2M5-OS-32b-2

マルチエージェント強化学習における エピソードの順序が獲得方策に与える影響

Influence of sequence of episodes on acquisition strategy in multi agent reinforcement learning

木村 祥 *1　荒井 幸代 *1
Sho Kimura　Sachiyo Arai

*1千葉大学大学院融合理工学府 地球環境科学専攻 都市環境システムコース
Department of Urban Environment Systems, Division of Earth and Environmental Sciences,
Graduate School of Science and Engineering, Chiba University

In multi-agent reinforcement learning, convergence to optimal policy is made difficult by non-Markov processes caused by simultaneous learning problems and incomplete perception problems. Conventionally, the way for avoiding divergence and for promoting convergence depends on symptomatic treatment such us reducing the exploration rate in action selection. In this research, we focus on the episode experience sequence and consider the effect of convergence of learning and the influence on convergence improvement in the Predator Pray.

1. はじめに

エージェントの知識設計が困難といわれるマルチエージェント環境において教師信号を必要としない強化学習が期待されているが，マルチエージェント強化学習では，学習するエージェントが複数存在することに起因する「同時学習問題」が存在する．「同時学習問題」に対しては協調を前提としないタスクにおいても個々が独立に学習した場合は，学習環境の非マルコフ性から得られる解の最適性は保証されない．これまでは，学習初期の exploit を大きくすることによって各エージェントの方策を早い段階で収束させる対症療法的な措置にとどまっている．そこで，本研究では蓄積した経験サンプル（以下，バッチと記す）を繰り返し利用できるバッチ強化学習に着目する.exploit が初期乱数に収束を委ねるのに対して，バッチを事前に分類し，分類したバッチの入力順序を操作して学習した結果，従来の学習以上の学習性能が得られることを計算機実験を通して確認し，提案手法の有効性を検証する．

以下，2 章では，マルチエージェント強化学習について説明する．3 章では，提案手法としてバッチ強化学習を用いたエピソードの入力順序を考慮した学習手法を説明する．4 章では，計算機実験の結果を示し，5 章において本研究をまとめる．

2. 対象問題

2.1 マルチエージェント強化学習の同時学習問題

本研究ではマルチエージェント環境における学習を対象とする．マルチエージェント環境とは複数の自律エージェントから構成される環境のことである．マルチエージェントの学習段階では，各エージェントの方策の更新によって状態遷移の定常性が保証されない．これに起因する問題を同時学習問題と呼ぶ．同時学習問題に対するアプローチとしては，学習率を制御するものがある [野田 14]．これは学習率を制御することによってエージェントの方策更新を低減し，収束性を保証するための提案である．

これらに対して本研究では，エピソードの入力順序を操作し方策の強化順序を考慮することによって学習性能を向上させることを考える．

連絡先: 木村祥，千葉大学大学院融合理工学府，千葉市稲毛区弥生町 1-33，kimusho@chiba-u.jp

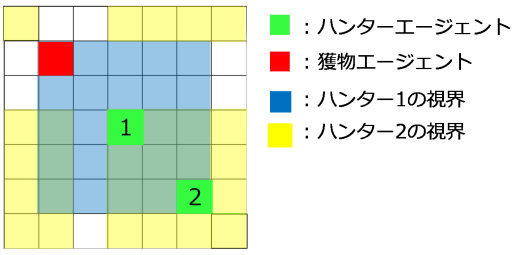

図 1: 追跡問題の環境

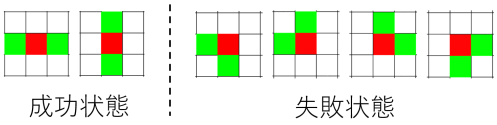

図 2: 終端状態の分類

2.2 追跡問題

追跡問題とは複数のハンターエージェントが，環境を探索して獲物エージェントを捕獲する問題である．追跡問題における環境を図 1 に示す.7 × 7 格子状のトーラス環境を設定し，2 体のハンターと獲物を配置したものである．各エージェントは，ハンター 1，ハンター 2，獲物の順番で行動し，上下左右の方向へ 1 マス進むか，その場にとどまる行動を一つ選択する．獲物は上下左右にハンターがいる場合にハンターから遠ざかる行動をとる．複数のエージェントが同じマスに存在することは許されない．ハンターの視界は 5 × 5 で，自分の周囲 24 マスを見ることができる．2 人のハンターが獲物に隣接した状態を終端状態とし，2 人のハンターに報酬が与えられる．初期状態から終端状態までを 1 エピソードとする．本研究では，エピソードの学習順序の影響を明らかにするため，追跡問題における終端状態を成功状態に加えて，失敗状態を導入する．成功状態では正の報酬を，失敗状態では負の報酬である罰をそれぞれ与える．図 2 に成功状態，失敗状態を示す．

1: $Q \leftarrow Q_0$.
2: **repeat**
3: $D \leftarrow \emptyset$.
4: $t \leftarrow 0$.
5: $episodes \leftarrow 0$.
6: **repeat**
7: $t \leftarrow t + 1$.
8: **if** new episode **then**
9: $s_t \leftarrow getStateFromEnvironment()$.
10: $episodes \leftarrow episodes + 1$.
11: **end if**
12: $a_t \leftarrow selectAction(Q, s_t)$.
13: Execute a_t.
14: $r_t \leftarrow getRewardFromEnvironment()$.
15: $s_{t+1} \leftarrow getStateFromEnvironment()$.
16: $d_t \leftarrow (s_t, a_t, r_{t+1}, s_{t+1})$.
17: $D.append(d_t)$.
18: **until** $episodes = m$.
19: **for all** $t \in [1..|D|]$ **do**
20: $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\Big(r_{t+1} + \gamma \arg\max_a Q(s_{t+1}, a) - Q(s_t, a_t)\Big)$.
21: **end for**
22: **until** Q has converged.

図 3: batchQ のアルゴリズム

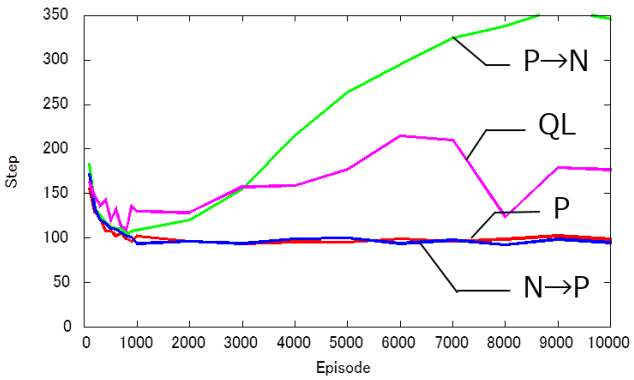

図 4: 成功状態に至るまでのステップ数
(P:正例だけ,P → N:正例の後に負例,N → P:負例の後に正例)

表 1: 成功状態に至るまでのステップ数 (step)

	Episode	1000	2000	5000	10000
QL	average	133	116	179.4	169
	SD	76.1	33.7	93.5	94.8
P	average	103	98	95.8	98
	SD	4.6	7.9	8.4	5.8
P → N	average	104.8	120.6	261.2	348.4
	SD	5.3	4.1	20.0	8.1
N → P	average	96.2	98.2	100.2	95.4
	SD	5.5	4.9	3.1	7.2

3. 提案手法

本研究では，2.2 節で示した追跡問題を対象とし，エピソードを分類したバッチを学習器に与える順序が学習に及ぼす影響を考察する．

3.1 バッチ強化学習

バッチ強化学習 [Lange 12] はある時刻 t においてエージェントの経験したサンプル (状態 s_t，行動 a_t，次状態 s_{t+1}，報酬 r_{t+1}) を蓄積し，蓄積したサンプルの集合（以下，バッチと記す）に基づいて方策 π を更新するアルゴリズムである．方策の更新に Q-Learning[Sutton 98] の更新式を用いる手法が batchQ である．

batchQ

batchQ は状態遷移を探索ではなくバッチに基づいて観測し，行動価値関数 $Q(s, a)$ を更新するアルゴリズムである.batchQ のアルゴリズムを図 3 に示す．

3.2 エピソードの入力順序の導入

提案法は収集したバッチの分類，バッチ強化学習における分類したバッチの入力順序の操作の二つのステップで構成される．あるエピソードの終端状態に到達した時刻を $t = T$ とすると，バッチとは，エージェントが初期状態から終端状態まで到達するサンプル (s, a, s', r) の集合 $d_t = (s_t, a_t, s_{t+1}, r_{t+1})$,$t = 1, 2, 3, \cdots, T$ のことである．はじめに，バッチを以下に定義する正例バッチと負例バッチに分類する．

・**正例バッチ：** 成功したエピソード (図 2 の成功状態)

・**負例バッチ：** 失敗したエピソード (図 2 の失敗状態)

次に分類したエピソードをバッチ強化学習に入力する．実験では以下四つを比較する．正例バッチと負例バッチの両方を入力する場合，正例:負例=9:1 の比率で入力する．

1. **正例だけ：** 正例バッチだけを学習
2. **負例だけ：** 負例バッチだけを学習
3. **正例の後に負例：** 正例バッチを学習後，負例バッチを学習
4. **負例の後に正例：** 負例バッチを学習後，正例バッチを学習

4. 実験

4.1 実験設定

図 1 に示す 7 × 7 格子状のトーラス環境にハンターエージェント 2 体と獲物エージェント 1 体を，それぞれ左上，右下，中央に配置したものを初期状態とする．従来法の探索による QL と提案手法における正例だけ，負例だけ，正例の後に負例，負例の後に正例の五つを実験する．

最初に追跡問題においてバッチをランダム方策で収集し蓄積する．次に蓄積したバッチを前章で解説したバッチの種類ごとに分類し，入力順序を考慮して方策を学習する．評価ステップは 1000 エピソード行い，成功状態までにかかったステップ数及び成功状態に至る確率（以下，成功確率と記す）から評価する．以上を 1 試行とし 5 試行で平均をとった結果を示す．

batchQ の行動選択にはボルツマン選択を用いる．成功状態時の報酬を 100，失敗状態時の罰を-1，割引率 γ を 0.99，学習率 α を 0.01，温度 τ は 0.1 に設定した．

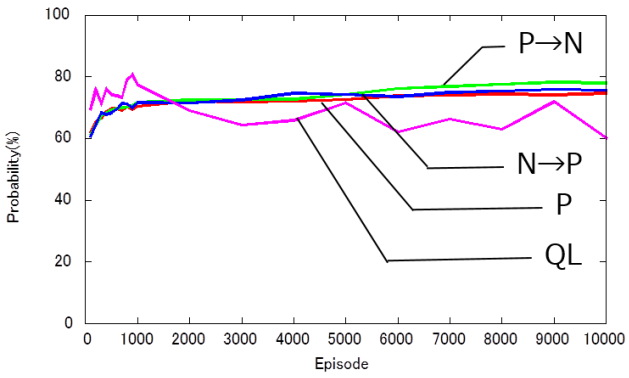

図 5: 成功確率
(P:正例だけ,P → N:正例の後に負例,N → P:負例の後に正例)

表 2: 成功確率 (%)

	Episode	1000	2000	5000	10000
QL	average	72.8	68.2	75.0	57.0
	SD	3.9	11.1	11.4	12.1
P	average	70.4	72.4	72.8	74.7
	SD	2.3	2.0	1.5	2.3
P → N	average	71.9	72.6	74.3	78.0
	SD	1.5	1.6	0.5	1.3
N → P	average	71.7	71.5	74.5	75.8
	SD	1.0	2.0	2.4	2.3

4.2 実験結果および考察

「負例だけ」で学習した場合，学習が収束しなかった．本節では，QL,「正例だけ」,「正例の後に負例」,「負例の後に正例」の四つの実験結果を示す．

図 4 に成功状態に至るまでのステップ数の変化，表 1 に学習エピソード数ごとのステップ数の平均と分散を示す．「正例の後に負例」で学習した場合は学習が収束しなかったが,「正例だけ」,「負例の後に正例」で学習した場合に学習が収束し，QL 以上の学習性能が得られた．また，10000 エピソード学習後の成功状態に至るまでのステップ数において，QL と「正例だけ」,「正例の後に負例」,「負例の後に正例」のそれぞれについて t 検定を行ったところすべての入力順序で有意差が認められた ($p \prec 0.05$).

次に，図 5 に成功確率の変化，表 2 に学習エピソード数ごとの成功確率の平均と分散を示す.QL では学習が振動しているが，学習性能は,「正例の後に負例」,「負例の後に正例」,「正例だけ」の順によいことがわかる．10000 エピソード学習後の成功確率において，QL と「正例だけ」,「正例の後に負例」,「負例の後に正例」のそれぞれについて t 検定を行ったところすべての入力順序で有意差が認められた ($p \prec 0.05$).

提案手法の性能について考察する．三つの入力順序では，それぞれエージェント自ら状態遷移した後に成功状態に至る Q 値が高く，他エージェントの行動に依存せずに成功状態に至る方策を獲得できたことが性能の向上に影響したと考えられる．また，正例と負例の両者で学習した方が性能が向上することは自明であるが，負例を後にした方が成功確率が高いのは，例えば図 6 に示す状態について考える．この状態ではハンター 1 は

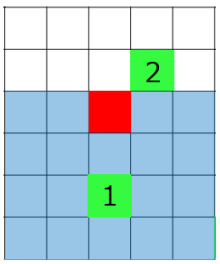

図 6: 緑:ハンター, 赤:獲物, 青:ハンター 1 の視界

上へ移動するのが成功状態に近づく行動であるが，成功状態に至るか失敗状態に至るかはハンター 2 の行動に依存する．「負例の後に正例」,「正例だけ」で学習した場合，ハンター 1 は上へ移動する方策を学習したが,「正例の後に負例」で学習した場合はランダムに行動する方策を学習した．このランダム方策は負例による学習時の負の報酬に依存し, 失敗状態に陥るリスクを避け，成功状態に至る最短経路から離れた状態を経て成功状態に至る方策をであると推測できる．また，この状態では不完全知覚問題も学習に悪影響を及ぼしていると考えられる．図 6 に示す状態ではハンター 1 の視界にハンター 2 が入っていないため，ハンター 1 は上へ移動する行動が成功状態に近づくと知覚できない．

5. 結論

本研究では，マルチエージェント強化学習におけるエピソードの入力順序の影響を考察するため，追跡問題を対象とした計算機実験を行った．エピソードのタイプを明確にするため，成功状態と失敗状態を導入し，エピソードを正例と負例に分類した．エージェントの学習性能は両者を与えることによって向上することを示した．しかし，正例と負例，いずれを先に与えるかについては，成功状態と失敗状態にそれぞれ設定した報酬値の影響があると考えられるため，現時点では明確ではない．

今後の課題として，この報酬設計問題とともに，本論文で用いたブートストラップ法である Q-learning 以外の強化学習アルゴリズム，たとえば，状態遷移確率に依らないモンテカルロタイプのアルゴリズムに対するバッチ順序の効果を検証することを挙げる．

参考文献

[野田 14] 野田五十樹, :マルチエージェント強化学習の最適 Exploration 率と各種パラメータの関連の実験的考察, population No.60, Vol.80, pp.100, (2014)

[Lange 12] Lange, Sascha, Thomas Gabel, and Martin Riedmiller :Batch reinforcement learning, Reinforcement learning. Springer Berlin Heidelberg, pp. 45-73, (2012).

[Sutton 98] Richard S. Sutton, Andrew G. Barto, : Reinforcement Learning: An Introduction , A Bradford Book, TheMIT Press (1998).