# ナッシュQ学習における協調行動の生成

北原頌士（情報システム工学科）
谷川裕一（情報システム工学科）
鶴岡久（情報システム工学科）

## Emergence of Cooperative Action in Nash-Q Learning

Shouji KITAHARA (Department of Computer and Systems Engineering)
Yuichi TANIGAWA (Department of Computer and Systems Engineering)
Hisashi TSURUOKA (Department of Computer and Systems Engineering)

**Abstract**

The effect of Nash-Q learning algorithm has not yet been confirmed in multiple experiments. We adopted a 5×5 grid world in which two agents started from opposite lower corners and tried to reach their respective goal cell. Experiments showed performance differences between single agent Q-learning and Nash-Q learning. In the Nash-Q learning, both agents obtained similar accumulated rewards; however, in the Q learning, each agent accumulated his reward differently. Findings of this research confirmed that when agents adopt Nash Q-learning to predict the other agent's behavior, not only is the performance of the agents better than their performance when using single-agent Q-learning, but the emergence of the cooperative action can also be observed.

Keywords: *Q learning, Nash-Q learning, grid world, agent, reward*

## 1．はじめに

### 1.1　研究背景

マルチエージェント学習手法の中で、環境を事前に知る必要がない強化学習技術は研究者の強い関心を惹いてきた。なかでもルコフ決定過程において状態と行動の空間が有限ならば、学習の収束が保障されているQ学習が特に注目を集めている。シングルエージェント学習であるQ学習は直接マルチエージェントに適用することはできないが、ロボットサッカーや、追跡ゲーム、インターネットオークション等へ応用にされてきた。しかし、従来のシングルエージェントを対象としたQ学習をマルチエージェント学習に適用した場合、他のエージェントの行動による環境の変化を無視しており、マルコフ性が成立せず、学習の収束は保障されない。

Q学習法をマルチエージェント環境に拡張するための有力な方法は確率ゲームの導入であり、その基本解であるナッシュ均衡を価値関数の更新式に活用するナッシュQ学習がJ. Fu等によって提案されている。

### 1.2　研究目的

マルチエージェント学習では、各エージェントの報

平成19年5月30日受付

酬はエージェントの行動の組み合わせで決まるため、確率ゲームの枠組みでとらえることが有用であり、その基本解はナッシュ平衡解である。J. Fu 等はナッシュ Q 学習の収束性の理論的証明と、その格子ゲームによる確認実験を行っているが、ここではマルチエージェント強化学習アルゴリズムとして提案されているナッシュ Q 学習をシングルエージェントを対象とした Q 学習と学習性能の点で比較評価することを目的とする。

## 2．Nash-Q 学習

### 2.1　Nash-Q 学習とは

シングルエージェント Q 学習では最適 Q 値は利得を最大化するものと考えられるが、マルチエージェント学習では Q 値は他のエージェントの方策に依存し、確率ゲームの枠組みでは最適 Q 値はナッシュ均衡点で受け取る Q 値となる。従って Nash-Q 学習とは、任意の推定値から出発し、エージェントは試行を繰り返すことにより、ナッシュ均衡点を学習することである。このナッシュ均衡点とは､他のエージェントにとっても自己にとっても最良の行動をとった際の行動の組み合わせである。ナッシュ均衡点を行動価値関数のバックアップとして用いるため、お互いに最良な行動を選択し相手の行動に干渉しないため各エージェントが獲得できる報酬も高くなると考えられる。このため他のエージェントの Q 値を推測する、言い換えれば他のエージェントの行動を予測する機構が必要になる。またすべてのエージェントは自己の利得を最大化するよう行動する（自己犠牲などを目的としない）という合理性を有することを仮定する。

Nash-Q 学習の行動価値関数の更新式を以下に記す。

$$Q_{t+1}^{i}(s, a^1,..., a^n)=(1-\alpha_t)Q_t^{i}(s, a^1,..., a^n)+\alpha[\gamma_t^{i}+\beta NashQ_t^{i}(s')] \quad \cdots(1)$$

$$NashQ_t^{i}(s')=\pi^1(s')...\pi^n(s')\cdot Q_t^{i}(s')$$

$t$ ：現在の時刻（ステップ）
$i$ ：エージェント
$s$ ：現在の位置
$a$ ：現在の位置で取る行動
$\alpha$ ：正しいと推論された行動選択の修正率（学習係数）
$\gamma_t^{i}$ ：ステップ $t$ でエージェント $i$ が、行動 $a$ をとった時得る報酬
$\beta$ ：将来の報酬が現在においてどれだけの価値があるかを決定する率（割引率）
$NashQ_t^{i}(s')$ ：全てのエージェントが行動できる方向についてナッシュ均衡点を求める

今回の実験においてナッシュ均衡点とは、エージェントがその行動以外を選択した場合、獲得できる報酬が減少する行動の組み合わせをさす。ナッシュ均衡点の例を図 1 に示す。

（エージェント A，エージェント B）

| | 右 | 左 |
|---|---|---|
| 上 | (47，47) | (90，10) |
| 下 | (10，90) | (60，60) |

図 1．ナッシュ均衡点の例

このとき、エージェント A はもし右を選択してしまうと、どちらも得られる報酬は減少するため選択しない。また、エージェント B も同様に上を選択すると、得られる報酬は減少するため選択しない。よって、（左、下）がナッシュ均衡点となる。

### 2.2　Nash-Q 学習のアルゴリズム

図 2 に Nash-Q 学習のアルゴリズムを記す。

Q(s,a)を任意に初期化
各エピソードに対して繰り返し：
　s を初期化
　エピソードの各ステップに対して繰り返し：
　　Q から導かれる方策を使って、s での行動 a を選択する
　　行動 a をとり、r、s'を観測する
　　$Q(s,a)\leftarrow Q(s,a)+\alpha[r+\gamma \max_{a\in A} Q(s,a)-Q(s,a)]$
　　s←s'
　　s が終端状態なら繰り返し終了

図 2．Nash-Q 学習のアルゴリズム

## 3．行 動 推 測

### 3.1　行動推測の目的

強化学習において、エージェントの環境はエージェントの行動によって遷移する。マルチエージェント学習では、複数のエージェントが存在するため対象エージェントが置かれる環境は、他エージェントの行動によっても遷移するため行動決定に必要な情報が不十分になり、マルコフ性を維持できなくなり、学習の収束

が保障されない。

そこで、エージェント同士がお互いの行動を観測し、その観測情報を基にして相手の政策を推測するという方法をとる。これにより環境遷移の情報の精度をより高めることができ、マルコフ的なモデルに近づけることができる。

### 3.2 行動推測の手法

エージェントkがエージェントoの行動を予測する手順を以下に記す。

1．エージェントkが推定した他エージェントoの政策を$I^k(a^o \mid S)$とし、関数$\bar{Q}^k(S,a^k)$を、

$$\bar{Q}^k(S,a^k) \equiv \sum_{a^o \in A^o} I^k(a^o \mid S) Q^k(S,a^k,a^o) \quad \cdots(2)$$

とする。現在の状態$s_t \in S$において、エージェンkは政策（ε -グリーディ）に従って行動を確率的に選択する。

2．エージェントkは、手続き1で選択した行動を実行する。ここで、他のエージェントも同時に行動を選択し実行する。両エージェントの行動により、状態は現状態$s_t$から次状態$s_{t+1}$に移行し、エージェントkは環境から報酬$r^k_{t+1}$を受け取る。

3．エージェントkは、状態$s_t$、行動$a^k_t$、$a^o_t$に対する行動価値関数を式(4)に従い更新する。エージェントkは状態$s_t$においてエージェントoが選択可能な全ての行動$a^o \in A^o$に対して、式に従い関数$I^k$を更新する。

$$Q(s_t,a^k_t,a^o_t) \leftarrow Q(s_t,a^k_t,a^o_t) + \alpha[r^k_{t+1} + \gamma \max_{a \in A^k} \bar{Q}(s_{t+1},a^k_{t+1}) - \mathrm{Q}(s_t,a^k_t,a^o_t)] \quad \cdots(3)$$

$$I^k(a^o,s_t) \leftarrow (1-\theta) I^k(a^o \mid s_t) + \begin{cases} \theta\,(a^o = a^o_t) \\ 0\,(ohterwise) \end{cases} \quad \cdots(4)$$

ここで、式(4)中の$\theta$は観測した行動を将来の行動予測時にどれくらい考慮するかを決定するパラメータである。

4．学習の終了条件を満たしていれば終了する。そうでない場合tに1を加え手続き1に戻る。

以上の学習法をエージェントが自律的に行う。

## 4．実験概要

本実験では、3×3のフィールドと2体の学習エージェントを用いる。2体の学習エージェントは対角にある各ゴールを目指す。この際、エージェントは同時ゴールする必要はなく単独でもゴールすることが可能である。もし一方が先にゴールすれば、そこでゲームは終了であり、あとからゴールに入るはずのエージェントには報酬は与えられない。2エージェントが同時にゴールすれば両エージェントに報酬が与えられる。しかしゴールへ向かう経路がクロスしているため、互いの利益を尊重して協力しなければならない。

本実験において学習エージェントは対角にあるゴールを目指すため、シングルQ学習では自分の価値を最大化するように学習するため、相手の行動に干渉してしまいゴールまでの最短ルートの邪魔をしてしまう可能性がある。一方、Nash-Q学習では、お互いに干渉しないような最短ルートは両エージェントにとって最良行動であると考えられるため、ナッシュ均衡解に収束することが期待できる。

### 4.1 実験1：ゴールまでが同じ距離の場合

実験は3×3のフィールドで行い、学習エージェントは2体使用する。図3に示す。図中のGAとGBはそれぞれエージェントAとエージェントBのゴール地点である。

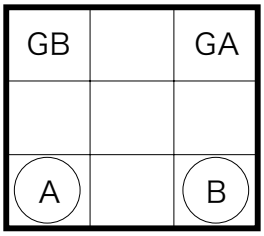


図3．実験フィールド

- エージェントは同時行動をとる。
- エージェントの行動は「上下左右止」のいずれかを選択し、実行する。
- エージェントが壁に激突した場合−20の報酬を得、エージェント同士が激突した場合は−10の報酬を得る。いずれの場合もエージェントは行動前の状態に戻される。

- １エピソードは100ステップに達した時点で強制的に終了する。
- エージェントが協力してゴールした場合、両エージェントは＋100の報酬を得る。また、片方のエージェントが単独でゴールした場合、その対象エージェントは＋100の報酬を得る。エージェントが一方でもゴールに到達したらエピソードを終了する。

### 4.2 実験結果

以上の条件で15000エピソードを10回試行し100エピソード毎に、各エージェントのゴール回数、累積報酬値、をそれぞれ調べた。

図４は15000エピソードを10回行い、100エピソード毎に平均した各エージェントのゴール回数をグラフ化したものである。図５は15000エピソードを10回行い100ステップ毎に平均した累積報酬値をグラフ化したものである。

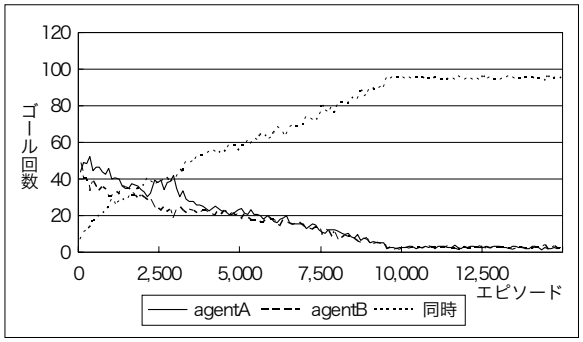


図４．Nash-Q 学習における各エージェントのゴール回数

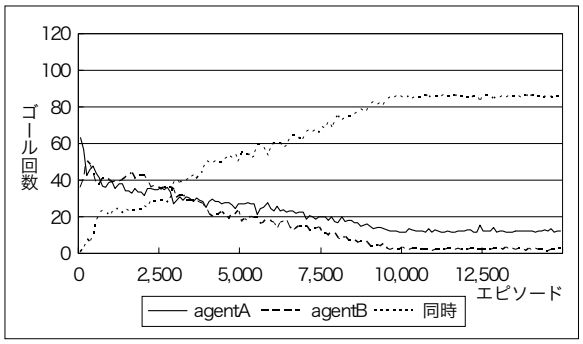


図５．Q 学習における各エージェントのゴール回数

グラフより、各エージェントのゴール回数は、Q 学習のほうが Nash-Q 学習に比べエージェント A の単独ゴール回数が多く見られる。これは、Q 学習では各エージェントがそれぞれ利益を最大化しようと行動した結果であり、２体のエージェント A、B が最短ルートをめぐり競合したためであると考えられる。

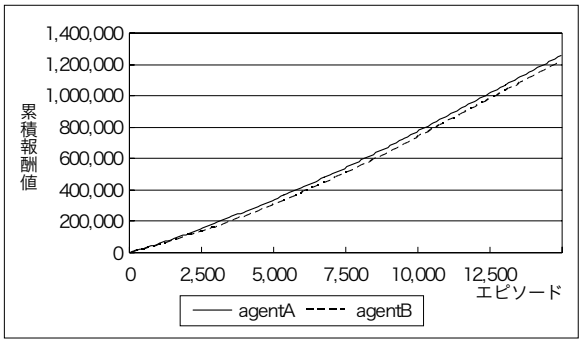


図６．Nash-Q 学習における累積報酬値

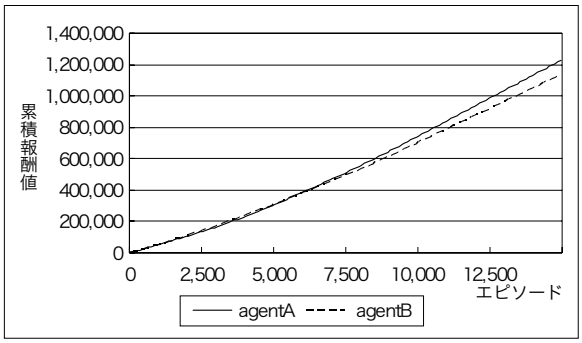


図７．Q 学習における累積報酬値

表１．Nash-Q 学習と Q 学習の累積報酬値の比較

| | Nash-Q 学習 | Q 学習 |
|---|---|---|
| エージェント A | 1256521 | 1226306 |
| エージェント B | 1217672 | 1133322 |
| 合　計 | 2474193 | 2359628 |

Nash-Q 学習では、２体のエージェントが獲得した累積報酬値の総計は Q 学習のそれより大きいことがわかる（図６、図７、表１）。これは、２体のエージェントが互いに競合しないように最短ルートを通り報酬を獲得したためであると考えられる。図８にエージェントが通ったゴールまでの経路を示す。

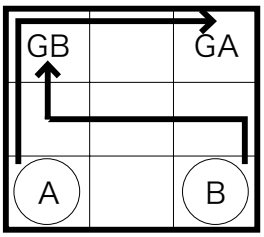

図８．最短経路

### 4.3　実験２：ゴールまでの距離が違う場合

実験１ではゴールまでの距離が同じであり協力が発生しやすい環境であった。実験２では、ゴールまでの距離が違う場合にも協力が起こるか実験を行う。エージェント A のゴール地点が違うだけでその他の条件は実験１と同じである。図９に示す。図中の GA と GB はそれぞれエージェント A とエージェント B のゴール地点である。

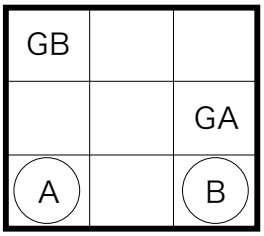

図９．実験フィールド

エージェントは同時行動をとる。

- エージェントの行動は「上下左右止」のいずれかを選択し、実行する。
- エージェントが壁に激突した場合−20の報酬を得、エージェント同士が激突した場合は−10の報酬を得る。いずれの場合もエージェントは行動前の状態に戻される。
- １エピソードは100ステップに達した時点で強制的に終了する。
- エージェントが協力してゴールした場合、両エージェントは＋100の報酬を得る。また、片方のエージェントが単独でゴールした場合、その対象エージェントは＋100の報酬を得る。エージェントが一方でもゴールに到達したらエピソードを終了する。

### 4.4　実験結果

以上の条件で15000エピソードを10回試行し100エピソード毎に、累積報酬値、各エージェントのゴール回数、ゴールまでにかかったステップ数をそれぞれ調べた。

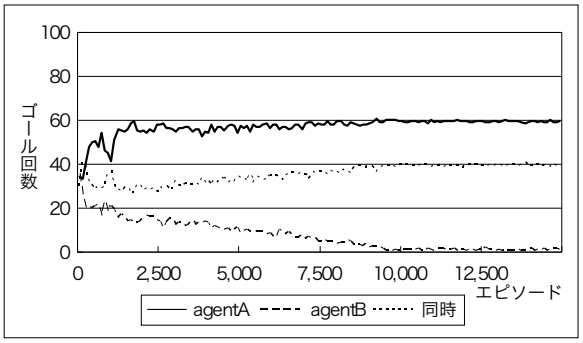

図10．Nash-Q 学習における各エージェントのゴール回数

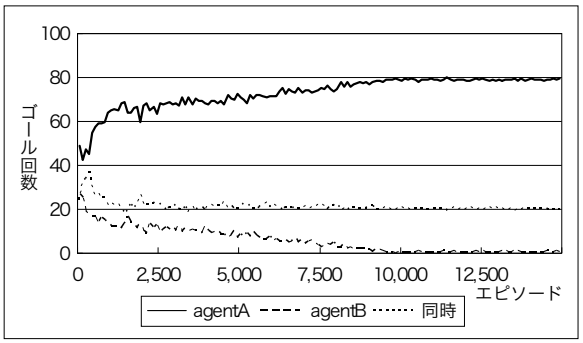

図11．Q 学習における各エージェントのゴール回数

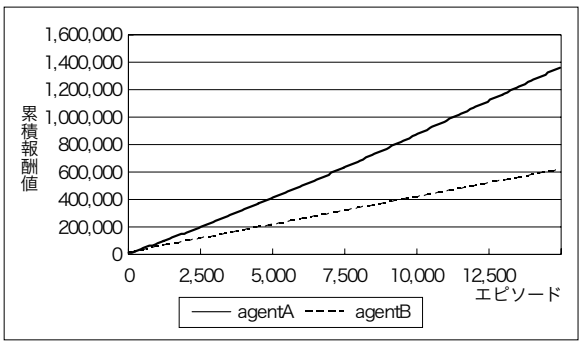

図12．Nash-Q 学習における累積報酬値

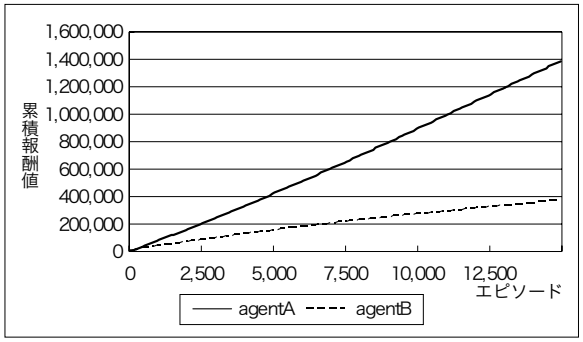


図13．Q 学習における累積報酬値

表２．Q 学習における累積報酬値

| | Nash-Q 学習 | Q 学習 |
|---|---|---|
| エージェント A | 1360466 | 1386043 |
| エージェント B | 614062 | 374866 |
| 合　計 | 1974528 | 1760909 |

図10、図11は15000エピソードを10回行い、100エピソード毎に平均した各エージェントのゴール回数をグラフ化したものである。図12、図13は15000エピソードを10回行い100ステップ毎に平均した累積報酬値をグラフ化したものである。表２にその累積報酬値の比較を示す。

グラフより、各エージェントのゴール回数を比較すると、Nash-Q 学習のほうがエージェントの同時ゴールの回数が多いことがわかる。一方、Q 学習ではエージェント A のゴール回数が多いことが目立つ。これは、エージェント A、B のゴールまでのステップ数が異なるためゴールまでのステップ数の短いエージェント A が早くゴールに到達できるためであると考えられる。Nash-Q 学習において、同時ゴールが多く起きているのは両エージェントにとって最良の行動を選択した結果、協調の関係が発生したためエージェントの同時ゴールが Q 学習に比べ増加したものと考えられる。

累積報酬値では、１体のエージェントのみで見るならば Q 学習のほうが高い累積報酬値を獲得できていることがわかるが、２体のエージェントの累積報酬値の合計を比べると Nash-Q 学習のほうが多く獲得できていることがわかる。これは、両エージェントがお互いに最良の行動を選んだため Q 学習に比べ高い報酬を獲得できたと考えられる。

図14、図15に、エージェントがゴールした最短ルートを示す。図中の黒丸は行動の「停止」を意味する。図14実験経路(1)ではエージェント A がエージェント B と激突しないように行動「停止」を選択していることがわかる。図15実験経路(2)ではエージェント B はエージェント A と経路が交差しないように行動しているが、エージェント A のほうがゴールまでのステップ数が少ないためエージェント A が先にゴールしてしまうことがわかる。図14において、エージェント A が停止という行動を選択するのは、上を選択した場合ゴールまでのステップ数がかかってしまうためであると考えられる。

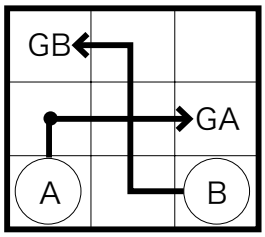


図14．最短経路(1)

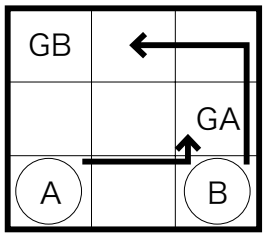


図15．最短経路(2)

## 5．結　　言

シングルエージェント Q 学習ではエージェントが互いの利益の最大化を図ればエージェント間に競合が発生する環境においては、片方のエージェントのみが多くゴールするという傾向が見られた。しかし、Nash-Q 学習では競合が発生するような環境においても２体のエージェントの累積報酬値を高められることがわかった。また、実験２よりエージェントの目標達成条件が異なる場合でも２体のエージェントの累積報酬値を高めることを確認できた。以上より、「Nash-Q 学習は競合が発生するマルチエージェント学習アルゴリズムとして有効であるといえる。」

今回ナッシュ均衡点は利得表におけるすべての格子点を逐一順番に均衡点の条件を満たすか、テストして発見する方法をとったが、エージェント数や行動数が多いと計算時間がかかり、今後の課題としては、効率的なナッシュ均衡点を求めるアルゴリズムに代える必要がある。

## 参　考　文　献

１）R. S. Sutton, A. G. Barto Reinforcement Learning, The MIT Press (1998)

２）J. Hu; Nash Q-Learning for General-Sum Stochastic Games J. M. L. R 4, (2003) pp.1039-1069