのスタートからゴールまでの 5 マス ($N = 5$) を 1 周とし，2 人のプレイヤは 1 から 5 を動く．人生ゲーム 1 セットを 10,000 周とし，人生ゲーム毎の報酬のばらつきを抑えるため，人生ゲームを 50 セット行うものとする．全てのマス目には指示を設定する．これは学習を加速するためである．前述の通り，指示は報酬 $V = 100$ と利得行列 M からなり，M の指示を乱数により生成する．M により，報酬 V は 2 のプレイヤに完全分配される．さらにナッシュ均衡解あるいはパレート最適解を予め求めておく．プレイヤ毎に振られたサイコロの目により，移動先と得られる報酬を決定する．ここでのサイコロは $-1, 0, +1$ のいずれか等確率により生成する．比較のため，Q 学習とナッシュ Q 学習の 2 つのケースについて実験する．プレイヤはどちらも Q 学習 (Q) か，どちらもナッシュ Q 学習 (NQ) に従うとする．

プレイヤは，全て同確率の行動選択をし無作為の戦略で行動をとる．本実験では，α, γ, T という 3 つのパラメタが存在する．学習で用いる更新式の係数パラメータは，学習率 $\alpha = 0.01, 0.1, 0.5$，割り引き率 $\gamma = 0.5, 0.9, 0.99$，行動選択で用いる温度パラメータを $T = 1, 10, 100$ とし，Q 学習・ナッシュ Q 学習の各々でテストを行う．従って，合計 $2 \times 3 \times 3 \times 3 = 54$ 回 テストを行う．

ゲームは，総収入 $H = 0$ と 手持ちの収入 $T_1 = 0, T_2 = 0$ でポジション 1 から 2 人のプレイヤがスタートする．一方のプレイヤが N に到達すればその周のゲームが終了し，その際に 2 人の手持ちの収入の差 $|T_1 - T_2|$ を H に加え，$T_1 = 0, T_2 = 0$ にして再度ゲームを行う．本実験では，学習と実験結果を得るため多くの回数を繰り返す．学習中は，指示 M・V は同じである．

Q と NQ の結果を互いにを比較するために，各回の手持ちの収入差の合計を用いる ($\sum_i |T_1^{(i)} - T_2^{(i)}|$)．ナッシュ Q 学習の利得が大きい場合には，協調動作により利得が改善したと考えられるためである．報酬を完全にわけあうことから，協調動作の評価には，エージェントの収入差 H を比較する．協調行動をとっている場合は，両者の収入は拮抗しこの値は小さくなると考えることができる．

3.5.2 実験結果

本稿で扱う実験のパラメタの記述では，$\{Q, NQ\} \times \{T = 1, 10, 100\} \times \{\alpha = 0.01, 0.1, 0.5\} \times \{\gamma = 0.99, 0.9, 0.5\}$ の 54 通りの組み合わせについて実験する．パラメタは横軸に $(Q, 1, 0.01, 0.99), (Q, 1, 0.01, 0.9), ..., (NQ, 100, 0.5, 0.5)$ を表している．

はじめ，サイコロを振る回数について示す．ゲームは 1 周 5 マスからなる．サイコロは $-1, 0, +1$ の目からなり， 平均 16.40 回振る必要がある．2 人のプレイヤが同一のマス目に存在する衝突は，5.85 回起こる．その結果 合計 163907 回学習し，そのうち 58486 回ナッシュ均衡を学習する．

衝突回数 (合計)	3158247/54 = 58486
衝突回数 (各周)	5.85
サイコロを振った回数 (合計)	8850983/54 = 163907
サイコロを振った回数 (各周)	16.40 (35.68%)

表 3.4 および 図 3.1 に収入差の総和を示す．全てのパラメタについて Q 学習による収入差総和の平均は 1,625,928.703 であり，ナッシュ Q の場合の平均値 1,497,067.363 は 7.9% の改善を示している．収入差総計の最大値 67,322,272 は $T = 1, \alpha = 0.5, \gamma = 0.99$ で生じ，ナッシュ均衡による行動原理の影響が強く表れている．他方，収入差総計の最小 -15,604,719 は $T = 10, \alpha = 0.01, \gamma = 0.99$ で生じてはいるがその差が小さく，ナッシュ均衡の影響がここでも強く表れている．

Q 学習による最小の収入差 1,423,773.64 ($T = 100, \alpha = 0.01, \gamma = 0.5$) はナッシュ Q 学習による最小収入差 1,418,064.8 ($T = 100, \alpha = 0.5, \gamma = 0.5$) を 0.4% しか改善していない．しかし標準偏差が Q 学習で 218,132.3 であるのに対しナッシュ Q 学習では 74,434.7 とおおよそ 34% と安定した値を示す．またいずれも $T = 100$ であり，均一的な行動をとろうし行動選択の差が出にくい状況にある．

パラメタ	Q	NQ
$T = 1, \alpha = 0.01, \gamma = 0.99$	1621752.001	1523870.563
$T = 1, \alpha = 0.01, \gamma = 0.9$	1662415.921	1517212.159
$T = 1, \alpha = 0.01, \gamma = 0.5$	1753247.96	1554814.36
$T = 1, \alpha = 0.1, \gamma = 0.99$	1988527.882	1630377.36
$T = 1, \alpha = 0.1, \gamma = 0.9$	2007028.081	1461387.12
$T = 1, \alpha = 0.1, \gamma = 0.5$	1556886.202	1616867.921
$T = 1, \alpha = 0.5, \gamma = 0.99$	2327870.077	1654647.357(28.9%)
$T = 1, \alpha = 0.5, \gamma = 0.9$	1934354.88	1424415.361
$T = 1, \alpha = 0.5, \gamma = 0.5$	1739201.08	1507142.601
$T = 10, \alpha = 0.01, \gamma = 0.99$	1470772.922	1626820.12
$T = 10, \alpha = 0.01, \gamma = 0.9$	1535244.161	1554105.398
$T = 10, \alpha = 0.01, \gamma = 0.5$	1604466.162	1572337.843
$T = 10, \alpha = 0.1, \gamma = 0.99$	1627458.842	1490574.921
$T = 10, \alpha = 0.1, \gamma = 0.9$	1587668.398	1489956.958
$T = 10, \alpha = 0.1, \gamma = 0.5$	1619084.044	1513985.362
$T = 10, \alpha = 0.5, \gamma = 0.99$	1606998.359	1488714.041
$T = 10, \alpha = 0.5, \gamma = 0.9$	1629115.682	1460924.44
$T = 10, \alpha = 0.5, \gamma = 0.5$	1628046.881	1516351.241
$T = 100, \alpha = 0.01, \gamma = 0.99$	1425705.2	1427943.441
$T = 100, \alpha = 0.01, \gamma = 0.9$	1423846.282	1428644.561
$T = 100, \alpha = 0.01, \gamma = 0.5$	1423773.64	1427864.038
$T = 100, \alpha = 0.1, \gamma = 0.99$	1436287.521	1422137.919
$T = 100, \alpha = 0.1, \gamma = 0.9$	1435243.881	1422395.678
$T = 100, \alpha = 0.1, \gamma = 0.5$	1432339.959	1424950.718
$T = 100, \alpha = 0.5, \gamma = 0.99$	1511830.799	1421701.117
$T = 100, \alpha = 0.5, \gamma = 0.9$	1458998.519	1418064.8
$T = 100, \alpha = 0.5, \gamma = 0.5$	1451909.639	1422611.401
平均	1625928.703	1497067.363(7.9%)
分散	218132.3345	74434.74902

表 3.4: パラメタによる収入差の総和

図 3.2 に $T = 1, \alpha = 0.1, \gamma = 0.9$ での Q 学習におけるゲーム，および $T = 10, \alpha = 0.1, \gamma = 0.9$ におけるナッシュ Q 学習のゲームの (10,000 回までの) 各回毎の収入差を示す．このパラメタ組は 10,000 周目で最も収入の差が小さかった 2 つである．Q 学習

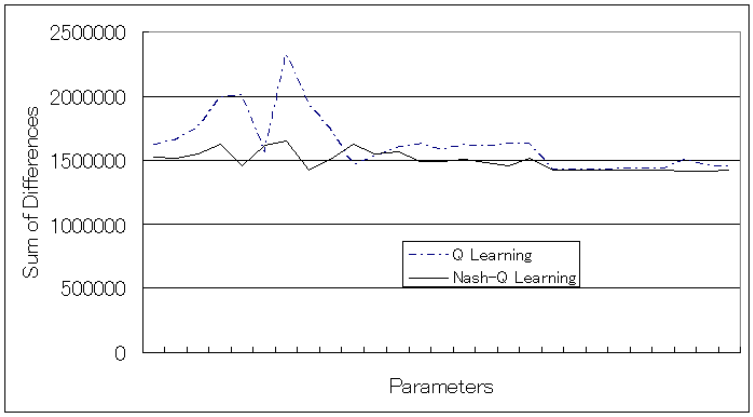

図 3.1: パラメタによる収入差の総和

での平均が 162.18，標準偏差が 23.74 であり，ナッシュ Q 学習での平均は 149.00 (8.8% の改善)，標準偏差が 20.97 となる．いずれも極めて安定的に ナッシュQ 学習での収入差を生成している

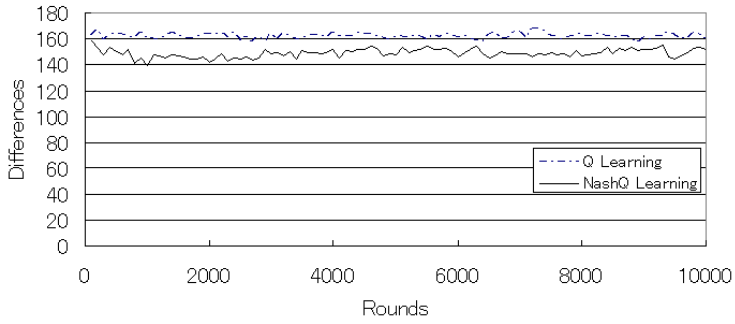

図 3.2: 各周における報酬差

最後に，学習した Q 値を直接比較する．2 つの Q 値を比較するのは困難であるため，多次元情報 N の **フロベニウスノルム** (Frobenius Norm) $||N||$ を定義する．$x_k = 1, .., n_k$ の 2 つの多次元データ $N(x_1, .., x_k)$ のフロベニウスノルム $||N||$ を以下に示す．

$$||N|| = \sqrt{\sum_{x_1=1,..,n_1,..,x_k=1,..,n_k} N(x_1, .., x_k)^2}$$

互いのプレイヤの Q 値が類似している場合には，ボルツマン分布に基づく 2 つの確率も近い値となるべきである．

本稿では Q 値 $Q_i(s, a_1, a_2), s = (1, 1), , (5.5), a_1, a_2 = I, II$ を Q 学習とナッシュ Q 学習でそれぞれ求め，$||Q_1 - Q_2||$ を求めて比較する．実験結果を 表 3.5 に示す．Q 学習におけるノルムはナッシュ Q 学習のそれとほぼ同じであるが，標準偏差が 7.5 倍程

度である．すなわち，ナッシュ Q 学習では Q 値は極めて安定的に求められていることがわかる．

パラメタ	Q	NQ
$T = 1, \alpha = 0.01, \gamma = 0.99$	448.6297137	218.8997415
$T = 1, \alpha = 0.01, \gamma = 0.9$	171.9696808	188.0851988
$T = 1, \alpha = 0.01, \gamma = 0.5$	28.84698752	179.3153638
$T = 1, \alpha = 0.1, \gamma = 0.99$	680.7310435	140.5620884
$T = 1, \alpha = 0.1, \gamma = 0.9$	216.4803873	205.2043583
$T = 1, \alpha = 0.1, \gamma = 0.5$	83.46349321	161.5906338
$T = 1, \alpha = 0.5, \gamma = 0.99$	417.5393315	195.5513214
$T = 1, \alpha = 0.5, \gamma = 0.9$	141.7941112	233.8467492
$T = 1, \alpha = 0.5, \gamma = 0.5$	94.05508483	188.4624046
$T = 10, \alpha = 0.01, \gamma = 0.99$	640.1305848	206.2215533
$T = 10, \alpha = 0.01, \gamma = 0.9$	179.6590745	212.8367659
$T = 10, \alpha = 0.01, \gamma = 0.5$	21.36131886	185.0914212
$T = 10, \alpha = 0.1, \gamma = 0.99$	593.764268	249.8484393
$T = 10, \alpha = 0.1, \gamma = 0.9$	215.1351854	222.4487387
$T = 10, \alpha = 0.1, \gamma = 0.5$	28.12856534	215.7609382
$T = 10, \alpha = 0.5, \gamma = 0.99$	294.313464	242.2002577
$T = 10, \alpha = 0.5, \gamma = 0.9$	109.4168758	224.9832451
$T = 10, \alpha = 0.5, \gamma = 0.5$	43.79241205	225.889786
$T = 100, \alpha = 0.01, \gamma = 0.99$	87.43635453	206.5370723
$T = 100, \alpha = 0.01, \gamma = 0.9$	59.06821519	205.9177496
$T = 100, \alpha = 0.01, \gamma = 0.5$	24.51310235	203.8776401
$T = 100, \alpha = 0.1, \gamma = 0.99$	109.325061	212.3167093
$T = 100, \alpha = 0.1, \gamma = 0.9$	67.13858751	217.1518182
$T = 100, \alpha = 0.1, \gamma = 0.5$	27.00344437	205.482255
$T = 100, \alpha = 0.5, \gamma = 0.99$	272.7306777	273.6228919
$T = 100, \alpha = 0.5, \gamma = 0.9$	149.2752252	204.9374215
$T = 100, \alpha = 0.5, \gamma = 0.5$	59.25212384	219.7724645
平均	194.9983102	209.1264825
分散	195.8486491	26.42443818

表 3.5: Q 学習 とナッシュ Q 学習の Q 値

3.5.3 考察

本実験に関する限り， Q 学習とナッシュ Q 学習の収入差の総和には 7.9% 程度の優位性はあるものの，極端な差はみられない．むしろ表 3.4 より，最大の収入差が $T = 1$ でみられるのはナッシュ均衡原理による行動選択に起因することのほうが興味深い．

図 3.2 から収入差を各回毎に調べると，最終的な収入差が最小のケースでさえも，平均して 149.00 (NQ), 162.18 (Q) と 8.8% 程度の優位性を保ったままに推移する．

これらの事実から，ナッシュQ 学習の効果は，収入差の極端な均衡化ではなくて，結果の安定化にあることがわかる．実際，表 3.5 からパラメタ設定によらずに安定した Q 値を学習していることがわかる．ナッシュ Q 学習は，パラメタ設定によらず安定的な Q 値を学習し，ばらつきの少ない協調的な結果を生成する戦略として有効である．

3.6 結論

本稿では， Q 学習とナッシュ Q 学習に基づく戦略を比較した．非零和確率ゲームにより実験において，ナッシュ Q 学習に対し Q 学習はプレイヤの Q 値の差が 6 倍よりも大きくなることを示した．従って，ナッシュ Q 学習は学習率・温度（行動選択手法ボルツマン分布）によらず安定した学習結果を示すことが確認できる．

今回の実験では，どのプレイヤも対等に振る舞う状況で，収入合計が Q 学習と比べ協調的に振る舞うことを確認した．今後は多数のプレイヤや行動の選択肢が多い場合に対して有用性を確認する必要がある．本実験では報酬を完全に分け合うという仮定を置いたため，ナッシュ均衡は存在すれば一意であり，ナッシュ均衡解が存在しなければどの選択もパレート最適性となる．また互いの情報は全て既知とした．しかし，一般にはこの仮定は限定的であり，ナッシュ均衡自体の問題にさかのぼった考察が必要である．

第4章　局所影響ゲームにおける相関均衡を用いた行動選択

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の有効性を検証する．行動選択手法としては，ナッシュ均衡を用いた手法が提案されているが，ナッシュ均衡とパレート最適が必ずしも一致しない，ナッシュ均衡が一意に定まらず確率的に行動選択をする必要が出てくる，という問題がある．本稿では，相関手法を行動選択手法に用い，局所影響ゲームに適用することで，本手法の有効性を検討する．

4.1　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．この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各エージェントには報酬が与えられる．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を獲得することが期待される．

本稿では，エージェントがどのようにして行動選択をするかに注目する．行動選択の際には，自分の行動だけではなく，他のエージェントの行動により自分の得る報酬が大きく変化する．そのため，自身だけの評価を考え行動しても報酬を大きくすることは難しい．互いの報酬を大きくするには，協調行動が必要だが，その選択は困難である．従って，行動選択の際にはマルチエージェントに対応した行動選択が必要である．

Hu らは，行動選択手法にゲーム理論のナッシュ均衡を用いたナッシュQ 学習を提案している [3]．ナッシュQ 学習は，本来シングルエージェントの学習手法である Q 学習を行動選択戦略・Q 値の更新にナッシュ均衡の利用することでマルチエージェントに対応させたものである．しかし，一方 Shoham はナッシュ Q 学習を批判し，ナッシュ均衡が一意解でないときの対応が明確ではないなどの問題を指摘している [17]．そこで著者らは，ナッシュ均衡が 1 つのみ存在するよう報酬を制約した，報酬完全分配法を提案している [5]．

ここでは，行動選択手法として，ゲーム理論の相関手法を考える．相関手法は，他のエージェントとの通信路を有し，共通の行動の選択機構により行動を決定する．そのため，エージェントとの協調行動をとることができる．本稿では，その有効性を実験にて確認することを目的とする．

Kakade らは相関戦略を Graphical Games での適用を検討し，Graphical Games の制約に依存することで，多項式時間で相関均衡を導出する手法を提案している [10].

第 2 章では，マルチエージェントシステムでの行動決定に関して述べ．第 3 章では，相関手法における行動選択について提案する．第 4 章では，本稿で用いる局所影響ゲームについて述べ，第 5 章では，実験手法および実験結果を示し，第 6 章で，本稿をまとめる．

4.2 マルチエージェントシステムでの行動決定

マルチエージェントシステムでは，自分の報酬の決定に，自分の行動のみならず他のエージェントの行動との組み合わせにより決定される．そのため，より大きなを報酬得るためには，マルチエージェントに対応した行動選択，協調行動をとることが必要である．

そこで，行動選択手法をゲーム理論に対応させて考える．ゲーム理論とは，複数の行動主体が選択する戦略の決定手法であり，互いに自分の報酬を最大化することを目的としている．本稿では 2 名のエージェントを仮定する．互いに交渉 (協力) すること無く同時に戦略選択するゲームは非協力同時的であると呼ばれる．

非協力 (同時) ゲームで行動 I, II のいずれかを選択するとき，各エージェントがそれぞれの行動で報酬を得る．報酬は (競合することから) 組み合わせによって異なり，これを 2×2 の表 (利得行列という) で表す．各項目 (α, β) はそれぞれ エージェント $1, 2$ が行動 I, II を選択したときに得る報酬を表す.

エージェントの行動が**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な反応となるときをいう．本稿でいう”反応”は報酬を意味し多いほど良い．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最適であるという．各エージェントは，自分の報酬を最大化するような合理的な行動を選択し，相手の行動に対して自分の報酬が最大となるような最適な反応をとる．

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

表 4.1: チキンゲーム

表 4.1. 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 を選択するほうが大きな報酬 5 が得られる．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 を選択するほうが大きな報酬 1 が得られる．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) と (II, I) との 2 つ存在する．

また，確率的に行動を選択することを許す混合戦略を考えた場合，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 とし，それぞれ確率的に行動するものとする．この場合，エージェント A の報酬の期待値 $f(p,q)$ は次の式で表すことができる．

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

これは p に関する関数とみて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．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な行動を選択する．各エージェントがどのような確率で行動を選択すべきかを示す，最適反応は，図 4.1. のように表される．ナッシュ均衡は，互いの行動が最適反応である行動の組み合わせであるため，混合戦略まで考えたナッシュ均衡は図 4.1. の交点である，$(I, II), (II, I), \left((\frac{1}{2}I + \frac{1}{2}II), (\frac{1}{2}I + \frac{1}{2}II)\right)$ の 3 つである．

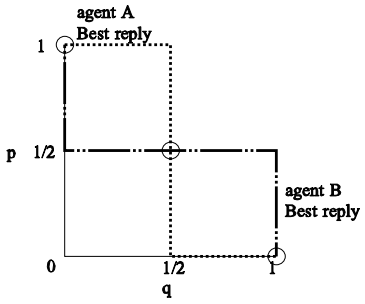

図 4.1: 各エージェントの最適反応

しかし，ナッシュ均衡が複数存在する場合には，どの行動を選べばいいかという方針が存在しない．ナッシュ均衡の前提として，エージェント間の通信路をを持たず，各々のエージェントが，混合戦略に従い行動 I と行動 II を確率 $\frac{1}{2}$ ずつで選択としたしても，確率 $\frac{1}{4}$ で非パレート最適解 $(0,0)$ となり，確率的に選択するのも問題がある．従って，ナッシュ均衡が複数存在する場合や混合戦略のナッシュ均衡が存在する場合は行動を選択手段が無い．また，ナッシュ均衡は誰もが損すること無く，報酬をあげることができるパレート最適を保証せず，最多報酬を得られるとは限らない．

そのため，ナッシュ均衡は行動選択の 1 つの指針として有効であるが，ナッシュ均衡のみにより協調行動の選択を行うことは難しい．

4.3 相関戦略

ここで，Aumann らにより提案された相関手法を考える [1]．相関手法では，ナッシュ均衡と同様な合理的な行動選択を前提とする．エージェント間は通信路を有し，エージェント間の相互作用を示す「共通の情報」を用い行動を選択する．この場合の報酬は，共通の情報を用いて決定される．

合理的な行動選択は，式 4.1．の様に定義される．

$$\sum_{s \in S_{-p}} u^p_{is} - u^p_{js} \geq 0 \tag{4.1}$$

ある空間に n エージェントが存在するとき，それぞれのエージェント p は行動の集合 S_p を持ち，ある行動を $s_p \in S_p$ と表す．エージェント p を除いた行動の集合を S_{-p} と表す．報酬 u_{is} はエージェント p の行動 i とエージェント p 以外の行動の組み合わせ $s = (s_1, ..., s_{p-1}, s_{p+1}, ..., s_n)$ で決定される．

式 4.1．では，エージェント p の行動 i は他の行動 j 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

表 4.1．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 1 は，行動 II を選択するほうが大きな報酬 5 が得られる．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 を選択するほうが大きな報酬 1 が得られる．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，相関均衡は (I, II) と (II, I) との 2 つ存在する．これは，ナッシュ均衡と一致し，逸脱する動機も持たない．

ナッシュ均衡との違いは，共通の情報を持つことにより，互いに協調する仕組みを持っている点である．このような仕組みを持つゲームを (非協力ゲームに対して)，協力ゲームと呼ぶ．共通の情報を持つことで，均衡を決定する関数が導出できる．この関数を用いることにより，エージェントが協調行動を獲得することを狙っている．

本稿ではこの相関手法を用いることで協調行動が獲得できるかどうかについて述べる．

4.4 ゲームモデル

4.4.1 混雑ゲーム

マルチエージェント環境での問題として，Rosenthal によって提案された，混雑ゲーム (Conestion Games) がある [15].

混雑ゲームとは，混雑すると 1 人当たりのコストが増大するという，ネットワークの混雑や施設の混雑などをモデル化したものである．ここでのコストとは，マルチエージェントに対応させた場合には，負の報酬に対応される．

混雑ゲーム G は，エージェント数 n，コスト $Cost$，要因 k で定義され，エージェント $i(i = 1...n)$ のコスト $Cost_i(a_1, ..., a_n)$ は式 4.2．のように表される．

$$Cost_i(a_1, ..., a_n) = \sum_{k \in a_i} F_k(D_k(a_1, ..., a_n)) \tag{4.2}$$

このとき，$a_i(a_i = 1...s_i)$ は選択可能なエージェント i の行動を表し，D_k は要因 $k(k = 1...t)$ を選択したエージェント数である，F_k は要因 k をとったときの 1 人当たりのコストを表す．

従って，コスト $Cost_i(a_1, ..., a_n)$ は，混雑ゲームは全エージェントの行動により決定され，混雑によりでコストが増えることを意味している．

このとき，エージェント間の相互作用が問題となる．このモデルの場合，全エージェントとの相互作用を計算する必要があり，エージェント数が多くなるとパラメタ数も増え，計算が膨大となる問題がある．

4.4.2 局所影響ゲーム

本稿では，Leyton-Brown らによって提案された局所影響ゲーム (Local Effect Games) を用いる [9]．局所影響ゲームは，混雑ゲームに条件を加え，直接影響を与えるエージェントの相互作用のみを考える．そのため，コスト計算に用いるパラメタ数を減少させることができ，混雑ゲームのような問題を避けることができる．

n エージェントの 局所影響ゲーム G は，$G =< A, F, n >$ で定義される．本ゲームは，各エージェントが獲得したコスト $(Cost)$ と呼ばれる負の報酬で評価される．

A は各エージェントが選択可能な行動集合を表す．$D(i)$ は，行動 i を選択しているエージェント数を表し，$F_{j,i}(D(i))$ は，自分が行動 i，他のエージェントが行動 j をとったときの，1 人当たりのコストを表している．つまり，コストは行動 j を選択しているエージェント数により変化する．

ある行動 i をとったときのコスト $Cost(i)$ は，”自分自身の行動”と”自分と他のエージェントとの行動の組み合わせ”で定義され，以下のよう表される．

$$Cost(i) = F_{i,i}(D(i)) + \sum_{j \in A, j \neq i} F_{j,i}(D(j)), \tag{4.3}$$

$(\forall a, a' \neq a \in A \ : \ F_{a,a'}(0) = 0)$

ここで，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ことに注意したい．

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

表 4.2: 局所影響関数 (局所影響ゲーム)

ここで，表 4.2. という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 をとった場合，

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

また，エージェント A が行動 I，エージェント B が行動 I をとった場合，

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

となる．コストは負の報酬を意味し，その結果得られる報酬は表 4.3. の利得行列で表す通りである．

	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 4.3: 局所影響ゲームの利得行列

4.4.3 双方向局所影響ゲーム

本実験では単純化のため，局所影響ゲームの 1 つ，双方向局所影響ゲーム (Bidirectional Local-Effect Games : B-LEG) を用いる [9]．B-LEG では，上記の条件に”相手に与える影響”と”自分が受けとる影響” が等しいく，エージェント数に関して線形関数であるという制約が加わる．この制約の結果，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，Leyton-Brown らにより示されている．[1]

これは，

$$F_{a,a'} = F_{a',a} \ (\forall a \in A, \forall a' \neq a \in A) \tag{4.4}$$

という式で定義される．

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

表 4.4: 局所影響関数 (双方向局所影響ゲーム)

ここで，表 4.4. という Local Effect 関数がが与えられたとき，エージェント A が行動 I，エージェント B が行動 I をとった場合，

$$Cost = F_{I,I}(1) + F_{II,I}(0) = 200$$

また，エージェント A が行動 I，エージェント B が行動 I をとった場合，

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

	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 4.5: 双方向局所影響ゲームの利得行列

となる．その結果得られる報酬は表 4.5．の利得行列で表す通りである．

この B-LEG で，相関戦略を考える場合の共通の情報を，本稿では，局所影響関数を共通の情報として扱う．

協力ゲームの場合は，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から行動を選択する．共通の情報を含む利得行列は表 4.5．の通りであるから，表 4.5．に従い混合戦略ナッシュ均衡により，行動を選択する．今回の場合，混合戦略に従った場合に得られる報酬の期待値は $(-187.5, -187.5)$ となる．局所影響ゲームでは共通の情報を持つため，非協力ゲームの場合よりも，協調行動を獲得しやすいことが期待できる．

4.4.4 ナッシュ均衡が複数存在する条件

行動選択にナッシュ均衡を利用する際の問題の 1 つが，均衡が複数存在する可能性があり確率的な行動選択をする必要がある点である．ここでは，B-LEG でナッシュ均衡が複数存在し．行動の選択肢が一意に決定できない場合を考える．

	エージェント B		
エージェント A		I	II
	I	$(2F_{I,I}, 2F_{I,I})$	$(F_{I,I} + F_{I,II},$ $F_{II,II} + F_{I,II})$
	II	$(F_{II,II} + F_{I,II},$ $F_{I,I} + F_{I,II})$	$(2F_{II,II}, 2F_{II,II})$

表 4.6: 局所影響関数での利得行列表示

B-LEG での各エージェントが獲得する報酬は，表 4.6．の利得行列で表される．この場合，複数の純粋戦略ナッシュ均衡が存在する場合は，$(I, I), (II, II)$ と $(I, II), (II, I)$ の 2 通りの組み合わせが考えられる．

[1] 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必ず存在するとは限らない．しかし，表 4.3．のようなエージェント数が $n = 2$ の場合は，対称ゲームとなり純粋戦略ナッシュ均衡が必ず存在する．

$(I,I),(II,II)$ が純粋戦略ナッシュ均衡である場合,

$$\begin{cases} 2F_{I,I} > F_{II,II} + F_{I,II} \\ 2F_{II,II} > F_{I,I} + F_{I,II} \end{cases} \tag{4.5}$$

(i) もし,$F_{I,I} < F_{II,II}$場合

$$\begin{cases} 2F_{I,I} > F_{II,II} + F_{I,II} > F_{I,I} + F_{I,II} \\ 2F_{II,II} > F_{I,I} + F_{I,II} \end{cases} \tag{4.6}$$

$$\Rightarrow \begin{cases} F_{I,I} > F_{I,II} \\ 2F_{II,II} > F_{I,I} + F_{I,II} \end{cases} \tag{4.7}$$

$$\Rightarrow \begin{cases} F_{I,I} > F_{I,II} \\ 2F_{II,II} > F_{I,I} + F_{I,II} > 2F_{I,II} \end{cases} \tag{4.8}$$

$$\therefore \begin{cases} F_{I,I} > F_{I,II} \\ F_{II,II} > F_{I,II} \end{cases} \tag{4.9}$$

(ii) もし,$F_{I,I} > F_{II,II}$場合

$$\begin{cases} 2F_{I,I} > F_{II,II} + F_{I,II} \\ 2F_{II,II} > F_{I,I} + F_{I,II} > F_{II,II} + F_{I,II} \end{cases} \tag{4.10}$$

$$\Rightarrow \begin{cases} 2F_{I,I} > F_{II,II} + F_{I,II} \\ F_{II,II} > F_{I,II} \end{cases} \tag{4.11}$$

$$\Rightarrow \begin{cases} 2F_{I,I} > F_{II,II} + F_{I,II} > 2F_{I,II} \\ F_{II,II} > F_{I,II} \end{cases} \tag{4.12}$$

$$\therefore \begin{cases} F_{I,I} > F_{I,II} \\ F_{II,II} > F_{I,II} \end{cases} \tag{4.13}$$

式 4.9. 式 4.13. より，純粋戦略ナッシュ均衡が複数存在する条件は以下の通り となる.

$$\begin{cases} F_{I,I} > F_{I,II} \\ F_{II,II} > F_{I,II} \end{cases} \tag{4.14}$$

$(I,II),(II,I)$ の場合も同様に導出される．これら結果をまとめると以下の通りである.

$$\begin{cases} F_{I,II} > F_{I,I} \\ F_{I,II} > F_{II,II} \end{cases}$$
$$\text{もしくは} \tag{4.15}$$
$$\begin{cases} F_{I,I} > F_{I,II} \\ F_{II,II} > F_{I,II} \end{cases}$$

従って，式 4.15. の条件を満たしているパラメタにおいて，協調行動が獲得しているときには，B-LEG において確率的な行動選択を行う場合にも，協調行動が獲得可能であることがわかる．

4.5 実験

ここでは，上記で示した B-LEG を協力ゲームとして，比較のため一様分布に従い利得行列を生成したものを（互いの情報を共有しないため）非協力ゲームとして実験を行う．本実験では，エージェント数を $A \cdot B$ の $n = 2$ とし，選択可能な行動数を $I \cdot II$ の 2 通り とする．協力ゲーム · 非協力ゲームともに 100 種類のパラメタを用意する．それぞれの種類を 1 ゲームとし，1 ゲームを 10 回繰り返すことで合計 100×10 回を実験し，それぞれ 1 ゲームで獲得した報酬を記録する．

本実験の協力ゲーム (B-LEG) は，多くのパターンのパラメタで検証するために，以下の表 4.7. の様に 0 ∼ 10 の間で一様分布に従い生成する．この生成したそれぞれ 100 種類のパラメタを局所影響関数として B-LEG で用いる．

非協力ゲームでは，互いの行動の組み合わせにより得られる報酬が決定される．得られる報酬は，一様分布に従い以下の表 4.8. の様に 0 ∼ 10 の間で生成する．

2 エージェントの戦略は，互いに同一の行動決定手法を用いる．協力ゲームとして相関手法で行動決定を行うものと，非協力でナッシュ均衡を用いて行動決定をする 2 種類で比較する．

	$F_{I,I}$	$F_{I,II}$ $=F_{II,I}$	$F_{II,I}$	$F_{II,II}$
0	2.79	0.64	0.64	4.72
1	2.25	9.74	9.74	5.96
2	0.37	2.39	2.39	8.09
3	7.71	9.19	9.19	9
4	6.08	3.82	3.82	3.63
5	2.91	0.67	0.67	6.77
6	7.52	2.75	2.75	9.05
7	6.91	6.07	6.07	3.62
8	2.72	2.74	2.74	5.46
9	1.82	1.49	1.49	4.82
...	...	...	...	...
分散	6.84	8.03	8.49	6.94
平均	4.87	4.97	4.97	5.45

表 4.7: B-LEG(協力ゲーム) に用いるパラメタ 10 件抜粋

	エージェント A の報酬				エージェント B の報酬			
行動	(I,I)	(I,II)	(II,I)	(II,II)	(I,I)	(I,II)	(II,I)	(II,II)
0	0.32	5.39	3.18	6.89	0.31	6.96	7.72	4.78
1	2.91	2.78	6.78	9.66	1.51	0.07	3.54	7.55
2	0.47	9.12	2.1	1.75	8.09	4.12	0.71	5.08
3	6.41	1.64	6.71	0.69	8.25	6.72	7.07	1.23
4	8.55	3.79	6.01	0.21	8.93	4.04	9.61	7.1
5	1.2	3.5	5.83	7.58	8.39	6.46	7.01	3.65
6	9.26	6.92	1.31	3.77	5.63	6.71	8.85	2.02
7	9.84	2.64	8.2	6.74	3.79	6.84	8.81	3.03
8	9.63	3.39	5.55	0.84	6.69	2.16	2.09	0.18
9	8.73	5.06	0.96	7.46	0.66	3.41	9.46	2.6
...	...	...	...	...	...	...	...	...
平均	5.46	4.81	4.54	4.61	4.92	5.18	5.17	5.10
分散	8.16	7.86	6.83	9.08	7.83	8.36	9.32	8.32

表 4.8: 非協力ゲームに用いるパラメタ 10 件抜粋

評価として，行動選択の戦略が協力ゲーム（相関手法）の場合と非協力ゲーム（ナッシュ均衡）の場合とを比較する．

1 ゲーム を 10 回繰り返すとき報酬の平均が多いエージェントを”勝ち”と，互いのエージェントがともに同じ報酬の場合は”引き分け”と定義する．

今回の実験は，一方のエージェントのみが多くの報酬を得るということが無く，互いエージェントが協調行動を獲得できるかを確認する．従って，協調行動により”引き分け”数が多くなることが期待できる．

表 4.9．に非協力ゲームと協力ゲームにおける エージェント A の勝敗数を，表 4.10．には，”引き分け”の判定を エージェント A の報酬の $\pm 5\%$ まで広げた場合の エージェント A の勝敗数を示す．

	引き分け	勝ち	負け
非協力ゲーム	0	48	52
協力ゲーム	73	13	14

表 4.9: エージェント A の勝ち負けの結果

	引き分け	勝ち	負け
非協力ゲーム	8	44	48
協力ゲーム	96	2	2

表 4.10: エージェント A の勝ち負けの結果 (報酬差 $\pm 5\%$ 以内も引き分けとした場合)

表 4.9．から，非協力ゲームの”引き分け”数は 0 であるのに対し，協力ゲームの”引き分け”数は 73 であり，7 割 以上のゲームに関して協調行動が獲得できている．

また，表 4.10．から，少しの報酬の差を許し”引き分け”を エージェント A の報酬の $\pm 5\%$ まで広げた場合では，非協力ゲームでの”引き分け”は 8 であり，ほとんど協調行動

が獲得できていない．それに対し，協力ゲームでは表4.9．よりも増え96であり，96%とというほぼ全てのゲームが協調行動を獲得できていることがわかる．

実験結果の表4.10．において，協力ゲームで協調行動が獲得できていないと判断されたゲームについて考える．そのときのパラメタは以下表4.11．に示す．また，このときの利得行列を表4.12．に，10回繰り返した実験結果を表4.13．に示す． 表4.11．より，

$F_{I,I}(x)$	$4.61x$
$F_{II,I}(x) = F_{I,II}(x)$	$8.38x$
$F_{II,II}(x)$	$2.68x$

表 4.11: 協調行動が獲得できていないゲームの Local Effect 関数

		エージェント B	
		I	II
エージェント A	I	$(9.22, 9.22)$	$(12.99, 11.06)$
	II	$(11.06, 12.99)$	$(5.36, 5.36)$

表 4.12: 協調行動が獲得できていないゲームの利得行列

	エージェント A	エージェント B
0	12.99	11.06
1	11.06	12.99
2	9.22	9.22
3	11.06	12.99
4	9.22	9.22
5	9.22	9.22
6	11.06	12.99
7	11.06	12.99
8	9.22	9.22
9	9.22	9.22
平均	10.33	10.91
差	5.60% (エージェント B の勝ち)	

表 4.13: 10 回繰り返し実験結果

このゲームでの混合戦略ナッシュ均衡を計算すると，$((0.81I, 0.19II), (0.81I, 0.19II))$ であり，期待値 $(9.95, 9.95)$ を得る．しかし，純粋戦略ナッシュ均衡の $(I, II), (II, I)$ の報酬のほうが大きく，混合戦略ナッシュ均衡は非パレート最適である．

ここで，表4.13．の平均報酬で比較すると，エージェント B が報酬 10.91 を獲得し，”勝ち”と判定している．これは混合戦略ナッシュ均衡により，行動が確率的に決定され，行動が一意に決定できないことから，得られる報酬が安定していないためである．

ただし，エージェント A に対して エージェント B の報酬が高々5.60% 多いことからでは，一方のエージェントのみが大きな報酬を得ているとはいえず，相関手法の効果により互いの報酬が近くなっている．

上の例のように，ナッシュ均衡が複数存在し行動が一意に決定できないものについて考える．この条件式 4.15. に当てはまる本実験のパラメタ（表 4.7.）は，73% 存在する．ただし，今回の実験では全体の 73% （”引き分け”を報酬の ±5% まで広げた場合では全体の 96%）が協調行動を獲得している．このことから，確率的行動選択をした場合でも協調行動を獲得することが可能であることがわかる．

4.6 結論

本稿では双方向局所影響ゲームにおいて，ある条件下において，行動選択戦略にとして相関手法を用いることにより，協調行動をとることが可能であることを示した．従って，相関手法を行動選択戦略として用いることは有効である．特に本実験では，相関手法において，複数のナッシュ均衡が存在し確率的に行動決定した場合でも，協調行動が獲得できていることを確認した．

本実験では，エージェント数および行動数を 2 に制限している．従って，エージェント数，行動数が 3 以上に場合に関しても検討する必要がある．

また，本稿で用いた双方向局所影響ゲームにも，各エージェントが選択可能な行動が共通であるなど，多くの制約がありのこの制約を緩和した場合や，他のモデル，一般的なゲームに適用した場合に関してさらに考察が必要である．

第5章　繰り返し局所影響関数を用いた行動選択

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．

5.1　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 [3]．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．[17]．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 [5]．

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用する．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 学習を適用するかを提案する．第 5 章では実験結果を示し，第 6 章で本稿をまとめる．

5.2 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，個別およ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 5.1. 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 をとることで 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 5.1. 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		
エージェント A		I	II
	I	$(4, 4)$	$(1, 5)$
	II	$(5, 1)$	$(0, 0)$

表 5.1: チキンゲーム

表 5.1. 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わけではなく，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 5.1. 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 $f(p,q)$ および，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 5.1. 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．

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．

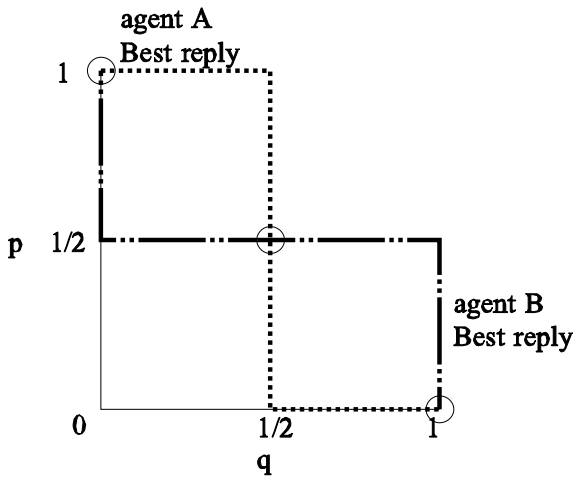

図 5.1: 各エージェントの最適反応

5.3 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 5.1．の様に定義する．

$$\sum_{s \in S_{-i}} u^i_{as} - u^i_{a's} \geq 0 \tag{5.1}$$

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 5.1．) 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 5.1．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．

5.4 局所影響ゲームモデル

5.4.1 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 [15]．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 5.2. のように表される．

$$r_i(a_1, ..., a_n) = \sum_{k \in a_i} F_k(D_k(a_1, ..., a_n)) \tag{5.2}$$

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．

5.4.2 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 [9]．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 をとるエージェント $D(x)$ 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．

$$r_i(a', a) = \sum_{y \in \mathcal{B}^{n-1}, x \in y} F_{y,a}(D(x)) \tag{5.3}$$

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 5.2: 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5.2. 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 をとったときの 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 をとったのときの 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5.3. で示す．

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

表 5.3: 局所影響ゲームの利得行列

5.4.3 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 [9]．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”自分が受けとる影響”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 [9] 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 $a' =< b_1, ..., b_{n-1} > \in \mathcal{A}^{n-1}$ をとったとき，B-LEG における局所影響関数を以下の式で定義する．

$$F_{a',a} = F_{a''} \quad a'' \in \mathcal{B}^n, \text{ i.e., } a'' \text{ は } \mathcal{B} \text{ を置換したもの} \tag{5.4}$$

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

表 5.4: 局所影響関数 (B-LEG)

例えば，表5.4. 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 をとるとき，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 をとるとき，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5.5. に示す．

		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 5.5: 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5.5. 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．

5.5 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，各エージェントがと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 をと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 5.5 ．に従い Q 値を更新する．

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a)) \tag{5.5}$$

状態 s で行動 a をとるとき， 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．$\alpha(0 \leq \alpha < 1)$ 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，$\gamma(0 \leq \gamma \leq 1)$ を**割り引き率** 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 [19].

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．[3]．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 5.6 により Q 値 $Q_i(s, a_1, ..., a_n), i = 1, ..., n$ を更新する．

$$\begin{aligned} Q_i(s,a_1,...,a_n) &\leftarrow Q_i(s,a_1,...,a_n) \\ &\quad +\alpha\{r_i + \gamma NashQ_i(s') - Q_i(s,a_1,...,a_n)\} \\ NashQ_i(s') &= \pi_1(s')...\pi_n(s')Q_i(s') \end{aligned} \tag{5.6}$$

π_i は，エージェント i のと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 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 5.6: Q 値行列

学習の結果，表5.6. 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[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 [17]．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．

5.6 実験

5.6.1 準備

本実験では，R-B-LEG によるナッシュ Q 値の収束，およ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] × 10[エピソード] × 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，ともに 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．

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．

[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．

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とも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．

5.6.2 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 5.7: 報酬の平均と分散

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 5.8: 5000 回 (4501-5000 平均) との比較

R-B-LEG および 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 5.7. に示す．表 5.7. 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 5.7. を用いる．

$$差\,(reward)[\%] = \frac{reward - Reward}{Reward} \times 100 \tag{5.7}$$

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 5.8. 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．

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 5.9: 5000 回 (4501-5000 平均) との比較した場合の χ 2 乗検定

5000 回 までの差の比較結果を表 5.9. に示す. 表 5.7. の 5000 回 までの結果を用い, 式 5.7. の結果を ±4% で 0.5% ずつ 18 分割した結果が, 表 5.9. 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, はじめの 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.

5.6.3 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.

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 5.10: 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 5.11: エージェント A の勝ち負け数 (RandG)

R-B-LEG および RandG における エージェント A の勝ち負け数を，表 5.10. (R-B-LEG) および表 5.11. (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 5.12: 繰り返し $1 \sim 50$ 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 5.12. 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 5.9. では，R-B-LEG の繰り返し回数が 2500 回 以上で差が $\pm 2\%$ 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．

5.7 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．

第6章　結論

本研究では，マルチエージェント環境においての行動選択について議論した．特にマルチエージェント環境に対し報酬分配手法・相関手法の適用を提案し，非協力ゲーム・協力ゲームにおいて協調行動の獲得に対し有効に作用することを確認した．

本研究では，まず人工知能学会により提案された複雑なマルチエージェント環境である”Happy Academic Life 2006: 研究者の人生ゲーム”を基にした人生ゲームをモデル化し，ナッシュ Q 学習を適用した．そして，2 エージェントが対等な場合と対等ではない場合において，効果的に学習し協調行動が獲得できることを確認した．

次に，報酬分配手法を人生ゲームに適用した．報酬分配手法では，エージェントに与える報酬に制約を加えることでナッシュ均衡が一意に存在するようにしている．その上， Q 学習とナッシュ Q 学習を比較することで，複雑なゲームにおいて有効な戦略的知識を獲得できるか検証し，マルチエージェントシステム環境でナッシュ Q 学習が適用できることを示した．

協力ゲームでは，エージェント間に通信路を仮定し相互作用を表す共通の情報を許容した，相関手法を考えた．本研究では，相関手法に対応した局所影響ゲームを用いた．局所影響ゲームは局所影響関数を共通の情報として持つ．また，確率的に行動決定することを許容する混合戦略ナッシュ均衡を用いた．これらにより，相関手法が行動選択手法として有効であることを実験により示した．

繰り返し協力ゲームに拡張した場合では，相関手法に対応した，繰り返し双方向局所影響ゲームを用いた．繰り返し双方向局所影響ゲームでは，局所影響関数を共通の知識としてエージェント間で共有し相関手法に対応させている．一般のナッシュ Q 学習では収束性の保証が無いが，本研究では，局所影響関数を用いた相関手法に基づく混合ナッシュ戦略を提案し，実験により収束性と協調行動が獲得できることを示した．

本研究で提案したマルチエージェント環境における行動選択手法，特に最後の実験で用いた，局所影響ゲームはより実社会に近いゲームモデルであり，協調行動をとることが可能になるだろう．例えば，カーナビゲーションシステムにおいて，相関手法に対応させそれぞれ通信路持ち，確率的に最適な経路を指示することにより，一極集中による渋滞を軽減させることが期待できる．このように，提案手法は幅広い分野への応用が期待できる．

今後の課題の 1 つとして，エージェント数を増やす場合がある．本研究では，マルチエージェント環境として 2 エージェントを扱ってきた．そのエージェント数を 3 以上にした場合における振る舞いの変化について考える必要がある．本研究では，局所影響ゲームの中でも制約の強い双方向局所影響ゲームを用いている．そのため，その

制約を緩和した場合の振る舞いに関しても考察が必要である．また，ナッシュ Q 学習の収束性についは実験により示しているが，今後は，収束性についての証明をする必要がある．

謝辞

本研究を遂行するにあたり，日頃より適切な御指導，御鞭撻をいただいた，法政大学工学部情報電気電子工学科 三浦孝夫教授に深く御礼申し上げます．

並びに，産業能率大学情報マネジメント学部 塩谷勇教授にも多くの有益なご助言をいただきました．深く感謝いたします．

データ工学研究室の先輩方，同輩，後輩たちにも，本研究の遂行にあたって数多くの助言と快適で能率的な研究室環境を整えていただきました．御礼申し上げます．

修士論文として私の研究をまとめることができたのも，多くの皆様方の御支援，御協力の賜物であります．この場をお借りしまして，厚く御礼申し上げます．

最後に，今までの学生生活を支えてくださった私の両親，兄弟に深く感謝したいと思います．

参考文献

[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: Multiagent Reinforcement Learning ? Theoretical Framework and An Algorithm, ICML, pp.242-250, (1998)

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

[4] 井越 一穂, 三浦 孝夫: Q 学習のナッシュ均衡による戦略的知識の獲得, 第 19 回データベースワークショップ (DEWS), (2008).

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

[6] 井越 一穂, 三浦 孝夫: Local Effect Game における相関均衡を用いた行動選択, 電子情報通信学会 データ工学研究会, 電子情報通信学会技術研究報告, Vol.109(153), pp.7-11,(2009).

[7] 井越 一穂, 三浦 孝夫: 繰り返し局所影響関数を用いた行動選択, 第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM), (2010).

[8] K. Igoshi, T. Miura: Selecting Behavior on Repeated Local Effect Functions to appear in "2010 International Symposium on Collaborative Technologies and Systems (CTS 2010)", (2010).

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

[10] S. Kakade, M. Kearns, J. Langford, and L. Ortiz: Correlated Equilibria in Graphical Games, ACM Conference on Electronic Commerce (EC), (2003).

[11] S. Kitahara, Y. Tanigawa and H. Tsuruoka: Emergence of Cooperative Action in Nash-Q Learning, Research Bulltin of Fukuoka Inst.Tech. 40-1, pp.15-20, (2007).

[12] M.L. Littman : Markov Games as A Framework for Multi-agent Reinforcement Learning, ICML, pp.157-163, (1994)

[13] T.M. Mitchell : Machine Learning, McGraw Hill Companies, (1997)

[14] 武藤 滋夫: ゲーム理論入門, 日本経済新聞社, (2001).

[15] 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).

[16] S.J. Russel and P. Norvig: Artificial Intelligence - A Modern Approach (2/e), Prentice Hall, (2003)

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

[18] 高玉 圭樹: マルチエージェント学習 相互作用の謎に迫る, コロナ社, (2003).

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

[20] 国土交通省 道路局: 道路 IR・道路整備効果事例集/道路関連データ. (http://www.mlit.go.jp/road/index.html)

[21] 人工知能学会: Happy Academic Life 2006:研究者の人生ゲーム –ゲーム型キャリアデザイン学習教材の開発–, 人工知能学会誌 Vol. 21, No. 3, (2006). (http://academiclife.jp/)