状態空間が異なる強化学習エージェント間の学習結果の伝達

上田 展久　　　佐藤 泰介
東京工業大学
理工学研究科　　情報理工学研究科
電気電子工学専攻　　計算工学専攻
e-mail:{ueda,sato}@cs.titech.ac.jp

abstract

Transmitting learned results between agents is a way to reduce the time for learning. With this transmission, the closer what agents have to learn are, the longer time it reduces. To solve cooperative problems for multiple agents, what one agent has learned is expected to resemble what the others have to learn.

On the other hand, in reinforcement learning, agents learn estimated values for their actions on their states. Their states generally consists of the observations of the other agents. One agent observe himself in a fixed place. But the others observe him in an arbitrary place. Then their state-spaces must be different to distinguish one agent from the others. Even if they learn the same task, the estimated values on different state-spaces are different. Hence they cannot share directly learned results with the others.

We propose here a method to transmit one agent's learned estimated values to the others. We assume that any two agents have at least one observation in common of the agents which constitutes their state-spaces. We also show an experimental result of applying this method to a pursuit problem that multiple agents cooperate to capture a prey.

1　はじめに

各エージェントが学習結果を互いに伝達すると, 互いに学習していない部分を補う. そのとき, その学習すべきものが似ているほど, 伝達により学習時間が短縮できる. そして, 複数エージェントによる協調問題の解は, 各エージェントで似ている可能性がある. 例として, 複数のハンターが協力して, 獲物を捕獲する追跡問題 [1] が挙げられる. その問題では, 4 人のハンターが獲物に近付き, 協力して獲物を捕獲することはすべてのエージェントに共通している.

一方, 強化学習では, エージェントは各状態での行動の評価値を学習する. その状態空間は一般に他のエージェントの観測から構成される. このとき, 各エージェントは自分の観測は一定である. しかし, このエージェントの観測は他のエージェントにとって一定ではない. そのため, 観測するエージェントを区別すると, 状態空間が異なる. 加えて, エージェントが同じタスクを学習していても, 異なる状態空間での評価値は異なる. よって, このようなエージェントは学習結果を直接伝達できない.

そこで本稿では, 状態空間の異なるエージェント間での学習結果の伝達方法を提案する. また, 実験により学習時間の短縮を検証する.

2　状態空間

状態の例として, 追跡問題の 4 人のハンターエージェントを考える. 各エージェントの状態は, 自分以外の全エージェントの観測から構成されるとする. このとき, エージェント 4 以外の総てのエージェントが, エージェント 4 の観測を用いて状態を構成する. しかし, エージェント 4 の状態には, エージェント 4 の観測は関係ない. このように, 各エージェントの状態空間は異なる場合がある. これは, 観測しているエージェントを区別していることによる. [2] は, 状態空間を等しくするため, 他のエージェントを区別せずに現在の状態を決めている.

しかし, 各エージェントを区別したことで状態空間が異なっても, 各エージェントの状態空間は共通性を含む. 上の例でも, エージェント 1,2 両方の状態は, エージェント 3,4 や獲物の観測から構成される. そこで, エージェント間で共通している部分のみの状態空間を考え, その状態での評価値を伝達する.

3　伝達する評価値

学習した評価値を伝達する例として, 追跡問題でハンター 1 からハンター 2 へ評価値を伝達する場合を説明する. 説明のために, 状態を構成する値のベクトルで状態を表す. 各ハンターを区別するために, 獲物, ハンター 1,2,3,4 の相対位置 z_p, z_1, z_2, z_3, z_4 から状態を構成する. この状態を次のように表す.

$$s = [z_p, z_1, z_2, z_3, z_4].$$

ハンター i が観測できないことを,z_iの代わりに$-$で表す.

ハンター 1 の状態 $[z_p, -, z_2, z_3, z_4]$ は, 自分以外のエージェントの観測から成る. そして, その状態での行動 a に対する評価値 $Q_1([z_p, -, z_2, z_3, z_4], a)$ を学習する. ハンター 2 も同様に評価値 $Q_2([z_p, z_1, -, z_3, z_4], a)$ を学習する. このとき, ハンター 1 とハンター 2 の状態で共通なのは, 獲物とハンター 3,4 の観測から状態が構成されている点である. そこで, 獲物とハンター 3,4 の観測から成る状態空間の評価値 $Q_1([z_p, -, -, z_3, z_4], a)$ を次の近似式で計算する. 以下では, この値を伝達評価値と呼ぶ. この近似式では, 学習した評価値と学習中に訪れた状態の履歴 e_1 を用いる. そして, 計算された伝達評価値をハンター 2 へ伝達する.

$$\begin{aligned}
& Q_1([z_p, -, -, z_3, z_4], a) \\
\stackrel{\mathrm{def}}{=}\ & \underset{z_2}{E}[Q_1([z_p, -, z_2, z_3, z_4], a) | e_1] \\
=\ & \sum_{z_2} P([z_p, -, z_2, z_3, z_4] | e_1) Q_1([z_p, -, z_2, z_3, z_4], a) \\
\simeq\ & \sum_{z_2} \frac{N_1([z_p, -, z_2, z_3, z_4], e_1)}{\sum_{z'_2} N_1([z_p, -, z'_2, z_3, z_4], e_1)} \\
& \times Q_1([z_p, -, z_2, z_3, z_4], a).
\end{aligned}$$

ただし, $N_1(s, e_1)$ はハンター 1 が e_1 で状態 s を訪れた回数を示す.

そして, ハンター 2 は, 伝達された値を基に次の式で評価値を書き換える.

$$\begin{array}{rl} & Q_2([z_p, z_1, -, z_3, z_4], a) \\ := & (1-R)\cdot Q_2([z_p, z_1, -, z_3, z_4], a) \\ + & R\cdot Q_1([z_p, -, -, z_3, z_4], a). \end{array}$$

ただし,Rは伝達された値を取り込む率 $(0 \leq R \leq 1)$ とする.

そしてこの伝達は, 評価値の変化が閾値を示す任意の定数Cを越えたときに行なうとする. これは, 各ハンターが学習により, 今までの学習結果と異なる結果を得たときに伝達を行なうためである.

学習を行なっているときの流れは次のようになる.

0. 閾値Cを設定. 各ハンターの状態の構成要素を把握.

1. 各ハンター i が t ステップ目での, 現在の状態 s_iに置ける行動 a_iの評価値 $Q_i^t(s_i, a_i)$ を学習.

2. **if** $[|Q_i^t(s_i, a_i) - Q_i^{t-1}(s_i, a_i)| > C]$ **then** 各ハンターへの伝達評価値を計算, 送信

3. **if** 共通評価値が他ハンターから伝達 **then** 伝達評価値を基に評価値を書き換え

4. **1.** へ戻る.

4 特徴

本方法では, 状態をつくる観測が共通しているならば, 異なる状態空間のエージェントへ学習結果を伝達できる. しかし, 共通の状態空間を考え, 学習した値の平均を伝達評価値としている. このため, 正確な評価値を伝えることは出来ない.

そして, 学習とは別に伝達評価値の計算に時間が必要になる. 状態の数 m, 状態を構成する要素数 n に対して, $O(mn)$ の時間が掛かる. 状態の構成上, n に対して m は一般に指数倍になる. よって, 伝達には n とある定数 c に対して, 最悪 $O(c^n)$ の時間が必要になる. このため, 計算機実験のような計算時間と探索時間に差がない場合では, 本手法による学習時間の短縮を伝達評価値の計算時間が打ち消す. しかし, 実ロボットのような, 計算時間よりエージェントの探索時間が非常にかかる場合には有効である.

また, 本手法は一方向の伝達のみである. よって, 他のエージェントが観測しているエージェントが事前にわかっていれば, 伝達評価値を計算できる. 相互にコミュニケーションを取る必要はない.

5 実験結果

追跡問題を次のように改変した問題に本方法を適用する. そして, 異なる 2 通りの捕獲条件で実験する. 獲物と 3 人のハンターが, 10×10 のトーラス構造の格子平面にいる. 1 ステップで, 獲物は前後左右に 1 桝, ランダムに移動する. 1 ステップに, 各ハンターは格子平面上を前後左右に移動するか, その桝に留まる. 各ハンターは自分の周囲 5×5 の範囲を観測して, 獲物と自分以外のハンターの相対位置から状態を構成する. そして, 観測した状態での行動の評価値を学習し, その値を基に行動を決定する. そして, 獲物の捕獲がハンターの目標である. 獲物を捕獲したステップに, 捕獲したハンターへ報酬値 1 を与える. 捕獲の条件を実験 1 では 3 人のハンター, 実験 2 では 2 人のハンターが獲物に隣接することとする. それ以外のステップでは常にハンターは報酬値-0.01 を得る. 各エージェントは, t ステップで状態 s で行動 a を取り, 報酬 r を得て状態 s'への遷移する. そのとき, 次の Q 学習の更新式により状態 s での行動 a の評価値を学習する.

$$Q^{t+1}(s, a) := (1-\alpha)Q^t(s, a) + \alpha(r + \gamma \max_{a'} Q^t(s', a')).$$

そして, 1 試行をハンターをランダムに配置してから, 獲物を捕獲するか, 2000 ステップ経過するまでとする. ただし実験では,u を試行の回数として, 減衰率$\gamma = 0.9$, 学習率$\alpha = 0.5 \cdot (0.997)^u$, , 伝達の閾値 $C = 0.72 \cdot \alpha$, 伝達を取り込む率 $R = 0.1$ とする. また, エージェントの行動は, 確率 0.9 で $Q^t(s, a)$ が最大とする行動 a を取り, 確率 0.1 でランダムな行動選択を行なう.

計算機による 10 回の実験結果のグラフを以下に示す. 図 1, 図 2 の横軸は試行回数を示す. 図 1の縦軸は, 学習に掛かった総ステップ数の最大値, 最小値, 平均を表す. 図 2の縦軸は, 各試行に置ける獲物の捕獲までのステップ数の平均を示す. 実験 2 に対する同様のグラフが図 3と図 4 である.

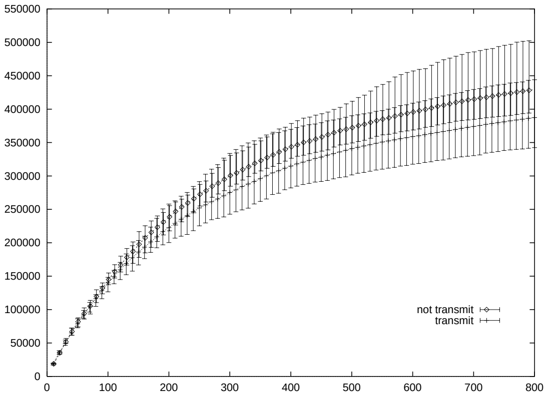

図 1: 実験 1 での総ステップ数

実験 1 では, 図 1から総ステップ数について, 伝達を行なうことで, 平均時間が若干短くなる. 800 試行までに掛かった総ステップ数の平均を比較すると, 伝達により約 9 パーセント短縮されている. しかし, 伝達を行なった方の総ステップ数の最大値は伝達を行なわない方の最小値を上回っている. よって, 必ずしも学習時間の短縮が期待できるわけではない. また, 図 2を見ると, 各試行のステップ数では差は見られない.

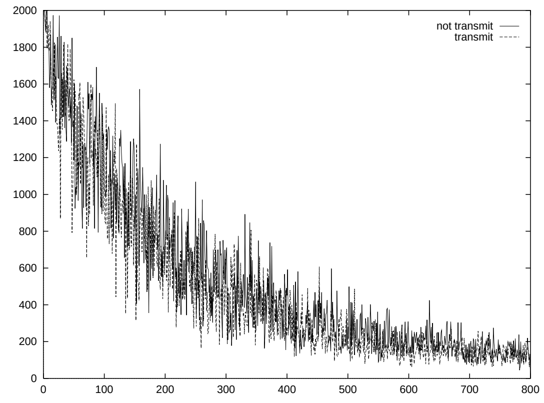

図 2: 実験 1 での各試行のステップ数

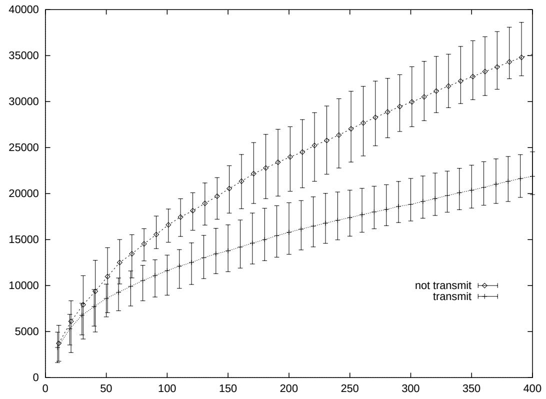

図 3: 実験 2 での総ステップ数

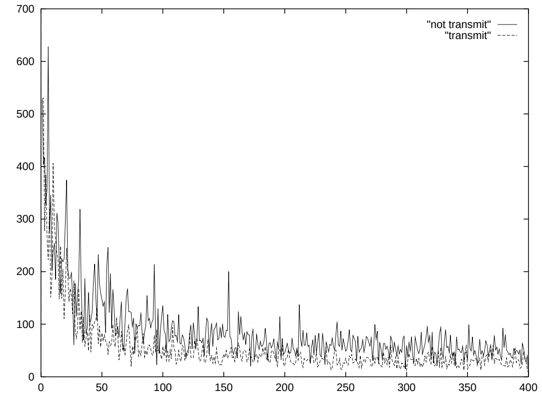

図 4: 実験 2 での各試行のステップ数

実験2では, 伝達を行なった方が行なわない方に対して, 総ステップ数で顕著な差が出ている. 400試行を終えるまでの総ステップ数の平均を比較すると, 伝達により約37パーセント短縮されている. そして, 伝達を行なった方の総ステップ数の最大値は伝達を行なわない方の総ステップ数の最小値を下回っている. つまり, 常に学習時間が短縮されている. しかし, 各試行のステップ数では大きな差は見られない.

実験1では, 3人のハンターが協力して獲物を捕獲するため, どのハンターも均等に報酬を得る. その均等に学習した結果をお互いに伝達するので, あまり大きな効果が現れていない, と考えられる. 実験2では, 3人のハンターのうち2人で捕獲できることから, 殆んどの試行で報酬を得られないハンターが出てくる. そのため, 学習の進んでいるハンターとそうでないハンターが現れる. その学習の進んでいないハンターに対して, 学習結果が伝達されることで学習の進んでいない部分が補われている, と考えられる.

6 まとめ

状態空間の異なるエージェント間で学習結果を共通な状態空間での評価値に変換することで伝達する方法を提案した. そして, 実験により, 学習に掛かる総ステップ数が平均的に少なくなること, 各エージェントの学習の進み方に偏りがあるときには学習に掛かる総ステップ数が大幅に少なくなることを確認した.

参考文献

[1] 大沢 英一, 沼岡 千里, 石田 亨. 分散人工知能における標準的小問題. コンピュータソフトウェア, Vol.10, No.3, pp.3-19. 1993.

[2] Ming Tan. Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents. In *Proceedings of the 10th International Conference on Machine Learning*. pp.330-337. 1993.

[3] Christopher J. C. H. Watkins and Peter Dayan. *Q*-learning. *Machine Learning*, Vol.8 No. 3, pp.279-292. 1992.