# Q学習に基づく一方向コミュニケーションの創発における行動選択の影響

# Effect of action selection on emergence of one-way communication using Q-learning

大分大学　○仲西賢展　柴田克成
Masanobu Nakanishi, Katsunari Shibata, Oita University

Abstract: In this paper, the effect of action selection policies in the learning of one-way communication between two agents using Q-learning is examined. The ratio of successful learning become larger when the receiver agent's action selection policy is greedy, and the transmitter agent's policy is not completely greedy, but with a small random factor. From the analysis of the learning process, it is known that inappropriate mapping from states to signals in the transmitter agent sometimes breaks the mapping from signals to actions severely in the receiver agents. Accordingly, the transmitter agent needs to find an appropriate mapping through exploration, while the receiver agents decides its action finally after the mapping is fixed in the transmitter. In that case, no exploration is necessary in the receiver agent.

## 1. まえがき

コミュニケーションを行うことは、群ロボットやマルチエージェントシステムにおいて、観測の不十分さを補うことや、利害の衝突を避けて協調的な行動を行うことなどに有効である。ロボットやエージェントに、目的に沿ったコミュニケーションを自律的・適応的に獲得させるために、進化的手法[1]や強化学習の適用[2][3]が試みられている。文献[1][2]では、受け手が行動をする際の観測の不十分さを補うための一方向コミュニケーションを進化的手法や学習によって獲得できる例が示されている。

これに対し筆者らは、コミュニケーション信号として、どういう情報を伝達すればよいのか、また、あらゆる場合に学習がうまくいくのかどうかといった点について検討するため、2 エージェント間の一方向コミュニケーションを、強化学習アルゴリズムの一つであるQ学習で獲得させる問題についてシミュレーションを行ってきた。そして、発信側が学習によって生成したコミュニケーション信号によって受信側が状態混同を起こし、受け手が部分観測状態に陥って行動が最適にならない場合があることを発見し、どのような場合に状態混同が発生するのか、いくつかのシミュレーション結果から考察した[4]。

これらの検証過程において、学習時の行動選択方法として用いたボルツマン選択の温度係数の下げ方が学習成功率に大きく影響することを発見した。

そこで、本稿では、いくつかのシミュレーション結果から、行動選択方法が学習に与える影響を検証するとともに、そのときのQ値の変化を解析し、その理由を考察する。

## 2. 一方向コミュニケーションの学習

### 2-1 構成

本稿では、文献[4]と同様に、文献[1][2]を参考に、オス、メスと呼ばれる2体のエージェントを仮定し、両者が接触したら、両者共に報酬がもらえるタスクを考える。構成図をFig.1 に示す。メスは視覚を持ち、オスの位置を特定することができるが、移動することができない。一方、オスは足を持ち移動できるが、視覚を持たず、メスの位置を直接特定することはできない。そして、メスは信号を発生し、オスはメスが発生する信号を正確に識別できるとした。また、オスの行動による状態遷移はすべて決定論的であるとした。ただし、コミュニケーション信号の意味は全く与えない。したがって、メスはどういう状態のときにどういう信号を送れば良いか、オスは送られてきた信号をどう解釈し、どう行動に反映させれば良いかを学習する。そして、オスとメスの間で何らかの共通の言語を確立することができれば、効率よく接触を繰り返すことが期待できる。

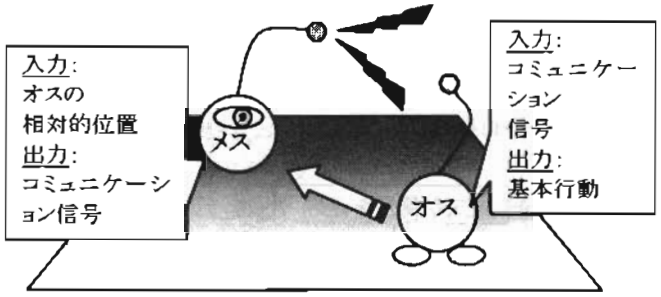



Fig.1 One-way communication problem between two agents

### 2-2 エージェントの学習

ここで、発信側および受信側のエージェント、すなわちオスとメスの学習について示す。本稿では、状態と行動を離散的としたため、両者の学習はQ学習で行った。Q学習は状態と行動の組に対して評価を行い、その評価値に基づいて確率的に行動を選択するとともに、その評価値であるQ値を学習していく方法である。

Q学習のアルゴリズムを以下に示す。

(1) エージェントは状態 $s_t$ を観測する。

(2) エージェントは任意の行動選択方法にしたがって行動 $a_t$ を実行する。


SY0011/04/0157～0162 ©2004 SICE

(3) 環境から報酬 $r_{t+1}$ を受け取る。

(4) 遷移後の状態 $s_{t+1}$ を観測する。

(5) 以下の更新式よりQ値を更新する。

$$Q(s_t,a_t) \leftarrow (1-\alpha)Q(s_t,a_t) + \alpha\left[r_{t+1} + \gamma \max_{i \in A} Q(s_{t+1},i)\right] \quad (1)$$

ただし $A$ は基本行動の集合、$\alpha$ は学習率 $(0<\alpha\leq 1)$、$\gamma$ は割引率 $(0\leq\gamma<1)$ である。

(6) 時間ステップ $t$ を $t+1$ に進めて(2)に戻る。

ここでは、メスの状態 $s$ はオスの相対的位置、行動 $a$ はコミュニケーション信号であり、オスの状態 $s$ はコミュニケーション信号で、行動 $a$ は基本行動である。

一方向コミュニケーションの学習のフローチャートをFig.2 に示す。オス、メスともに相手の行動が決まってから次の自分の状態が決まるため、ゴール時以外の Q 値の更新は相手の行動後に行った。

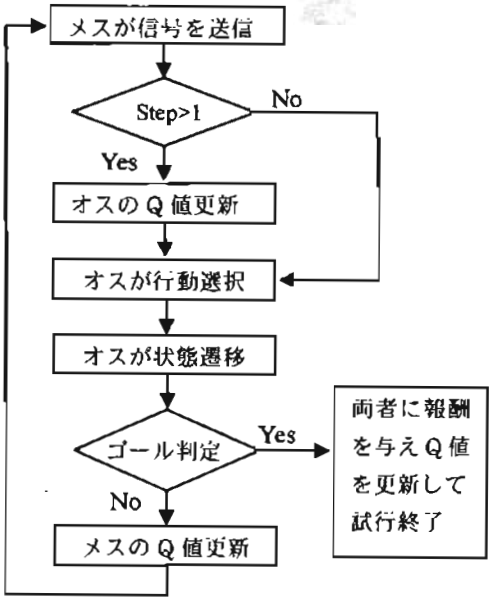



Fig.2 Flowchart of the learning of one-way communication

### 2-3 行動選択方法

・ボルツマン選択

Q 学習の学習過程における行動選択方法としてよく使われる方法にボルツマン(Boltzmann)選択がある。この方法は状態 $s$ において行動 $a$ を選択する確率 $p(a\mid s)$ を以下のように定義したものである。

$$p(a\mid x) = \frac{\exp(Q(s,a)/T)}{\sum_{i\in A}\exp(Q(s,i)/T)} \quad (2)$$

ただし、$T$ は温度係数である。$T$ は値が大きいほど選択はランダムになり、積極的に探索を行うことになる。逆に、$T$ を0に近づけると、わずかなQ値の差が行動選択に大きく影響し、極限では、最大の Q 値を持つ行動を選択することになる。本稿のシミュレーションでは、温度係数 $T$ は学習開始時に 1.0 とし、任意の試行回数で 0.005 になるように、指数関数的に徐々に小さくした。そして残りの試行では 0.005 に固定した。0.005 で止めたのは、それより小さくすると(2)式の指数関数の計算が困難になるからである。

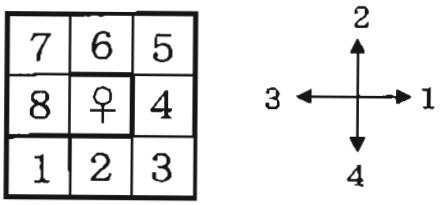
・Greedy 選択

Greedy 選択とは、確率的な要素を排除し、常にその状態におけるQ値が最大の行動を選択し続ける方法である。

## 3. シミュレーション

### 3-1 状態数より信号数、行動数が少ない場合

Fig.3 にシミュレーション環境を示す。




**Female** state:8 action:4
**Male** state:4 action:4

Fig.3 Simulation environment 1

ここで、太い線は壁で、左下がスタート、中心がゴール、すなわちメスの位置である。オスの行動は上下左右のセルに移動する 4 通りで、壁に向かう行動はその場にとどまるとする。この環境では、メスの状態は、オスの位置である 8 通り、行動は 4 種類のコミュニケーション信号である。一方、オスの状態は、メスから受信する 4 種類のコミュニケーション信号で、行動は上下左右のセルへの移動の 4 通りである。1セットの学習は 1000 試行行う。1試行の上限ステップを 1000 ステップと定め、1000 ステップ内にゴールしない場合は学習を打ち切り、スタートに再配置して次の試行へ移行する。また、Q 値の初期値はすべて 1.0 とした。これは、Q 値の初期値を高く設定する(Optimistic initial value)とgreedy の場合でも探索の効果があり、学習がうまくいく場合があるからである[5]。以下のシミュレーションにおいて、greedy な行動選択で、初期 Q 値を 0.0 とすると、学習は成功することが少なかった。また、ボルツマン選択の場合にも初期 Q 値の影響はあるものの、その差はあまり大きくなかった。また、本稿のシミュレーションでは、学習率 $\alpha$ を 0.1、割引率 $\gamma$ を 0.9、接触時の報酬 $r$ を 1.0 とした。

シミュレーションは、ボルツマン選択における温度の下げ方をメスとオスでそれぞれ変化させ、乱数系列を変えた1000 セットの学習における学習成功率を観察した。温度の下げ方は、ボルツマン選択の温度係数 $T$ を総試行回数のどの地点で 0.005 に収束させるかで変化させた。例えば、0.5 であれば、Fig.4 のように 500 試行で温度係数は指数関

数的に 0.005 に近づき、それ以降は 0.005 に固定される。$N$ 試行で温度を収束させるときの、第 $k$ 試行時の温度係数は次式で計算し、0.005 より小さい場合は 0.005 とした。

$$T(k) = 0.005^{\frac{k}{N}} \quad (3)$$

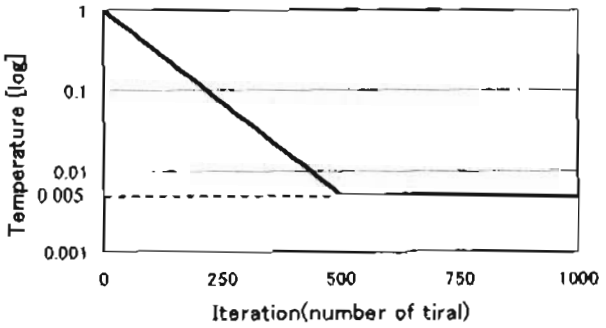



Fig.4 Schedule of the temperature coefficient in the Boltzmann selection

温度係数は、両者の収束点をそれぞれ 0.2 から 1.0 まで 0.2 ごとに変化させた全ての組み合わせを設定した。また、両者の行動選択方法を greedy 選択にした場合のシミュレーションも行った。各温度設定での学習成功率を Table 1 に示す。

Table 1 を見ると、オスが greedy 選択をし、メスがボルツマン選択を行う場合の学習成功率が最も高く、逆にメスが greedy 選択をしてオスがボルツマン選択をする場合は低くなっている。両者がボルツマン選択をする場合は、全体的にオスがメスより早く温度を下げたほうが学習成功率は高くなっていることがわかる。

メスとオスの収束点の差がなぜ学習成功率の差につながるのか考察した。メスは状態数が 8 通りあり、行動は 4 通りのコミュニケーション信号である。そのため、正しく学習できるように信号を決めるには、状態(1,2,8)、(3,4)、(5,6)、(7)の 4 つの組に対してそれぞれ一つずつ信号を割り当てなければならない。違う組の状態に同じ信号を割り当ててしまうと、オスはどう学習しても同一の信号に対しては同じ行動しかできないため、学習が成功することはない。それに対し、オスの状態は 4 通りであり、メスから受け取る 4 つの信号を 4 つの行動に割り当てるだけでよい。そのため、片方のエージェントが完全にランダムに行動を決定した後に、もう一方のエージェントが学習する場合では、メスがランダムに決定するよりも、オスがランダムに決定したほうが、学習がうまくいく確率が高いため、オスが先に温度を落とすことにより、学習が容易になるのではないかと推測した。そこで、メスとオスの条件を近づけるため、状態数と信号数と行動数を揃えた環境でシミュレーションを行った。

Table 1: Success ratio according to the number of trials when the temperature converged to 0.005 in the case of the environment 1 as shown in Fig.3.

| | | オスの温度収束点 | | | | | |
|---|---|---|---|---|---|---|---|
| | | greedy | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
| メスの温度収束点 | greedy | 99.7 | 48.1 | 35.2 | 15.4 | 4.1 | 1.4 |
| | 0.2 | 99.4 | 74.3 | 70.7 | 59.8 | 33.6 | 11.8 |
| | 0.4 | 99.2 | 76.3 | 71.2 | 63.6 | 36.7 | 12.6 |
| | 0.6 | 99.1 | 74.3 | 73.5 | 64.9 | 40.8 | 12.6 |
| | 0.8 | 98.5 | 72.7 | 71.2 | 64.1 | 44.2 | 14.6 |
| | 1.0 | 97.6 | 69.0 | 69.3 | 65.4 | 46.8 | 21.3 |

### 3-2 状態数と信号数、行動数が同じ場合

状態数と信号数と行動数を揃えた環境を Fig.5 に示す。このシミュレーション環境では、メスの状態はオスの位置である 8 通り、行動はコミュニケーション信号の 8 通りである。一方、オスの状態はメスから受信する 8 通りのコミュニケーション信号で、行動は Fig.5 に示すように状態遷移する 8 通りである。この環境は Fig.3 の環境とは違い、メスの状態と行動、オスの状態と行動の数は全て同じである。この環境でも同様のパラメータ設定で、1000 セットのシミュレーションを行い、成功率を調べた。結果を Table 2 に示す。

この環境では、メスは 8 つの状態に 8 つの信号を、同じものが出ないように割り当てなければならないので、Fig.3 に示した環境と比べて学習が難しい。したがって、Table 2 に示した結果は、Table 1 の場合と比べて、平均的に成功率が低くなっている。

Table 2 を見ると、両者の状態数、行動数をそろえた場合でも、オスが greedy 選択を行い、メスはボルツマン選択を行う場合の学習成功率が最も高くなり、10 割に近い学習成功率を得られた。一方、メスが greedy 選択を行い、オスがボルツマン選択を行う場合の学習成功率は低かった。この結果から、オスは学習初期にランダムな行動を行い試行錯誤することにあまり意味はなく、始めから greedy 選択を行う場合に一番学習効率がよいと考えられる。

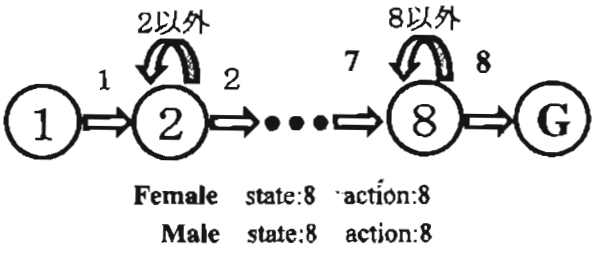



Fig.5 Simulation environment 2

Table 2: Success ratio according to the number of trials when the temperature converged to 0.005 in the case of the environment 2 as shown in Fig.5.

| | | オスの温度収束点 | | | | | |
|---|---|---|---|---|---|---|---|
| | | greedy | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
| メスの温度収束点 | greedy | 87.1 | 44.7 | 38.7 | 30.2 | 19.8 | 9.8 |
| | 0.2 | 100.0 | 82.0 | 77.0 | 61.1 | 32.2 | 7.6 |
| | 0.4 | 99.9 | 80.1 | 78.8 | 61.0 | 36.0 | 9.9 |
| | 0.6 | 99.1 | 64.8 | 62.4 | 57.4 | 34.9 | 9.0 |
| | 0.8 | 81.9 | 41.0 | 36.4 | 32.1 | 21.8 | 7.6 |
| | 1.0 | 62.8 | 6.1 | 6.9 | 5.6 | 2.7 | 0.4 |

Table 3: Success ratio according to the number of trials when action selection changed to greedy selection after the temperature converged to 0.005 in the case of the environment 2 as shown in Fig.5.

| | | オスの温度収束点 | | | | | |
|---|---|---|---|---|---|---|---|
| | | greedy | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
| メスの温度収束点 | greedy | 87.1 | 84.0 | 67.2 | 40.5 | 20.3 | 9.8 |
| | 0.2 | 90.5 | 86.9 | 75.9 | 47.9 | 22.4 | 10.4 |
| | 0.4 | 90.5 | 90.9 | 83.4 | 65.2 | 28.7 | 11.9 |
| | 0.6 | 90.4 | 87.9 | 85.9 | 77.8 | 53.0 | 14.4 |
| | 0.8 | 83.8 | 84.4 | 85.8 | 79.3 | 54.5 | 18.6 |
| | 1.0 | 62.8 | 61.5 | 56.8 | 36.7 | 6.8 | 0.4 |

また、メスの方は、ボルツマン選択の場合には、温度を早く下げたほうが成功率が高い傾向があるのに対し、greedy 選択の場合は、逆に成功率が落ちていることが分かる。このことから、メスにおいてはgreedy選択を行うことは、学習成功率を落とす結果となり、わずかでもランダムな行動を残しておくことが学習を成功させる秘訣ではないかと推測される。

そこで、次に、両者が greedy 選択を行う場合の学習成功率への影響を調べるために、Fig.5 に示した環境で学習を行う場合に、ボルツマン選択の温度係数が 0.005 になった後に、両者の行動選択を完全に greedy 選択とした場合について、シミュレーションを行った。Table 2 の場合は、ボルツマン選択の温度係数が 0.005 になった後にも、ほぼ greedy ではあるが、確率的要素のため Q 値が最大でない行動をとることがある。しかし、この場合には、いったん温度が落ちた後は確率的要素はなく、常に Q 値が最大の行動をとる。結果を Table 3 に示す。

Table 3 を見ると、Table 2 より、オスが最初から greedy 選択を行う場合の学習成功率が低くなっている。これは、メスが温度収束後に greedy 選択を行うためであると考えられる。逆に、オスが先に温度を収束させる場合には、Table 2 と比較して学習成功率が上がっている。これは、オスの行動選択が最終的に greedy 選択になるためであると考えられる。

この結果から、オスの greedy 選択は学習成功率を上げ、メスの greedy 選択は逆に学習成功率を落とすということがわかった。また、メスより早くオスの温度係数を下げる場合に成功率が高かったのは、結局オスが総試行回数の早い段階で greedy 選択に近くなるためであると考えられる。

そこで、オスが greedy 選択でメスがボルツマン選択を行う際になぜ学習がうまくいくのか、また、メスを greedy 選択にするとなぜ学習成功率が下がったのかを探るために、それぞれの場合で学習過程の Q 値の変化を観察した。

Fig.6, 7 に示したのは、オスの行動を greedy 選択にしたときについて、メスの行動選択を、greedy 選択にした場合(Fig.6)と、温度係数を初期値から 0.005 に固定してボルツマン選択を行わせた場合(Fig.7)の学習過程での Q 値の変化である。それぞれの行動選択方法で、(a)は、状態 8、すなわちゴールの 1 ステップ前の状態にオスが位置しているときの、メスの信号の Q 値の変化である。また、(b)は、メスから受け取る各信号を状態としたときの、行動 8、すなわちゴールへ向かうオスの行動の Q 値である。試行回数はそれぞれ 1000 試行行った。

Fig.6, 7 を見ると、ともに学習が少し進んだ頃から常にどれか 1 つの Q 値のみが高くなっている。メスにおける最も高い Q 値の信号とオスにおける最も高い Q 値の信号が対応しており、状態 8 では、メスの信号に従ってオスは正しい行動をとることができるようになっていることが考えられる。また、最も高い Q 値に対応する信号は、何回か入れ替わっていることが分かる。そして、最終的に信号をひとつに決め、オスもメスが最終的に選んだ信号を受け取り、両者の Q 値は 1.0 に収束している。

Fig.6 のメスの行動選択が greedy 選択の場合では、オスの Q 値は、試行回数の早い段階で大きく上がり、上がった Q 値が一気に下がるという変化が多く見られる。一方、Fig.7 のボルツマン選択の場合では、オスの Q 値は比較的低い値で振動している。また、greedy 選択の場合は、信号と行動の入れ替わりが激しく、最終的にゴール手前の信号が決まるのが遅い。それに対しボルツマン選択は信号が変わることが少なく、オスの Q 値が一度 1.0 まで上がると、その後 Q 値が下がったとしても、その行動が最終的に選ばれる。乱数系列を変えていくつかシミュレーションを行ったところ、ボルツマン選択でオスの Q 値の振動が多少大きくなる場合も確認されたが、基本的には Fig.6、Fig.7 のような傾向が確認できた。また、メスが greedy 選択で、学習に失敗したときは、オスの最大 Q 値の入れ替わりと Q 値の振動が続いて Q 値が収束しなかった。

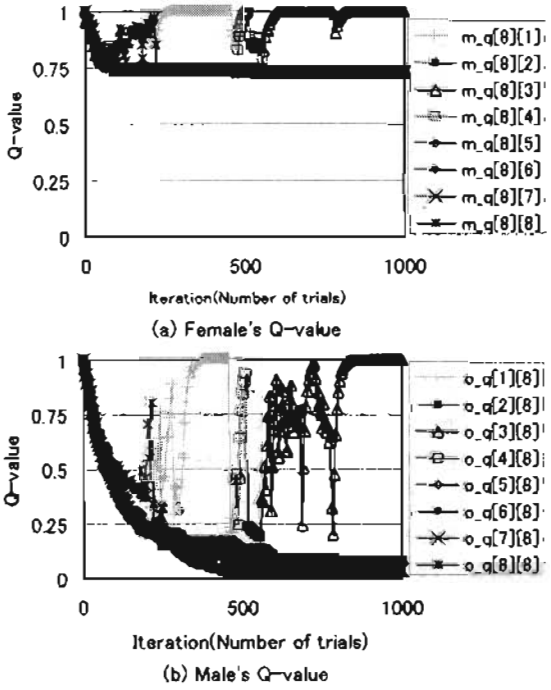



Fig.6 Change in the Q-value when the female's action selection is Greedy-selection

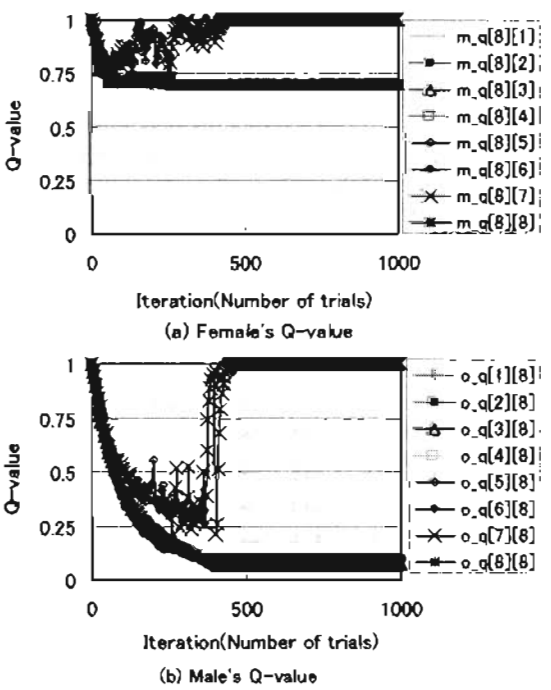
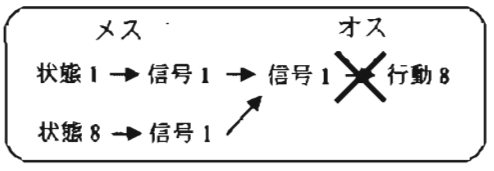



Fig.7 Change in the Q-value when the female's action selection is Boltzmann-selection

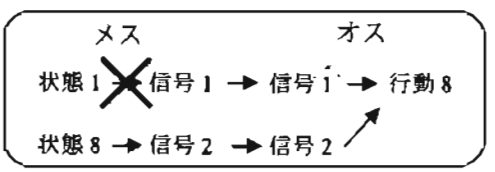
そこで、メスが greedy 選択の場合の、オスの Q 値が急激に下がるときの行動を観察した。Q 値が 1.0 となっているときは、ゴール手前の状態 8 で行動と信号のペアができている。しかし、ゴールから離れた別の状態で、メスがゴール手前の状態に割り当てていた信号と同じものを出したときに、オスは greedy 選択を行っているため、Q 値が下がりきるまで同じ行動を繰り返す。その結果、オスがゴールする行動の Q 値がその他の行動の Q 値と近い値になった後に、別の行動と入れ替わる。このようにして、メスが複数の状態で同じ信号を出すようになるときに、オスが最適ではない行動を取り続け、Q 値が大きく落ちることが分かった。このような組み合わせの探索を行う間に、全ての状態でメスが違う信号を割り当てることができたら学習はうまくいくと考えられる。そして、この組み合わせの探索を行える試行回数が多いほど、学習成功率があがることを確認した。一度全ての状態で信号と行動の組み合わせが完成すれば、両者の行動選択は greedy に近い状態なため、Q 値は落ちることなく学習は成功する。結局、メスがボルツマン選択を行うときには、信号と行動の入れ替わりが少なく、安定して学習ができるのに対し、greedy 選択を行う場合は、Q 値の変動が大きく、激しく信号が入れ替わるので、この差が学習成功率の差につながっているのではないかと考えられる。

そこで、メスの行動選択により、学習成功率と、Q 値の変化の様子に差がでる理由について考えてみる。Fig.8、に片方のエージェントの学習が、もう片方のエージェントに与える影響の例を示す。

メスの学習がうまく行ってないときは、複数の状態で同じ信号を選択する場合である。例えば、状態 8 において、メスの信号 1 と、オスの行動 8 が学習によりペアになっていたとき、メスが状態 1 でも信号 1 を出したとする。この場合は、オスは greedy 選択であるで、同じ信号を受け続けると、行動 1 をとり続ける。メスのこの失敗により、その状態 1 だけでなく、ゴール手前の状態 8 における信号と行動のペアも崩れ、学習をやり直すことになる。一方、オスの学習がうまくいってないときは、メスが全ての状態で違う信号を割り当てたにもかかわらず、違う信号を受信したときに同じ行動をとる場合である。例えば、メスが状態 1 で信号 1、状態 8 で信




(a) When the mapping in the female is not appropriate




(b) When the mapping in the male is not appropriate

Fig.8 The influence of one agent's learning to the other's learning

Table 4: The number of trials when an action was fixed for each state of each agent. The number before an arrow indicates the number of trials when the Q-value that is the maximum at last became the maximum. The number after an arrow indicates the number of trials when the maximum Q-value converged.

| state | メスの行動決定 | | | オスの行動決定 | | |
|---|---|---|---|---|---|---|
| 1 | 529 | → | 726 | 556 | → | 749 |
| 2 | 420 | → | 703 | 476 | → | 725 |
| 3 | 412 | → | 699 | 515 | → | 720 |
| 4 | 406 | → | 681 | 515 | → | 703 |
| 5 | 353 | → | 665 | 515 | → | 684 |
| 6 | 207 | → | 648 | 514 | → | 666 |
| 7 | 163 | → | 629 | 525 | → | 646 |
| 8 | 528 | → | 606 | 528 | → | 622 |

号2を出していたにもかかわらず、オスが信号1を受けても信号2を受けても、同じ行動8とっていたとする。このとき、オスの信号1における行動8のQ値が下がり、メスも状態1における信号1のQ値が下がる。しかし、この場合は、メスが状態8で信号2を出し、オスが信号2を受けて、行動8をとるところには影響はない。このように、メスは行動選択を失敗すると、他の状態でのオスの正しい行動を崩すことになるが、オスが行動選択を失敗する場合は、その状態のメスの信号しか崩さない。この差が、オスが先に greedy 状態になったときに学習が成功しやすく、逆に、メスは、わずかにランダム要素を残した行動選択を行うことが学習成功率を上げる原因になると推測される。

次に、オスとメスの行動は、学習過程でどのように決定されるかを調べるため、行動が決定するステップ数を調べた。このとき、学習は1000試行行い、メスの行動選択はボルツマン選択で、2割の地点で温度係数を0.005とし、オスはgreedy選択をする。結果をTable 4に示す。

Table 4の矢印の左に示した数字は、各状態で1つの行動が決まり、他の行動を起こさなくなった試行数を示している。また、矢印の右はQ値が収束したときの試行数を示している。

Table 4を見ると、よりゴールに近い状態から順にQ値が収束する傾向があるのが分かる。また、全ての場合で、オスが行動を決定するよりわずかにメスが行動を決定するステップが早いことがわかる。前述の推測のように、メスが失敗することが、学習を壊すことにつながるので、この場合はメスが先に信号を決定し、オスはそれに追従するように行動を決定することが、学習成功の秘訣になると考えられる。

## 4. まとめ

本稿では、2エージェント間で一方向コミュニケーションをQ学習によって学習させる問題について、行動選択法の影響を実験的に検証した。その結果、受信側のエージェントが学習初期にランダムな行動をとることはあまり学習に良い影響をあたえず、はじめからgreedyに行動を選択し、逆に送信側のエージェントは完全にgreedyに選択するよりはわずかにランダムな行動を行うほうが学習成功率は高いことがわかった。

また、学習過程のQ値の変化や、行動決定の様子を観察したところ、送信側が探索をし、受信側がそれに追従するという形で学習が進む様子が確認できた。このとき、メスの行動選択の失敗は、他の状態の学習にも影響を与えるのに対し、オスの行動選択の失敗は、その状態の学習を壊すだけである。そのため、メスはわずかにランダムな行動選択を残し、探索を行うほうがよく、オスはメスが決めた信号に追従するために、greedy に行動を決定することにより、学習効率が良くなると推測される。

本稿で示したシミュレーションは、推測が正しければ、ある程度一般的に言えることであるが、さらに環境やパラメータによる影響を検証していく必要がある。

謝辞

本研究は、日本学術振興会科学研究費補助金基盤研究(B)(14350227, 15300064)の補助の下で行われた。ここに謝意を表する。

参考文献

[1] G.M. Werner & M.G.Dyer: Evolution of Communication in Artificial Organizing System, Proc.of Artificial life II, 1/47(1991)

[2] N.Ono, T. Ohira, and A. T. Rahmani: Emergent Organization of Interspecies Communication in Q-Learning Artificial Organism,*Advances in Artificial Life*, pp. 396-405 (1995)

[3] 柴田克成, 伊藤宏司: 利害衝突回避のためのコミュニケーションの学習-リカレントニューラルネットワークを用いたダイナミックコミュニケーションの学習- 計測自動制御学会論文集 Vol.35, No.11, 1346-1354 (1999)

[4] Masanobu Nakanishi, Masanori Sugisaka & Katsunari Shibata: Occurrence of State Confusion in the Learning of Communication Using Q-leaning,Proc. of The 9th AROB (Int'l Sympo. on Artificial Life and Robotics), Vol. 2, pp. 663-666, (2004)

[5] R.S. Sutton and A.G.Barto,: Reinforcement Learning, The MIT Press,1998