# マルチエージェント強化学習における追跡問題の考察

菊地 敏博*　　神谷 昭基**

# A study of Pursuit Problem in Multi-Agent Reinforcement Learning

Toshihiro KIKUCHI　　Akimoto KAMIYA

**Abstract**- Reinforcement learning (RL) is a technique that allows an agent learn various action patterns in unknown environments, and has received a lot of attention around the world. RL, a kind of unsupervised learning method, is based on a reward scheme which is to learn action patterns such that the rewards are maximized. Multi-agent RL is a learning algorithm that allows agents learn how to work cooperatively. In this paper, we would like to propose a learning technique that can be more effective in solving the pursuit problem, a typical task for multi-agent RL. Our proposal has been verified by a successful simulation result. In this paper, we also examine the adaptability of agents to changing environments by changing the conditions of the pursuit problem.
**Key Words**: reinforcement learning, multi-agent, pursuit problem

## 1　はじめに

強化学習とは教師などの事前知識を用いずに，報酬という特殊な入力を手がかりに，エージェント (学習対象) が最適行動を学習する教師無し学習の一種である。

強化学習では，試行錯誤的に学習を行うことから人間が考えた以上の有効な行動を学習する可能性がある。また，環境に対する知識を必要としないため，環境変化に対応しやすいといえる。

これらの特徴から，強化学習は未知環境下での人工知能の行動獲得手法として注目されている。更に近年は複数のエージェントに協調動作を学習させるマルチエージェント強化学習の研究にも期待が寄せられている。

本論文では，より効率的な強化学習における学習高速化手法を追跡問題という問題において提案し，検証を行う。更に追跡問題の学習環境に変化を加え，それがもたらす学習への影響を調べ強化学習の環境変化への対応性を検証する。

* 釧路高専専攻科 電子情報システム工学専攻
** 釧路高専情報工学科

## 2　強化学習

強化学習とはエージェントと環境との相互作用，そして環境からエージェントに与えられる報酬を用いて，問題解決に適した行動をエージェントに学習させる方法である。

エージェントは限定された環境の下で行動を決定し，実行する。環境は，この行動により変化した状態をエージェントに提示し，エージェントはその新たな状態の下でまた新たに行動を行う。行動により変化した状態が問題解決に望ましい状態だった場合，環境はエージェントに報酬を与える。そして報酬を与えられた際に行っていた一連の行動は強化され，より高い頻度で実行されるようになる。

強化学習ではこの相互作用の繰り返しにより，得られる報酬が最大となる行動を学習する。相互作用の概念図を，図 1 に示す。

具体例として，人が餌を利用して鳥に飛行訓練をさせる場合を挙げる。この場合，鳥がエージェントであり，餌を与える人が環境である。初めは，鳥は飛び方が分からずに歩いたり，墜落したりを繰り返すが，いずれは偶然的に飛行が成功すると考えられる。そして

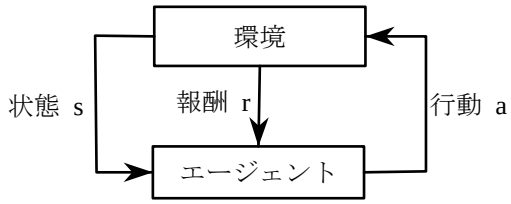


図 1: エージェントと環境の相互作用

飛行が成功したときに餌を与えると，鳥はより多くの餌を得ようとして，飛行に成功した時の羽や体の動かし方を繰り返し行うようになる。これを繰り返すことにより，最終的に，鳥は飛行するための一連の手順を学習することができる。

強化学習は，このように試行錯誤と報酬のみで問題解決のための行動を学習することができる。つまり強化学習は，問題の環境に対する事前的な知識や教師が存在しないような場合でも，問題に対する最適解を求めることが可能である。

強化学習ではエージェントは環境の状態を知覚し，その状態に適した行動を実行する必要がある。そのためには方策，報酬関数，価値関数の 3 つの要素が必要になってくる [2]。

方策は，エージェントが知覚したある状態に対して，どのような行動を行うか定義したものである。つまり，方策はエージェントの挙動を決定する要素であり，強化学習の中核を成す。

次に，報酬関数は知覚した状態を報酬と言う 1 つの数字として表すものである。つまりある状態がエージェントにとって望ましい状態か定義する関数ということになる。

また，価値関数は，現在の状態だけでなく，その後に続きそうな状態群から獲得が予想される報酬を考慮した，長期的な望ましさを定義する関数である。

強化学習はこれらの要素を用い，状態知覚，方策による行動決定，行動による環境変化，報酬の取得による方策 $\pi$ の最適化を繰り返していく。そしてどのような状態でも報酬値が最大となる行動を選択できる，方策 $\pi$ を求める。

マルチエージェント強化学習は，エージェントが複数存在する強化学習で，複数のエージェント間での協調動作を学習させ，より複雑な問題を解くことができる。しかし，エージェントの増加により様々な問題が発生してしまう [3]。それらの問題に関しては次節において述べ，問題解決方法に関しても案を挙げていく。

# 3　本研究における追跡問題の定義

本節では，本研究で行う追跡問題についての定義や，用いる学習アルゴリズムなどについて述べていく。

## 3.1　学習空間の定義

本研究では，マルチエージェント強化学習の標準的問題としてしばしば扱われる追跡問題において，学習高速化手法の提案を行う。更に，獲物の条件を変化させ，強化学習の環境変化への対応性を検証する。また追跡問題とは，複数のエージェントが協調しあい獲物を捕獲する問題である。

本研究で行う追跡問題を次のように定義する。

- 7 × 7 のトーラス状の二次元空間上で行う。
- エージェントは 3 体，獲物は 1 体もしくは 2 体配置する。
- 獲物やエージェントは，各隣接マスへの移動と待機の 5 種類から行動を選択する。各物体は同時に行動し，1 回の行動を 1 ステップとする。
- 同一の獲物 1 体に 2 体以上のエージェントが隣接した場合に捕獲とし各エージェントに報酬を与える。
- 各物体の初期配置はランダムとする。
- 各物体が同一のマスへ移動した場合，衝突として移動は行われない。エージェント間の衝突が起こった場合は，エージェントに負の報酬を与える。

この学習空間の概念図を，図 2 に示す。またこの図は，捕獲した状態を表している。

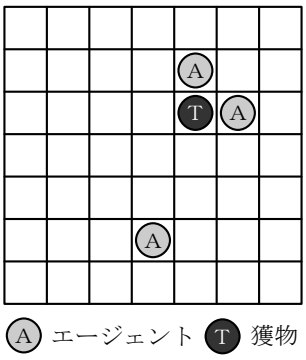


図 2: 学習空間の概念図

トーラス状の空間とは，端と端がつながっているドーナツ状の空間を指す。また，エージェントや獲物の行動決定アルゴリズムについては，次小節から述べていく。

## 3.2 Profit Sharing

強化学習に使用するアルゴリズムは多いが，環境を推定しながら各状態の価値を学習する環境同定型と過去の経験に基づいて行動を学習する経験強化型に分けることができる。

マルチエージェント強化学習では時間により各エージェントの方策が変化するため，次状態を推定することは非常に困難である。よって，環境同定型のアルゴリズムは学習を進めるのに適さない。

そこで本研究では，経験強化型の代表的アルゴリズムである Profit Sharing を用いて学習を行う。これはマルチエージェント強化学習に適した学習アルゴリズムであり，学習の収束が早いことが確認されている [4]。

### 3.2.1 Profit Sharing における行動選択

Profit Sharing では，状態 $s$ で行動 $a$ をした際の重要度 $V(s,a)$ を持っている。状態 $s$ における行動 $a$ の決定は，一般的に状態 $s$ の重要度 $V(s,a)$ を要素としたルーレット選択によって行われる [4]。これにより，重要度の高い行動ほど採択される確率が高くなり，低い行動ほど採択されにくくなる。

ルーレット選択により，ある状態 $s$ において次ステップで行動 $a$ が選択される確率 $P$ は次の式で表される。

$$P(a|s) = \frac{V(s,a)}{\sum_{a'} V(s,a)}$$

本研究の追跡問題においては，状態 $s$ は物体の空間への配置パターン，行動 $a$ は上下左右と待機の各移動となる。ルーレットには各状態で取るべき行動，つまり方策が格納されていることとなる。図 3 に，ある状態に対する重要度と行動選択ルーレットを示す。

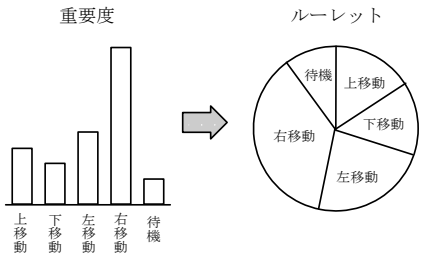


図 3: 重要度と行動選択ルーレット

### 3.2.2 Profit Sharing の概要

学習の開始から終了までの時系列を試行といい，初期状態から終端状態までの時系列をエピソードという。1 試行は，複数のエピソードで構成されており，Profit Sharing では，エピソードの終了時にのみ報酬が与えられる。試行とエピソードの関係図を図 4 に示す。本研究では，初期配置後の行動開始から獲物の捕獲までを 1 エピソードとする。

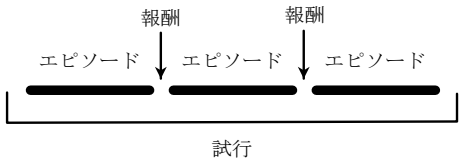


図 4: Profit Sharing における試行とエピソードの関係

Profit Sharing はエピソードの終了時に報酬を与える際，エピソード内で行った全行動の重要度を強化する。この点を考慮して，手順 3 で報酬を与えて重要度を更新する際は，次のような等比減少関数による更新式を用いる [5]。

$$V(s_t, a_t) \leftarrow V(s_t, a_t) + \gamma^{T-t-1} r$$

式中の $t$ は現在時刻 (現在のステップ数) を表す。$s_t$，$a_t$，$V(a_t, s_t)$ はそれぞれ時刻 $t$ での状態，行動，重要度である。$r$ は基本報酬であり，更に $T$ は終端状態に到達した際のステップ数で，$\gamma$ は割引率 $(0 \leq \gamma \leq 1)$ である。

この更新式より，時刻 $t$ が小さくなるほど (前の時刻になるほど) 報酬が割り引かれ，報酬が与えられる直前の行動ほど高い報酬が与えられるということになる。図 5 に，このステップ数と報酬値の関係を掲載する。

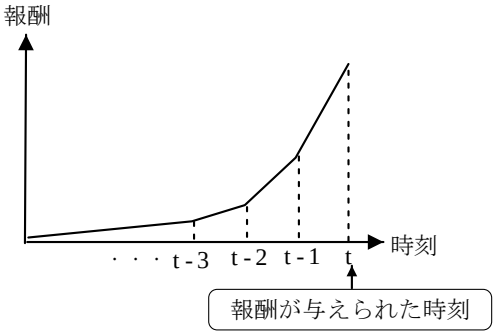


図 5: ステップ数による報酬値の推移

このような更新式で報酬を与えるのは，エピソード終了直前の行動ほど，問題解決に貢献した強化すべき行動という考えからである。

### 3.3 獲物の逃走アルゴリズム

本研究の追跡問題では，獲物はあらかじめ定められた逃走アルゴリズムに従って行動する。

獲物の逃走アルゴリズムは視界は縦横の十字方向のみに限定した場合と，全範囲に広げ死角をなくした場合の 2 つのパターンを用意した。

十字方向のみの限定視界では，獲物は「視界内のエージェントから遠ざかり，視界内にエージェントがいない場合はランダムに行動する」というアルゴリズムで逃走を行う。

視界が全範囲の場合は,「各行動選択後の獲物との距離を予測し，最接近している獲物の距離が最も大きくなるように行動する」というアルゴリズムで逃走を行う。

### 3.4 報酬の配分方法

マルチエージェント強化学習では，報酬の配分方法によっては問題解決に貢献しないエージェントや，報酬のみを求めて行動し，かえって問題解決を遅らせてしまう行動を学習するエージェントが現れる可能性がある。そこで，各エージェントに対する報酬配分を適切に定義する必要がある。

捕獲時に獲物に隣接したエージェントを直接エージェント，それ以外を間接エージェントとする。更に，直接エージェントに与えられる報酬を直接報酬，間接エージェントに与えられる報酬を間接報酬とする。

従来手法では間接報酬を割り引いて与えることが多かったが，本研究では，次のような報酬配分方法を用いることとする。この方法は，従来のものと比べて学習の早さ，協調行動の獲得が確認されている [1]。

- 捕獲時，直接報酬は無条件に与える。
- 間接エージェントが逃走要因となった獲物を逃走直後に捕獲した場合のみ間接報酬を与える。
- 直接報酬と間接報酬は同値とする。

間接報酬が制限されるため，問題解決に関係ない行動が強化されることはなくなる。また，間接報酬と直接報酬が同値であるため，直接報酬だけを優先させるような行動も抑制できる。

## 4 実験と考察

ここでは，前節で定義した問題，紹介した手法を用いて実験した結果を述べ，考察する。

また，本研究の実験では，Profit Sharing の各パラメータを割引率 $\gamma = 0.5$，基本報酬 $r = 2,000$，また重要度の初期値を 100 として実験を進めた。これらは何度か学習を行い調整した値である。

### 4.1 実験 1 学習高速化の検証

強化学習を実際に適用するためには，学習の高速化が重要である。それには，状態数の削減を行うことが必要となる。状態数が少ないと強化対象が少なくなり，わずかな試行回数で学習が完了できる。また，状態数が少なければ計算機の負荷が少なくなり，プログラムが高速に実行できる。

本研究では，エージェントの周りのマスの見え方（視界）を制限することによって状態数を削減する高速化手法を提案する。視界は，図 6 に示す，エージェントから遠い部分のみをエリア分けしたものである。

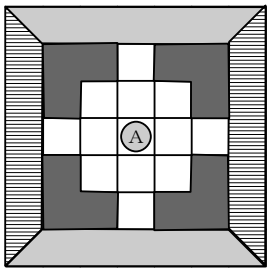


図 6: 提案手法の視界

以前に提案された視界 [1] は，エージェントに近い部分までエリア化したものであった。その視界では，獲物の位置が捕らえづらく最適行動が学習できないという問題点が生じていた。提案手法の視界はその点を考慮し，エージェントの行動決定に重要といえるエージェントの近傍視界を鮮明にしたものである。

#### 4.1.1 実験 1 の概要と結果

実験 1 は獲物を 1 体，獲物の視界は十字方向のみとして，従来視界と提案視界について 200,000 エピソードの学習を行った。

図 7 に，100,000 エピソードまでの学習経過のグラフを示す。横軸がエピソード数，縦軸が 1,000 エピソードごとの平均捕獲ステップ数である。

グラフより，提案手法の方が早く学習が収束しており，更に捕獲ステップ数の平均値も若干低くなっている。

次に，表 1 に 200,000 エピソード学習を行った際の最終結果を示す。学習時間は学習終了までにかかった実時間，捕獲ステップは最終 1,000 エピソードの捕獲ステップ数の平均，ステップ分散は最終 1,000 エピソードの捕獲ステップ数の最大値と最低値の差である。ま

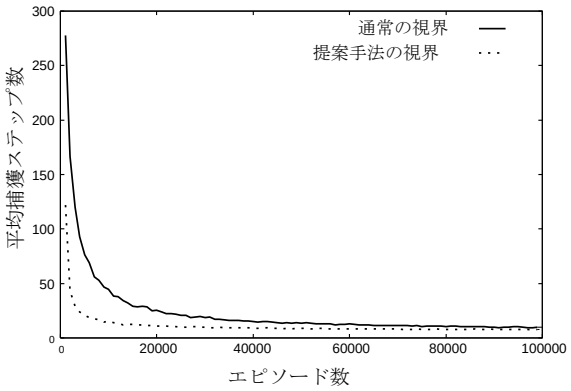


図 7: 実験 1 の学習経過グラフ

た，表 1 の値は，15 回の学習の平均値となっている。

表 1: 実験 1 の最終学習データ (200,000 エピソード後)

| | 学習時間 | 捕獲ステップ | ステップ分散 |
|---|---|---|---|
| 従来 | 66.60 秒 | 7.429 | 37.4 |
| 提案 | 28.65 秒 | 6.825 | 35.2 |

これを見ると，提案手法はすべての値が低くなっており，計算機にかかる負担，学習速度，学習精度全てにおいて優秀な結果となっている。

しかし提案手法では，学習後期でまれに分散が高くなる場合が見られることがあり，若干ではあるが通常の視界にくらべて安定性が低いところも見られた。

また，実験 1 において「エージェント 2 体が獲物の死角に待機し，残り 1 体が獲物の視界に入って獲物を誘導する」という協調行動が見られた。この行動を,「追い込み」と定義した。

通常の視界の場合，追い込みによる捕獲は高めの頻度で見られた。更に追い込みは提案手法においても学習していることを確認した。しかし，見られる頻度は通常の視界と比べてやや少なく感じられた。

#### 4.1.2 実験 1 の考察

結果により，提案手法では計算機にかかる負担の軽減や学習の高速化が確認できた。また分散も低く，学習の精度が上がっていることも分かった。

しかし，学習後期になっても分散が高くなる場合が見られ，更に追い込みの頻度が若干少なく，協調行動が通常の視界に比べて学習されていないことが見られた。これは，提案手法の場合は遠方の物体はどこにいても同様の位置にいると認識されるため，学習状況によって最適な行動を行わない場合があることが原因と考えられる。

この問題点を解決するためには，視界を更に細かく分けることが考えられる。この方法では状態数が増加するため，実行速度や学習速度はやや低下するが，問題の解決は期待される。

しかし，分散が高くなる現象はそれほど高い頻度で見られるわけではなかった。また，協調行動の学習頻度が下がってしまったとはいえ平均捕獲ステップ数が劣っているわけではなく，学習の精度は落ちていない。

### 4.2 実験 2 獲物の増加による影響の調査

従来の追跡問題は，ほぼ獲物を 1 体のみとして行っていた [1][4]。本研究では，獲物を 2 体に増加し，環境が変化した場合についても学習を行い，それが学習に与える影響を調査し，強化学習の環境変化への対応性を調べる。

#### 4.2.1 実験 2 の概要と結果

実験 2 ではエージェントの視界は提案視界，獲物の視界は十字方向のみとし，獲物が 1 体と 2 体の場合について 200,000 エピソードの学習を行った。

図 8 に，実験 2 の学習経過のグラフを示す。これは，100,000 エピソードまでの学習経過となる。

横軸，縦軸ともに図 7 のものと同様で，それぞれエピソード数と 1,000 エピソード毎の平均捕獲ステップ数である。

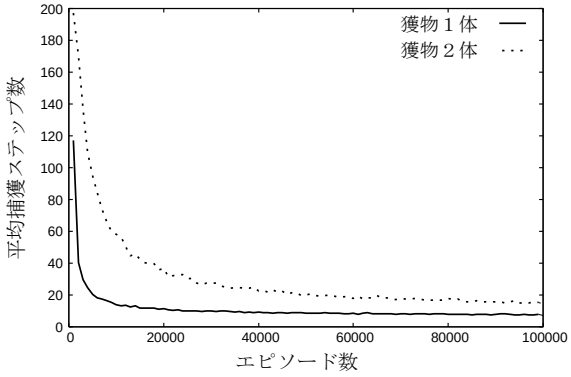


図 8: 実験 2 の学習経過グラフ

グラフからわかる通り，獲物が 2 体の場合は学習の収束が遅い。更に，明らかに平均捕獲ステップ数が獲物 1 体の場合より多くなっている。

次に，表 2 に，200,000 エピソード学習を行った際の，最終結果を示す。

捕獲ステップ，ステップ分散の意味に関しては表 1

表 2: 実験 2 の最終学習データ（200,000 エピソード後）

| | 学習時間 | 捕獲ステップ | ステップ分散 |
|---|---|---|---|
| 1 体 | 28.65 秒 | 6.825 | 35.2 |
| 2 体 | 113.32 秒 | 11.867 | 82.0 |

と同様である。また，15 回の学習データの平均という点も同様である。

これより，獲物 2 体の場合は，全ての要素が獲物 1 体の場合より大きくなってしまっている。

これは獲物が 2 体の際は，１体の場合と比べ状態数が増大しているためといえる。状態数が増えると，強化すべき対象が増えるため，学習回数がかかる。更に，保持する情報も増えるため，実行時間がかかってしまう。また，未強化の状態が多いため捕獲ステップの分散も高くなると考えられる。

また，獲物を 2 体にした場合も実験 1 で見られた追い込みを行っていることを確認した。更に，獲物が 2 体の際の新たな協調行動として,「複数の獲物に惑わされないよう，より多くのエージェントに接近されている獲物を追う」というものが見られた。この行動を「集中」と定義する。

集中は，実験 1 の追い込みと比べると，見られる頻度はあまり高くなかった。学習がまだ完全に収束していないことが原因と思われる。

#### 4.2.2 実験 2 の考察

実験 2 では，集中という獲物を増やした際に特有の行動を確認することができたが，あまり高い頻度で見られたわけではなかった。また，本来は獲物を増やすことにより捕獲対象が増え，平均捕獲ステップ数は獲物が 2 体の方が少なくなるはずである。

これは，先に述べたように状態数の増加により学習が十分に進んでいなかったことが原因と考えられる。学習のエピソード数を増加することで，より学習が進みこれらの問題点を解決できるが，学習完了までに時間がかかるためにあまり効率的ではない。

この問題を解消するために，まず獲物が 1 体の場合において学習をし，その学習結果を反映させた状態で獲物が 2 体の場合の学習を開始する方法が考えられる。また，初めに狭い学習空間で学習を進めその結果を利用するという考えもある。

しかし，集中は獲物を増やした際に有効な行動といえ，環境が変化してもそれに対する有効な行動を獲得できるということは確認できた。

### 4.3 獲物の視界拡大による影響の調査

従来の研究では，追跡問題の獲物は十字方向などのある方向のみに限定された視界で学習を進めることが多く [1][4]，エージェントは追い込みなど獲物の死角を利用した協調行動を学習していた。そこで実験 3 では，獲物の視界を空間全範囲に広げた場合にも学習を行い，学習への影響を調査し，強化学習の環境変化への対応性を調べる。

#### 4.3.1 実験 3 の概要と結果

実験 3 は獲物を 1 体，エージェントの視界は従来視界にとし，獲物の視界を十字方向のみと空間全範囲のそれぞれに設定して 200,000 エピソードの学習を行った。

図 9 に，実験 3 の学習経過のグラフを示す。これまで示してきた図 7 や図 8 のグラフ軸の意味は同様である。また，100,000 回の学習データであることも同様である。

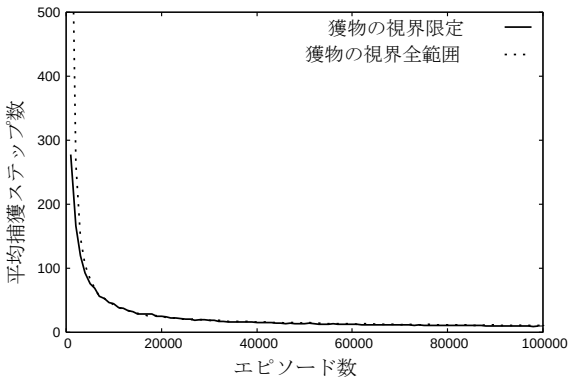


図 9: 実験 3 の学習経過グラフ

グラフより，獲物の視界が空間全範囲の場合は，学習初期の平均捕獲ステップ数が非常に高くなっていることがわかる。これは，学習初期のエージェントの行動はほぼランダムであることに対し，獲物の逃走にはほぼランダム要素がないため，捕獲する可能性が低いためと考えられる。

しかし，収束の速さや収束した際の平均捕獲ステップ数には大きな差はない。これは，どちらの条件でも状態数は等しく，学習の早さに開きがないためといえる。

次に，200,000 エピソード学習を行った際の最終結果を表 3 に示す。各要素の意味は表 1，表 2 と同様で，表 3 も 15 回の学習の平均である。

特徴的なのは，学習時間の差が大きくなっていることである。これは，前述した学習初期の捕獲の難解さが原因と見られる。

表 3: 実験 3 の最終学習データ (200,000 エピソード後)

| | 学習時間 | 捕獲ステップ | ステップ分散 |
|---|---|---|---|
| 十字 | 66.60 秒 | 7.429 | 37.4 |
| 全範囲 | 212.19 秒 | 8.735 | 35.7 |

また，実験 2 より両手法の平均捕獲ステップの差が少なくなっている。これも前述のとおり学習の早さにさほど差がないことが原因といえる。

捕獲ステップ分散に関しては，獲物視界が空間全範囲である方が少なくなっていた。これは逃走アルゴリズムにランダム要素が少なく，エージェントにとって獲物の行動が予想しやすいからと考えられる。

協調行動においては，獲物の視界を空間全範囲に広げた場合は死角を利用できず，これまでの実験で見られた追い込みは使えない。その代替として,「エージェント 3 体が，獲物とほぼ同じ距離を保ちながら，別方向から接近していく」という行動が確認された。この行動を「包囲」と定義した。

包囲はこれまでの協調行動と比べて非常に高い頻度で行われ，捕獲する際にはほぼ包囲の行動を取っていた。

#### 4.3.2 実験 3 の考察

学習経過としては，学習初期に時間がかかっても状態数が同じであれば，いずれは同じような値に収束することが確認でき，状態数の重要さを再認識させられる結果となった。

また，包囲という追い込みの代替となる行動が確認できた。包囲は獲物に死角がない場合に有効な行動であり，それはエージェントが高い頻度で行うことからもわかる。この包囲行動の獲得により，環境変化への対応性が確認できたといえるだろう。

## 5 今後の課題

本研究の目標の 1 つとして，最適行動を学習できる高速化手法を提案することがあった。

提案手法は通常の視界と比べて学習が早く，計算機にかかる負荷も少ない。若干分散が高くなるような場合も見られ協調行動を学習する頻度もやや低いが，学習精度自体は落ちていないことも確認できた。

更なる課題としては分散が高くなる，協調行動が学習されにくいといった欠点を解決するためにより効率のよい視界の分け方を調べることが考えられる。また，視界のエリア分け以外の状態数削減方法の考案も考えられる。

もう 1 つの目標は，獲物の条件を変化させることで，強化学習の環境への対応性を調査するというものであった。

実験 2 で獲物の数を増やした場合には集中，実験 3 で獲物の視界を拡大した場合は包囲と，それぞれ条件に適した行動を学んでいることが確認された。よって，強化学習は環境に合わせた行動を学習できるといえる。特に実験 3 の包囲については，かなりの頻度で行っていることが確認できた。

しかし，実験 2 では状態数の増加からさほど協調行動が見られなかった。この場合について実験を進めることで，新たな協調行動が確認される可能性が残されている。4.2.2 節で述べた「獲物が 1 体などの基本的な条件の学習結果を学習の初期値に反映させる」という高速化手法を導入し，新たな協調行動を調査する事が今後の課題として挙げられる。

また，割引率や基本報酬，初期重要度などのパラメータに対して更に微調整を行い，調査結果に関する考察を行う必要もある。

## 参考文献


[1] 奥空 武志.“3 次元空間マルチエージェントシステムの構築と考察”, 釧路工業高等専門学校卒業論文,(2005)

[2] 三上 貞芳, 皆川 雅章.“強化学習”, 森北出版 (2000)

[3] 荒井 幸代. “マルチエージェント強化学習 -実用化に向けての課題・理論・諸技術との融合-”, 人工知能学会誌,Vol.16,No.4,pp.476-481 (2001)

[4] 荒井 幸代. “マルチエージェント強化学習の方法論 -Q-Learning と Profit Sharing による接近-”, 人工知能学会誌,Vol.13,No.4,pp.609-618 (1998)

[5] 宮崎 和光. “Profit Sharing に基づく強化学習の理論と応用”, 人工知能学会誌,Vol.14,No.5,pp.800-807 (1999)