

Profit Sharing 法を用いた追跡問題における 視野領域を考慮した学習方法の提案

澤田 志門^{*1} 奥山 淳^{*2}

Proposal of a Profit Sharing Method for the Pursuit Problem Considering Field of View

by

Shimon SAWADA^{*1} and Atsushi OKUYAMA^{*2}

(Received on Mar. 31, 2023 and accepted on May. 25, 2023)

Abstract

Multi-agent reinforcement learning (MARL) has received considerable attention in recent years. MARL has often been applied to the pursuit problem, in which multiple hunters try to catch single or multiple prey. In this problem, hunters are usually considered to have an infinite range of field of view (FoV) to obtain the absolute or relative positions of the prey as the hunters' states. However, this is inappropriate owing to "the curse of dimensions" and the range of FoV should be restricted suitably. On the other hand, strict restrictions on the range of FoV would prevent hunters from observing their states and continuing learning because they cannot obtain the prey's position when it is outside their FoV. Therefore, in this study, numerical simulations were performed to: (1) propose a method to continue to learn when the prey is outside the hunter's FoV, and (2) evaluate the appropriate restriction of the range of FoV for a size of field. The simulation results validated the effectiveness of the proposed method and revealed the relationships between the range of the hunter's FoV and the size of the field.

Keywords: Multi agent system, Reinforcement learning, Profit sharing, Pursuit problem

1. 緒言

近年、マルチエージェントシステム (MAS: Multi Agent System) に関する研究が広く行われている。MAS は複数の自律的に行動する主体 (エージェント) が協調して動作することで目標の達成を試みるもので、作業効率や耐故障性やコストパフォーマンスに優れるとされている¹⁾。

この MAS の中でも特に、MAS へ強化学習を適用してその有効性を検証するマルチエージェント強化学習 (MARL: Multi Agent Reinforcement Learning)^{2,3)}の検討が活発である。例えば荒井らは追跡問題 (Pursuit Problem) に対して Q 学習と Profit Sharing 法 (以下, PS) を適用し、MARL においては PS の方が有効であることを示した⁴⁾。ここで、追跡問題は MAS における典型的な例題の 1 つであり、2 次元のセル空間上を移動する複数のハンターエージェント (以下, ハンター) および複数もしくは単数の獲物エージェント (以下, 獲物) を対象とする。また、ハンターが獲物を捕獲することを目的とし、MARL ではその捕獲戦略の学習を試みる。

ただし、従来は学習に用いる状態を明言せず、暗黙の了解として絶対座標系に基づいた獲物やハンターの位置

座標を状態とすることが多い⁴⁻⁸⁾。他方、状態として任意のハンターを基準とした相対座標系における、獲物や他のハンターとの相対位置を採用する検討もいくつか存在する⁹⁻¹⁴⁾が、これらはどれもハンターが無限大の視野領域を持っており、自分からみた他のハンターおよび獲物との相対位置がわかることを条件としている。しかし、このように無限大の広さの視野領域をもつことは実ロボットへの適用を考えると適切ではなく、また RL および MARL の課題の 1 つである次元の呪いを生じうる。このことからハンターの視野領域を適切に制限する必要があるが、過剰に制限すると状態の取得が行えず、学習へ悪影響を及ぼす可能性もある。このため、フィールドの大きさに応じて適切な視野領域を設計する必要があるが、そのような検討は少ない。

また、視野領域を制限した場合には獲物が視野領域外に存在することがあるが、その場合ハンターは自身と獲物との相対座標を取得できず、学習における「状態」を取得することができない。この場合学習を継続することができなくなるため、視野領域外に存在する獲物も「状態」として考慮することが必要になる。しかし、そのような手法についての検討は筆者の知る限り存在しない。

そこで本研究は限られた視野領域を持つハンターによる追跡問題を題材とする。また追跡問題においては、視

^{*1} 総合理工学研究科総合理工学専攻博士課程

^{*2} 工学部精密工学科教授

視野領域内の獲物の相対座標を状態として用いた PS による捕獲戦略の学習を試みる。加えて、獲物が視野領域外にいても学習を行えるような手法を 2 通り提案し、視野領域の広さを変更したシミュレーション検討を行うことで、フィールドの大きさと視野領域の広さの関係性および視野領域外の獲物の扱い方について考察する。

2. 強化学習

2.1 概要

強化学習は機械学習の 1 つであり、学習主体が試行錯誤的に行動を繰り返すことで適切な行動を学習していく手法である。なお、これ以降「強化学習によって学習を行う、自律的に行動する主体」をエージェントとする。

ここで、本研究ではエージェントが観測できる状態の集合を $\mathcal{S}$ 、実行できる行動の集合を $\mathcal{A}$ とする。ただし、行動の集合 $\mathcal{A}$ は状態の関数であるものとし、ある状態 $s \in \mathcal{S}$ において実行できる行動の集合を $\mathcal{A}(s)$ とする。また離散的な時間ステップを表す変数を t ($t = 1, 2, \dots$) とし、時刻 t における状態と行動をそれぞれ $s(t) \in \mathcal{S}$ 、 $a(t) \in \mathcal{A}(s(t))$ とする。ここで、エージェントは次の 5 つのステップを繰り返すことで学習を行う： (1) エージェントは時刻 t において、自身の状態 $s(t) = s_t$ ($s_t \in \mathcal{S}$) を環境から知覚する。 (2) エージェントは知覚した状態 s_t を元に、自身の方策 $\pi(s_t, a)$ から実行する行動 $a(t) = a_j$ ($a_j \in \mathcal{A}(s_t)$) を決定する。 (3) エージェントが (2) で決定した行動 a_j を行うことで、環境が変化して状態が $s(t+1)$ へ遷移する。 (4) エージェントが環境から報酬 $R(t+1)$ を受け取る。 (5) $t+1 \rightarrow t$ として (1) に戻る。

また、強化学習は Q 学習などの環境同定型 (Exploration-oriented method) と、PS などの経験強化型 (Exploitation-oriented method) に大別され、学習に要する時間や必要な条件などに違いがある¹⁵⁾。ここで、Q 学習は特定の条件下では最終的に最適な政策が得られることが証明されているが、その条件として状態遷移がマルコフ過程 (MDP : Markov Decision Process) であることを仮定している。しかし、MARL では複数のエージェントが同時に行動を行うために状態遷移が MDP にならず、MARL には適さない。

他方、PS は最終的に得られる政策が最適である保証がないものの、状態遷移が MDP でなくても学習を行うことができる。荒井らは追跡問題に Q 学習と PS をそれぞれ適用し、学習結果の比較から MARL に対しては PS の方が有効であることを示している⁴⁾。このことから MARL の検討では PS が多く用いられており、本研究でも PS を適用する。

最後に、強化学習の課題の 1 つに次元の呪いがある。これは状態数の爆発とも呼ばれ、状態数がきわめて多くなってしまうことによって学習が適切にできなくなるという課題である。MARL でも、エージェント数の増加が状態数の指数関数的な増加を引き起こしてしまうため、課題の 1 つとされている^{2,3)}。今、エージェントが自身を中心とした $m \times m$ マス ($m \in \mathbb{N}$) を知覚できる場合、視野領域の総マス数は m^2 となる。また視野領域内にいる他の

エージェントを知覚して状態とする場合、エージェントの総数を n ($n \in \mathbb{N}$) とすれば状態の総数が $(m^2)^n = m^{2n}$ となる。ここで指数の肩に n があるため、状態の総数はエージェント数の増加に伴って指数関数的に増加してしまう。このことから、これを次元の呪いまたは状態数の爆発と呼ぶ。

2.2 Profit Sharing 法 (PS)^{16,17)}

PS は経験強化型の強化学習手法の 1 つであり、状態と行動に対する重みを更新することで行動の学習を行う。ここで、エージェントの状態 s と行動 a の組を併せてルール $r = \overline{sa}$ とし、学習を開始してから報酬を獲得するまでのルール系列をエピソードと呼ぶ。また、あるエピソードにおいて同一の状態に対して異なる行動が選択されているとき、その間のルール系列を迂回系列と呼ぶ。学習を開始してから現在までのエピソードで常に迂回系列上にあるルールを無効ルール、そうでないルールを有効ルールと呼ぶ。このため、PS はルールに対する重みを強化する学習法ともいえる。

時刻 t における状態 $s(t)$ と行動 $a(t)$ によるルール $r(t) = \overline{s(t)a(t)}$ について、その重みを $w_{r(t)}$ とすると更新式は式 (1) で表される。

$$w_{r(t)} = w_{r(t)} + f(t) \quad (1)$$

ここで、 $f(t)$ は時刻 t における強化関数の値である。この強化関数によって重みを更新し、行動の学習を行う。

宮崎らは、無効ルールの学習を抑制するような強化関数の条件として合理性定理を提案している¹⁷⁾。式 (2) に合理性定理を示す。1 エピソードにおけるステップ数の上限を W とし、時刻 $t = W$ で報酬を獲得したとする。無効ルールを抑制して学習を行うには、時刻 $t = 1, 2, \dots, W$ において強化関数 $f(t)$ が式 (2) を満たせばよい。なお、各状態における有効ルールの最大個数を L とした。

$$L \sum_{\tau=t}^W f(\tau) < f(t-1) \quad (2)$$

この合理性定理を満たす強化関数として、式 (3) が提案されている¹⁷⁾。式 (3) において $U \geq L+1$ であり、本研究では行動集合の要素数 $|\mathcal{A}|$ を用いて $U = |\mathcal{A}|$ とした。ここで詳細は 3.1 節で述べるが、本研究ではすべての状態においてエージェントが最大で 5 つの行動を選択可能である。このためある状態 s において選択可能な行動の集合 $\mathcal{A}(s)$ について $|\mathcal{A}(s)| \leq 5$ が成り立つが、式 (3) においてはこの最大値 5 を用いれば強化関数は式 (2) を満たす。よって本研究では $U = 5$ とした。また、 R は報酬値である。

$$f(t) = \begin{cases} R & (t = W) \\ \frac{1}{U} f(t+1) & (t = 1, 2, \dots, W-1) \end{cases} \quad (3)$$

また、本研究ではハンターがルーレット選択を用いて行動を選択するものとする。今、時刻 t における状態 $s(t) = s_c$ において、エージェントが行動 $a(t) = a_k$ ($a_k \in \mathcal{A}(s_c)$)を選択する方策 $\pi(s_c, a_k)$ は、ルール $r_{ck} = \overline{s_c} a_k$ に対する重み w_{ck} を用いて式(4)で表される。なお w_{ch} はルール $r_{ch} = \overline{s_c} a_h$ ($a_h \in \mathcal{A}(s_c)$)に対する重みである。

$$\pi(s_c, a_k) = \frac{w_{ck}}{\sum_{a_h \in \mathcal{A}(s_c)} w_{ch}} \quad (4)$$

3. 追跡問題

3.1 基本的な問題設定

追跡問題は MAS の基本的な例題の 1 つであり、複数のハンターによって単数もしくは複数の獲物を捕獲することを目的とする。本研究では従来研究⁴⁻¹⁴⁾を参考に、以下の設定とする。

まず、フィールドは $p \times p$ マスのセル状であるとし、壁のないトーラス形状であるとする。ここで、 $p \times p$ マスのフィールドの大きさを p で表す。さらに本研究では 2 台のハンターと 1 台の獲物が存在し、どちらも上下左右マスへの移動およびその場での静止という 5 つの行動を行えるものとする。ただしハンターと獲物はそれぞれ固有の視野領域をもつ。すべてのハンターと獲物は各エピソードの開始時に重複しないランダムなセルへ配置され、そこから 1 ステップごとに 1 つの行動を選んで一斉に実行するものとする。

また、本研究では獲物の上下左右マスのうち、いずれか 2 マスに全ハンターが位置すれば捕獲を達成した、として全ハンターへ等しい大きさの報酬を与える。

さらに、本研究ではフィールドがトーラス形状であるため移動や視野領域に回り込みが生じる。Fig. 1 を用いてこれの詳細を述べる。ここで、Fig. 1 はあるタイミングにおけるハンターおよび獲物の位置関係の一例である。Fig. 1 はフィールドの大きさが $p = 5$ であり、獲物を四角形、ハンターを三角形およびひし形で表している。また、Fig. 1 ではハンターの視野領域を 3×3 としており、塗りつぶしたマスがハンター1の視野領域を表している。このとき、フィールド形状からハンター1の視野領域は下方が反対側へ回り込むようになる。また詳細は後述するが、獲物はこのとき上下左マスへの移動または静止の 4 つからランダムに行動を選択する。もし左または下へ移動する場合、獲物はその反対側である黒い星型のあるマスへ回り込むように移動を行う。

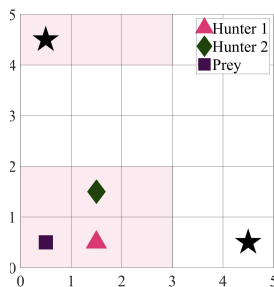


Fig. 1 An example of positions of predators and a prey.

3.2 獲物の詳細な定義

獲物は自分を中心とした 3×3 マスの視野領域をもち、視野領域内のハンターから逃避的な行動を取るものとする。例えば Fig. 1 の場合、獲物はハンター1 と重なることを避けるため、右マスを除いた残りの上・下・左マスへの移動およびその場での静止という 4 つの行動から等確率に行動を選択する。なお、視野領域内にハンターが存在しない場合は移動するマスへの制限がないため、上下左右マスへの移動および静止という 5 つの行動から 1 つを等確率で選択する。

3.3 ハンターの詳細な定義

ハンターは自分を中心とした $q \times q$ マスの視野領域をもち、その中にいる獲物および他のハンターの相対位置を知覚できる。また、自身を基準にした獲物の相対位置を学習における状態として用いる。ここで、 $q \times q$ マスの視野領域の広さを q で表す。Fig. 1 の場合は $q = 3$ であり、ハンター1 は自身の左マスに獲物が、上マスにハンター2 がいることを知覚できる。またハンター2 は自身の左下マスに獲物が、下マスにハンター1 がいることを知覚できる。

ここで、各ハンターは獲物のいるマスへの移動を行わないため、獲物と重複しないマスのみを候補とし、方策に従って行動を選択する。これの詳細を Fig. 1 のハンター1 を例として述べる。今、ハンター1 の上下左右マスへの移動および静止の行動を順に a_U, a_D, a_L, a_R, a_S とし、ハンター1 を基準にした獲物の相対位置を学習における状態とする。このとき獲物がハンター1 の 1 マス左に位置しているため、状態を s_L で表すこととする。また s_L におけるルール $r_h = \overline{s_L} a_h$ ($h = U, D, L, R, S$)に対する重みを w_h とする。このときハンター1 は左マスに移動すると獲物と重なってしまうため、残りの a_U, a_D, a_R, a_S のうちから行動を選択する。つまり、状態 s_L において実行可能な行動の集合は $\mathcal{A}(s_L) = \{a_U, a_D, a_R, a_S\}$ となる。このときハンターは式(4)に示すルーレット選択によって行動を選択するため、例えば上マスへの移動である行動 a_U を選択する確率 $\pi(s_L, a_U)$ は次式で求められる。ここで実行できない行動 a_L は集合 $\mathcal{A}(s_L)$ に含まれておらず、その重み w_L は行動選択に用いないことに注意が必要である。

$$\pi(s_L, a_U) = \frac{w_U}{w_U + w_D + w_R + w_S}$$

ただし、ハンター2 においては獲物が自身の上下左右には存在しないため、上下左右マスへの移動および静止の 5 つから方策ののって行動を選択する。一方、もし獲物が上への移動、ハンター2 が左への移動を選択すると同じマスへ移動することになる。このとき、本研究ではこのような場合の例外処理は行わない。つまり、ハンター2 と獲物は重複するマスへ移動するが、その後獲物は静止以外の 4 つから行動を選択し、ハンター2 は獲物が同じマスにいるという状態を認識するものとする。これは、Fig. 1 においてハンターの視野領域は 3×3 であるため合計 9 マスが認識できる一方、ハンター・獲物の

行動は上下左右および静止の 5 つであり、視野領域のうち斜めの 4 方向はハンター・獲物の行動を制限しないためである。

なお、獲物と重複するマスへの移動については従来研究でも「重複したマスへ移動しようとした場合はその場で静止させる」とするものがある。しかしその場合は本来ハンターが選択した行動を実行させておらず、必要以上に静止行動の重みが強化されてしまい学習に悪影響を与えると考えた。そこで、本研究ではハンターが初めから獲物と重なるマスへは移動せず、獲物と重複しないよう行動する中で学習を行うものと定義した。

3.4 従来研究での追跡問題における課題

本研究では、追跡問題における視野領域の広さに注目する。1 章および 3.1 節でも述べたように、追跡問題は MAS における典型的な問題の 1 つであり、MARL の研究でも広く用いられている。しかし、MARL の従来研究の多くはハンターが無限大の広さの視野領域をもっており、他のハンターや獲物の位置を常に正確に把握することができるとしている。これは、視野領域に制限を設けた場合ハンターが視野領域外にいる獲物を認識できず、学習する際に「状態」が不明になってしまうためである。したがって、従来研究の学習方式では獲物が視野領域外になった瞬間から学習を継続することができなくなる。

この解消を図るため、本研究では獲物が視野領域外に存在する場合にもハンターが学習を継続できる手法を提案する。

4. 提案手法

4.1 概要

本研究では、獲物がハンターの視野領域外にいる場合にも、ハンターが学習を継続できる手法を提案する。獲物がハンターの視野領域外に存在する場合、ハンターの視野領域内には獲物がいなくなるため、ハンターは自身からみた獲物の相対位置を取得することができない。

このため、本研究では視野領域をいくつか分割し、視野領域外の獲物はその分割した領域のいずれかにいるものとして扱う。この視野領域を分割した領域を写像領域と呼び、「視野領域外に存在する獲物がいずれかの写像領域へいるものとして扱う」ことを「獲物の位置を任意の写像領域へ写像する」と表現する。

しかし、視野領域の分割方法には自由度が存在するため、本研究ではこの分割方法を 2 つ提案し、効果的な学習を行うことができる分割方法についてシミュレーション検討の結果から考察する。

4.2 提案手法 1

提案手法 1 は、写像領域として視野領域の外周部のみを分割したものをを用いる手法である。詳細を Fig. 2 に示す。Fig. 2 ではフィールドの大きさを $p = 7$ 、視野領域の広さを $q = 5$ とし、ハンターを三角形で示している。また、写像領域を分割ごとに塗りつぶし、対応する番号を振っている。本手法では視野領域の外周部のみ上下左右と斜

めの 8 領域に分割して写像領域を定義している。このとき、詳細は 4.4 節で述べるが各写像領域を構成する視野領域内のマス全てに対してルール重みの強化が行われる。

本研究では式(3)で定義される強化関数によってルール重みの強化を行うが、その値は学習開始直後のステップが最も小さく、報酬を獲得したステップで最も大きい。このとき視野領域の外周部分はハンターから $(q-1)/2$ マスだけ離れているため、「最後に獲物が外周部分にいたステップ数」は視野領域広さ q が大きくなるほど「報酬獲得ステップ」から遠ざかる。このため視野領域の外周部分のルール重みを強化する強化関数の値は q が大きくなるにつれて小さくなるため、本手法によって視野領域の外周部分を多く強化すれば効果的な学習が実現できると考えた。なお、この分割方法は渡邊ら¹³⁾における視野の制限方法を参考に決定した。

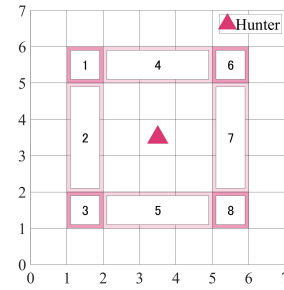


Fig. 2 Correspondence between Mapping area and hunter's FoV in proposal method 1.

4.3 提案手法 2

提案手法 2 は写像領域として視野領域内を十字に分割したものをを用いる手法である。この詳細を Fig. 3 に示す。Fig. 2 と同様に $p = 7$ 、 $q = 5$ とし、ハンターを三角形で示した。また Fig. 2 と同様、写像領域を分割ごとに塗りつぶし、番号を振っている。Fig. 3 のようにハンターの上下左右マスおよびそれ以外の 8 つに視野領域を分割し、写像領域としてこの 8 領域を用いる。このとき、詳細は 4.4 節で述べるが提案手法 1 に比べて視野領域内のより多くのマスを強化できるため、視野領域内に獲物が存在する場合に適切な行動を選択することができ、学習に良い影響を及ぼすと考えた。なお、この分割方法は伊藤ら⁹⁾および片山ら¹²⁾を参考に決定した。

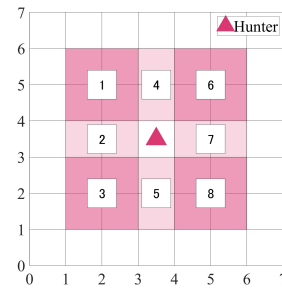


Fig. 3 Correspondence between Mapping area and hunter's FoV in proposal method 2.

4.4 提案手法の適用方法

本研究では 2 つの提案手法で視野領域を分割し、写像

領域のいずれかに視野外の獲物が存在するものとして推定（写像）する．ただし、どの領域へ写像されるかはランダムに決定し、それぞれの領域へ等確率に、12.5%の確率で写像されるものとする．

加えて、各領域を選択した場合に学習をどのように行うか述べる．今、Fig. 4 のようにハンター・獲物が存在し、ハンターは提案手法 2 を用いるとする．また $p = 7$ 、 $q = 5$ であり、ハンター・獲物をそれぞれ三角形・四角形で示した．さらに、ハンターは視野領域内にいる獲物の相対的な位置を学習の状態として用いるが、ハンターが認識する相対位置と「状態」の番号を対応付けて Fig. 4 に示した．例えば獲物がハンターの右方 1 マス、上方 2 マスに存在した場合、Fig. 4 よりハンターの状態は状態 15 となる．ハンターと獲物が Fig. 4 の位置関係にあるとき、獲物はハンターの視野領域外に存在するため、状態の番号は取得できない．そこで写像領域を 1 つランダムに選択するが、ここでは赤い線で囲った写像領域 1 が選ばれたものとする．このとき、写像領域 1 は Fig. 3、4 より状態 1, 2, 6, 7 から構成されているため、式(1)を用いた重みの更新をこの 4 状態に対して行う．ここで、このように視野領域外に獲物がある場合は各領域を構成するすべてのマスに対して重みの更新を行うのに対し、視野領域内に獲物がある場合は獲物が位置するマスが正確に取得できるため、その特定のマスに対して重みの更新を行うことに注意が必要である．

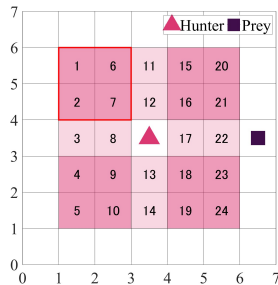


Fig. 4 An example of how to reinforce a Hunter.

続いて、提案手法におけるハンターの行動選択手法について述べる．ハンターは式(4)に示すルーレット選択によって行動を選択するが、獲物が視野領域内にいる場合は獲物の位置するマスが正確にわかる．したがってそのマスにおけるルール重みの値から式(4)を用いて行動を選択する．他方、獲物が視野領域外にいる場合は前述のように写像領域がランダムに選択されるため、「獲物の位置するマスのルール重み」を用いることができない．このとき、ハンターは写像領域を構成するマスのルール重みを平均した値を用いて式(4)からルーレット選択を行い、行動を選択する．例えば Fig. 4 における写像領域 1 が選択された場合、写像領域 1 は状態 1, 2, 6, 7 から構成されるためこれらの状態におけるルール重みの値を平均し、その平均値を用いてルーレット選択から行動を選択する．

最後に、視野領域の大きさの制限について述べる．本研究では Fig. 2, 3 のように視野領域の分割方法を 2 つ提

案するが、視野領域の広さ q は $q \geq 5$ とする．これは Fig. 5 のように、 $q = 3$ であるときは視野領域が 2 つの提案手法で同じように分割されてしまうためである．なお、Fig. 5 は $p = 5$ かつ $q = 3$ とし、ハンターを三角形で示した．

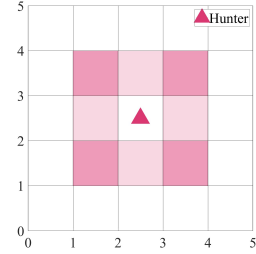


Fig. 5 Divided hunter's FoV area when $q = 3$.

5. シミュレーション条件と結果および考察

5.1 シミュレーションの環境と条件

本研究では数値計算用ソフトウェアである MATLAB 2022a を用い、 $p = 11$ のフィールドを対象としてシミュレーション検討を行った．各シミュレーションにおけるパラメータを Table 1 に示す．Table 1 より報酬値は 1000 であり、これは獲物の捕獲達成時に各ハンターへ与えられる値である．またこの値を用いて式(3)から強化関数の値が導出され、式(1)によってルール重みの強化が行われる．

Table 1 Learning conditions.

Field size	11 × 11
Size of hunter's FoV	5, 7, 9, 11
Number of episodes	3.0×10^5
Maximum number of steps in each episode	100
Reward value	1.0×10^3
Initial value of rule weight	200

5.2 学習の継続性からみた提案手法の有効性の検証

本節では獲物がハンターの視野範囲外に出てしまうエピソード数およびステップ数について述べる．3.4 節でも述べたように、従来手法ではハンターが獲物の視野領域外に出てしまうとハンターは状態を取得できず、学習を継続できない．このため、本研究ではそうした状況でも学習を継続できるような手法を提案する．学習の継続性という観点から提案手法の有効性を検証するため、学習の中で獲物がハンターの視野範囲外に出てしまうエピソード数およびステップ数を調べた．

ここで、本研究ではハンター・獲物の初期位置や行動選択などにランダム性が関係するが、このランダム性は乱数の種によって制御される．そこで、本研究ではこの乱数の種を 6 つ設定し、それぞれにおいて視野領域の広さを $q = 5, 7, 9, 11$ と変化させて追跡問題の学習を試みる．

各乱数の種を用いた追跡問題において、視野領域の広さを $q = 5, 7, 9$ としたとき、獲物がいずれかのハンターの視野領域外に出たエピソード数とステップ数を Table 2 に示す．なお、Table 2 では各乱数の種および視野領域の広さ q に対して、視野領域外に出たエピソード数 $n_{Episode}$ 、

ステップ数 n_{Step} を $(n_{Episode}, n_{Step})$ として示した．例として，Table 2 において 1 番目の乱数の種に対して視野領域の広さ q が $q = 9$ であるとき，値が $(2, 1)$ となっている．これは，この条件では 2 エピソード目の 1 ステップ目に獲物がいずれかのハンターの視野領域外に出てしまうことを意味する．なお，1 エピソード目の 1 ステップ目で視野領域外に出ていない表の項目を薄く塗りつぶしてある．また，Table 2 に示した各乱数の種の値は筆者が任意に定めた．

Table 2 Number of episodes and steps when the prey is outside of the field of view of any Hunters.

Number	Value	$q = 5$	$q = 7$	$q = 9$
1	1109	(1,1)	(1,1)	(2,1)
2	31	(1,1)	(1,1)	(1,5)
3	323	(1,1)	(1,1)	(1,1)
4	11	(1,1)	(1,1)	(1,1)
5	9	(1,1)	(1,1)	(1,1)
6	1205	(1,1)	(1,1)	(1,1)

Table 2 より，すべての乱数の種において $q = 5, 7$ の場合は 1 エピソード目の 1 ステップ目に獲物がいずれかのハンターの視野領域外に位置していることがわかる．特に 3・4・5・6 番目の乱数の種では $q = 9$ の場合にも同様であり，1・2 番目の乱数の種もそれぞれ 2 エピソード目の 1 ステップ目，1 エピソード目の 5 ステップ目に獲物が視野領域外に位置している．

つまり，すべての乱数の種において少なくとも 2 エピソード目以降はハンターが学習における状態を取得できないため，従来手法を用いた場合に捕獲戦略の学習ができなくなることがわかる．他方，提案手法を用いればこうして視野領域外に出た獲物についても学習に用いることが可能であるため，Table 1 に示した全 30 万エピソードによる学習が可能である．

5.3 シミュレーション結果・考察

次に，提案手法による学習性能についての検討を行う．

結果を述べる前に，まず本研究では 3.1 節および 5.2 節でも述べたようにエピソードごとに獲物・ハンターの初期位置がランダムに設定される．このため，エピソードによっては初期位置の影響で捕獲が容易または困難になることがある．このような初期位置への依存性をなくすため，全 30 万エピソードを 1000 エピソードごとに分割し，その分割ごとに捕獲までのステップ数の平均を取った．以降，1000 エピソードごとに平均を取った，1 エピソードにおける捕獲までのステップ数を単に「捕獲ステップ数」と呼ぶ．

ここで，Fig. 6 (a)~(f)はそれぞれ 1~6 番目の乱数の種を用いて追跡問題の学習を行った際の，各視野領域の広さ・提案手法における捕獲ステップ数の推移である．それぞれ横軸がエピソード数 ($\times 1000$)，縦軸が捕獲ステップ数を表す．また，各図の凡例として，視野領域の広さ q の値と共に提案手法 1 を用いたものを A，提案手法 2 を

用いたものを C とした．例えば Fig. 6 (a) の中央に存在する紫色で三角形のマーカーを付けたものは凡例が $q = 7C$ となっているが，これは視野領域の広さを $q = 7$ として，提案手法 2 を用いた時の結果であることを意味している．なお， $q = p$ であるときはフィールドを完全に知覚でき，提案手法の影響を受けないため，単に $q = p$ とした．

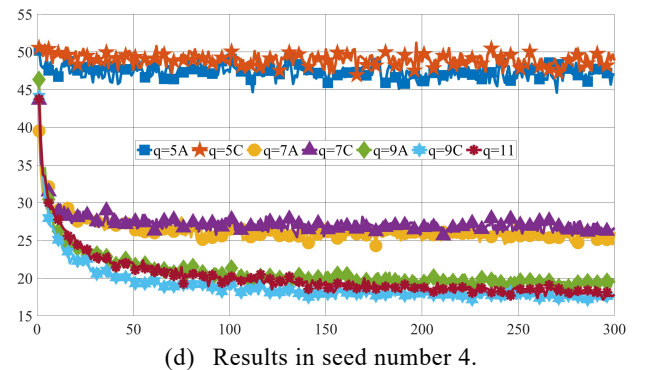
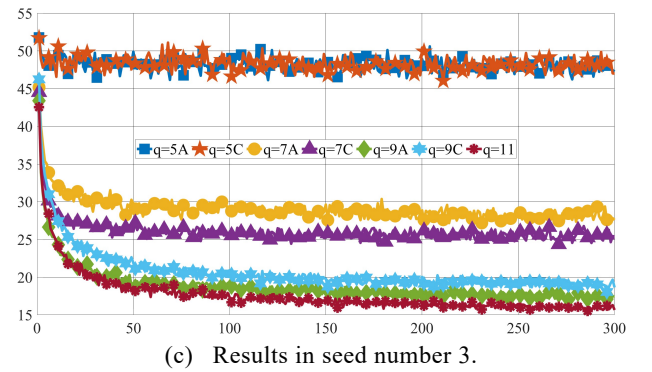
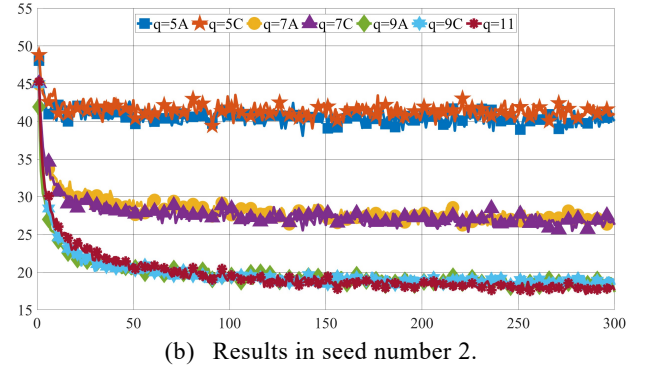
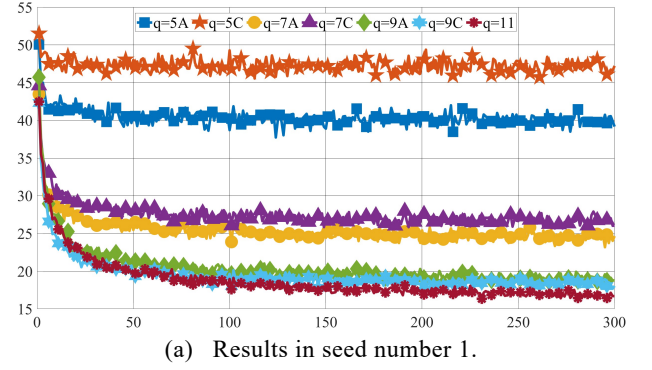
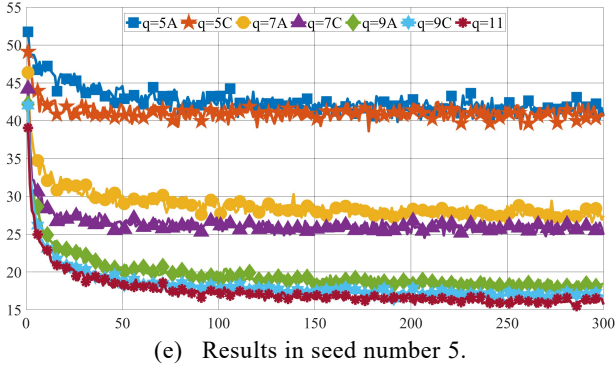
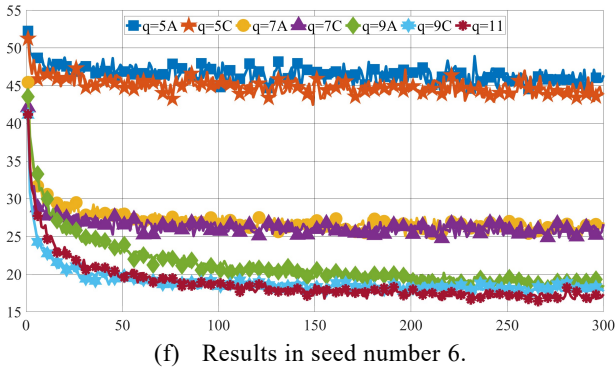


Fig. 6 Number of averaged captured steps in each FoV.



(e) Results in seed number 5.



(f) Results in seed number 6.

Fig. 6 Number of averaged captured steps in each FoV.

Fig. 6 より、いずれの乱数の種を用いた場合も視野領域の広さ q がフィールドの大きさ p に等しいとき、捕獲ステップ数が小さい値に収束していることが確認できる。また、この収束値は約 17 ステップであり、学習を始めたばかりの約 40~45 ステップよりも少ないステップ数で捕獲が達成できている。ただし Fig. 6 (d)より、乱数の種によっては最小値には収束していない。この原因の 1 つとして、学習エピソードの数が不足していることが考えられる。Fig. 6 (d)において、 $q = 9A$, $q = 9C$ の値はどちらも約 150($\times 1000$)エピソード以降でそれぞれおよそ 19, 17 ステップに収束している。他方、 $q = 11$ の値は 150($\times 1000$)エピソード以降も緩やかな減少を続けており、300($\times 1000$)エピソードにおける値が約 17 ステップになっている。ここで $q = 9$ の場合は視野領域を構成するマス数が $q^2 = 81$ マスである一方、 $q = 11$ の場合は $q^2 = 121$ マスとなる。つまり、 $q = 11$ の場合は視野領域を構成するマス数が $q = 9$ の場合より 40 マスだけ多いため、 $q = 9$ の場合より多くのマスに対してルール重みの学習を行う必要がある。このことから学習にはより多くのエピソード数が必要であり、 $q = 9$ のときは本研究で定めた全 30 万エピソードで学習が収束する一方、 $q = 11$ のときはエピソード数が不足していたと考えられる。

なお、本研究では Table 1 に示したように全 30 万エピソードで学習を行ったが、この 30 万という値は筆者が試行錯誤的に決定したものである。Fig. 6 (b) より、 $q = 11$ の場合の値は 200($\times 1000$)エピソードで値が収束しているものの、 $q = 9A, 9C$ の場合は約 150($\times 1000$)エピソードで値が収束している。また、荒井らの検討では全 10 万エピソードでの学習を行っており⁴⁾。これらの点を考慮

して本研究ではエピソード数を 30 万とした。

続いて、 $q < p$ の場合は $q = 9, 7, 5$ の順に捕獲ステップ数が小さい値へ収束していることがわかる。まず $q = 9$ の場合、いずれの乱数の種を用いた場合も提案手法 1・2 の両方で捕獲ステップ数が $q = p$ の場合の収束値に近い値に収束していることがわかる。また、3・4 番目の乱数の種を用いた場合はそれぞれ提案手法 1・2 を用いた方がより小さい捕獲ステップ数に収束している。ただしその差は約 3 ステップほどであり、またこれ以外の 1・2・5・6 番目の乱数の種を用いた場合はほとんど収束値に差がない。このことから、両提案手法で捕獲ステップ数の収束値に大きな差はないと考えられる。なお、捕獲ステップ数の値の推移に着目すると、2・4・6 番目の乱数の種を用いた場合は提案手法 2 が、1 番目の乱数の種を用いた場合は提案手法 1・2 の両方が $q = p$ の場合より早く値が収束している。このことから、視野領域の広さが $q = 9$ の場合は提案手法 2 を用いることで比較的早く学習が収束するものの、学習性能に大きな差はなく、両手法とも $q = p$ の場合に近い学習性能を持っているといえる。

続いて $q = 7$ の場合、両提案手法とも捕獲ステップ数が約 25~28 ステップ程度に収束しており、2・4・6 番目の乱数の種を用いた場合はほとんど同じ値に収束していることがわかる。また 1・5 番目の乱数の種を用いた場合は提案手法 2 を、3 番目の乱数の種を用いた場合は提案手法 1 を用いた方がより少ないステップ数に収束している。このことから、 $q = 7$ の場合も $q = 9$ の場合と同様に両提案手法で学習性能に大きな差はないと考えられる。

ここで $q = 5$ の場合を考えると、まず $q = p, 9, 7$ の場合は捕獲ステップ数が学習初期は約 40~45 ステップとなっており、学習によってそこから約 20 ステップまたは約 30 ステップに収束している。一方 $q = 5$ の場合、学習初期は捕獲ステップ数が約 50 ステップとなっており、そこから 40~45 ステップの値に収束していることがわかる。特に 3・4 番目の乱数の種を用いた場合は収束値がほとんど変化していないため、 $q = 5$ の場合学習そのものは行えるものの、学習性能は低いと考えられる。

最後に、視野領域の広さについて考察する。今回のシミュレーション検討では対象とするフィールドの大きさを $p = 11$ としたが、この場合フィールドの総マス数は $11^2 = 121$ マスである。他方、視野領域を構成するマス数はその広さ q を用いて q^2 で求めることができる。ここで、フィールドの総マス数に対して視野領域が占めるマス数の割合を、視野領域の広さ q の関数として $H(q)$ で定義する。このとき、フィールドの大きさ p が $p = 11$ であるため $H(q) = q^2/121$ であり、 $q = 5, 7, 9$ とした場合の値を Table 3 にまとめる。このとき、 $q = 5$ の場合は学習の性能が低く、 $q = 7$ 以上であれば性能に差はあれ学習が行えるため、フィールドの大きさに対しておよそ 40%の領域が知覚できるとき、提案手法によって学習が行えると考えられる。また、視野領域の広さが $q = 9$ であるとき、すなわち約 66%が知覚できれば、提案手法によってフィールド全域が知覚できる場合と同等の学習性能を発揮できると考えられる。

Table 3 Ratio of each field of view to the field.

q	5	7	9
q^2	25	49	81
$H(q)$	0.2066	0.4049	0.6694

6. 結言

本研究では、視野領域を制限したハンターによる追跡問題に対して、視野領域外に存在する獲物を学習に用いるための手法について検討を行った。従来手法では獲物が視野領域外に存在することがないように無限大の広さの視野領域を持つことがほとんどであるが、視野領域外の獲物の位置を「写像領域」のいずれかへ写像し、その写像した状態を学習に用いることで視野領域外に獲物が存在するときでも追跡問題の学習が行えることをシミュレーション検討によって示した。

シミュレーション検討においては 11×11 のフィールドを対象とし、複数の乱数の種を用いて追跡問題の学習を行い、その学習性能の評価を行った。まず学習の継続性から考えると、従来手法では少なくとも2エピソード目までにいずれかのハンターの視野領域外に獲物が存在してしまい、学習が継続できなくなるのに対して、提案手法ではその視野領域外の獲物に対しても学習が行えることを示した。次に、フィールド全域が認識できているときに最も捕獲までのステップ数は少ない値に収束するものの、視野領域が 9×9 であれば、提案手法を用いることでフィールド全域が知覚できている場合とほぼ同じ捕獲ステップ数へ収束することが分かった。また、 5×5 の視野領域しか認識できない場合も性能は劣るものの、学習によって捕獲ステップ数が小さい値へ収束することが分かった。加えて、提案手法を用いれば少なくともフィールドの40%程度が知覚できれば学習を行うことができ、また66%程度が知覚できればフィールド全体が知覚できる場合と同等の学習性能を示すことが示された。

今後の展望として、現在は獲物の相対位置のみを状態として用いているが、例えば他のハンターの状態を加味しての学習や、情報通信を考慮した学習などを行い、ハンター間の協調の創発を試みる。また、フィールドをさらに大きくした上でのハンター・獲物の台数の増加、連続時間系への理論の拡張などを行う予定である。

参考文献

- 橘義博, 森翔平 and 滑川徹: 外乱を考慮したマルチエージェントシステムの協調制御, 計測自動制御学会論文誌, Vol.50, No.9, pp. 669-676 (2014).
- L. Canese, G. C. Cardarilli, L. D. Nunzio, R. Fazzolari, D. Giardino, M. Re and S. Spanò: Multi-Agent Reinforcement Learning: A Review of Challenges and Applications, Applied Sciences, Vol.11, No.11, p.4948 (2021).
- 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).
- 荒井幸代, 宮崎和光 and 小林重信: 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.5, pp.609-618 (1998).
- 伊木美太輔, 松本啓之亮 and 森直樹: 追跡問題における状態数とメモリ量の削減に着目した Q 学習, システム情報学会論文誌, Vol.25, No.10, pp.266-271 (2012).
- 安田真, 古橋武, 橋山智訓 and 大熊繁: マルチエージェント系のポテンシャル関数による制御と強化学習の適用に関する検討, 計測自動制御学会論文集, Vol.36, No.8, pp.698-706 (2000).
- 保知良暢, 松井藤五郎, 犬塚信博 and 世木博久: マルチエージェント強化学習における報酬発生条件に基づく貢献度判別と報酬分配, 人工知能学会全国大会論文集, Vol.JSAI02, p.129 (2002).
- 宮崎和光, 荒井幸代 and 小林重信: Profit Sharing を用いたマルチエージェント強化学習における報酬分配についての理論的考察, 人工知能学会誌, Vol.14, No.6, pp.1156-1164 (1999).
- 伊藤僚, 吉川毅 and 野中秀俊: マルチエージェント強化学習における知覚情報の適応的粗視化, 日本知能情報ファジィ学会 ファジィ システム シンポジウム 講演論文集, Vol.26, p.18 (2010).
- 長行康男 and 伊藤実: マルコフゲームにおける環境モデルの推定を利用したマルチエージェント強化学習法, 情報処理学会研究報告知能と複雑系, Vol.2001, No.73(2001-ICS-125), pp.29-36 (2001).
- 長谷川雄吾, 高田沙都子, 中野秀洋, 荒井秀一 and 宮内新: 行動選択確率に基づく動的強化関数を用いた強化学習手法, 信学論, J.89-D-I, No.4, pp.788-796 (2006).
- 片山謙吾, 興石尚宏 and 成久洋久: 強化学習エージェントへの階層化意志決定法の導入--追跡問題を例に--, Vol.19, No.4, pp.279-291 (2003).
- 渡邊俊彦 and 和田竜也: マルチエージェント追跡問題のための相対座標系に基づく階層型モジュラー強化学習, バイオメディカル・ファジィ・システム学会誌, Vol.12, No.2, pp.65-74 (2010).
- 玉城斉, 遠藤聡志 and 山田孝治: Q-learning を用いたマルチエージェントにおける協調行動獲得に関する研究, 琉球大学工学部紀要, No.54, pp.101-108 (1997).
- 山村雅幸, 宮崎和光, 小林重信: エージェントの学習, 人工知能学会誌, Vol.10, No.5, pp.23-29 (1995).
- 玉嶋大輔, 小坪成一, 岡本卓 and 平田廣則: 期待強化値を考慮した動的強化関数を用いる Profit Sharing, 電気学会論文誌 C (電子・情報・システム部門誌), Vol.129, No.7, pp.1339-1347 (2009).
- 宮崎和光, 山村雅幸, 小林重信: 強化学習における報酬割当ての理論的考察, 人工知能学会誌, Vol.9, No.4, pp.580-587 (1994).