F-003

PS-GTR を用いたマルチエージェント強化学習システムにおけるロバスト性

Robustness of multiagent reinforcement learning systems using PS-GTR

中畑 一宏† 延澤 志保‡ 太原 育夫†‡

Kazuhiro Nakahata Shiho Nobesawa Ikuo Tahara

1. はじめに

近年，複数のエージェントの協調による問題解決を目指したマルチエージェントシステムが注目されている．マルチエージェント強化学習における協調的な行動の学習は，全てのエージェントが正常に動作する限り有効に機能する．しかし学習途中で，あるエージェントが故障した場合，今まで築いてきた協調関係が崩れ，故障後の学習効率は低下する可能性がある[1]．故障後に協調的な行動を再び学習することは有用ではあるが，学習には時間が必要となる．そこで本稿では，学習途中においても学習のやり直しが行える学習手法 PS-GTR[2]に着目し，この手法がマルチエージェント環境下においても適用可能であることを示し，故障発生後の学習において PS より有効な手法となることを実験により示す．

2. 学習アルゴリズム PS-GTR

強化学習の代表的な手法である Profit Sharing(PS)は，合理性を完全に保証している反面，報酬を目標状態から遠いルールまで伝搬することができない．また環境がマルコフ決定過程であっても，最適性が保証されるとは限らない．こうした短所は，評価値の初期値に依存している問題である．そこで各状態ごとに評価値の初期値を適切に設定する方法として，PS-GTR が提案されている．PS-GTR では，状態 s の最短経路長を l_{min}，適切に学習が進む比率を h とし，状態におけるルールの数を L とすると，評価値の初期値 $w_{init}(s)$を，

$$w_{init}(S) = \left(\frac{r \times f(l_{\min})}{L} \right) \times h$$

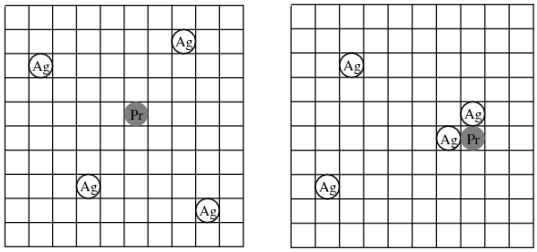
と設定する．ただし学習初期には l_{min} の値はわからない．そこで既知の最短経路長に合わせて評価値の初期値を設定し，より短い経路が見つかるとその経路に合わせて初期値を設定し直し，学習を 0 からやり直す．見つからない場合は，報酬の分配量を累積することで経験を利用する．このように PS-GTR では，学習をやり直すといった再学習機能を持つ．環境変化としてエージェントの故障を想定した場合，環境変化前の過去の経験に固執するのは望ましくない．そこで，再学習機能を持つ PS-GTR を使用することにより，環境変化に柔軟に対応できると期待される．

†東京理科大学大学院理工学研究科情報科学専攻
‡武蔵工業大学知識工学部
†‡東京理科大学理工学部

3. 実験

3. 1 マルチエージェント環境下での PS-GTR の適用

マルチエージェント環境下において，PS-GTR が適用可能かどうか，PS と比較したときどうかを調べるため，追跡問題を用いて実験を行った．10×10 の 2 次元トーラス平面上にハンター4 体，獲物 1 体を配し，図 1(b)のように 2 体以上のハンターが獲物に隣接することを目標とする．ハンターの視界は 5×5 とし，獲物は 3×3 とする．ハンターの行動は上下左右と停止であり，獲物は逃避的な行動をとる．



(a) 初期状態 (b) 終了状態

図 1：追跡問題

実験では，図 1(a)のように初期配置はランダムに設定する．目標達成したとき，獲物に隣接したハンターに報酬 10^6 を与える．PS では各ルールの初期の重みは 10^6 とする．PS-GTR では h を 1 とした．ハンターの学習はそれぞれ独立して行われ，互いに情報共有は行わないものとする．以上の設定の下，10 万エピソードを 1 試行として 100 試行の実験を行った．

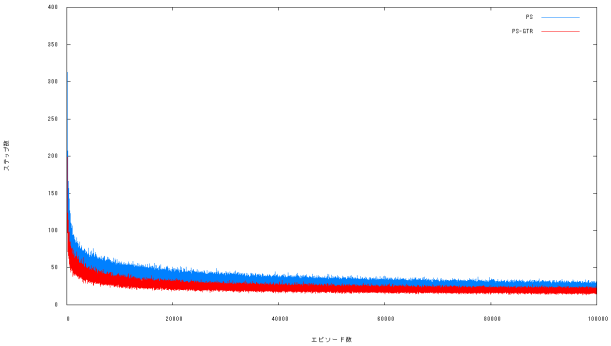


図 2： PS と PS-GTR の比較

図 2 は，PS と PS-GTR の 2 パターンで学習させた結果である．この結果から，PS-GTR はマルチエージェント環境下において適用可能であることがわかる．また PS-GTR は，PS より学習初期では早い学習が可能となり，学習後期においても良い性能に収束した．この理由は，PS-GTR には評価値の初期値を更新し学習をやり直す機能があるからだと考えられる．PS-GTR では，今までの経験よりも短い経路を経験すると，その状態では新しい初期値を設定し再学習が行われる．さらに報酬獲得に成功したエージェントのペアは今まで築いてきた協調的な行動も再学習することができ，以前よりも優れた協調行動を獲得する可能性がある．そのため，PS よりも性能が向上したと考えられる．

3．2　故障発生後の学習効率

PS-GTR がシステムのロバスト性に有効か検証するため，PS との比較実験を行った．実験環境は図 1 の追跡問題を用い，また PS，PS-GTR の初期設定も同様にする．

ハンターエージェントの故障人数はランダムに 2 体とし，故障発生後はその場に停止し続けるものとする．また故障したエージェントに捕獲能力はないものとする．以上の設定の下，10 万エピソードを 1 試行として 100 試行の実験を行った．

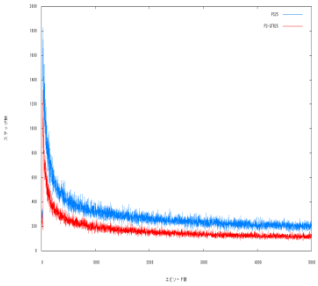


図 3： 25 エピソードで故障

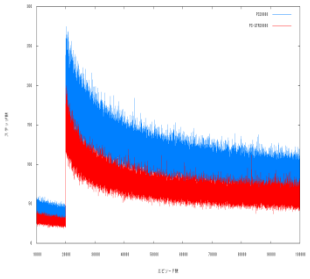
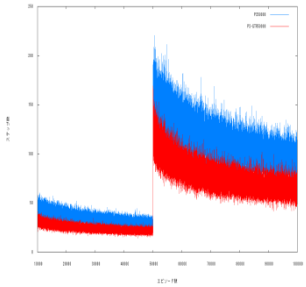


図 4：2 万エピソードで故障



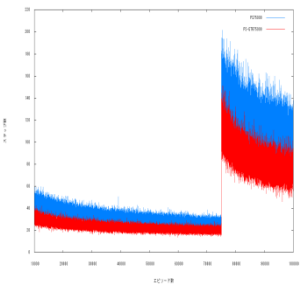


図 5：5 万エピソードで故障　図 6：7 万 5 千エピソードで故障

図 3 では，故障を 25 エピソードで発生させ，0 から 5000 エピソードまでの範囲を表示している．また図 4 から図 6 では，故障を 2 万，5 万，7 万 5 千の各エピソードで発生させ，1 万エピソードから 10 万エピソードまでの範囲を表示している．

表 1：故障発生直後の平均ステップ数

	図 3	図 4	図 5	図 6
PS	1829.95	238.65	167.80	123.17
PS-GTR	1151.95	162.36	131.99	109.19

図 3~図 6 と表 1 の結果から，PS-GTR がシステムのロバスト性に有効であることがわかる．また PS-GTR は PS よりも高いロバスト性を確保していることがわかる．PS-GTR が故障発生直後においてロバスト性を確保できた理由は，再学習の機能により，故障前での協調的行動の学習が PS よりも進んでいたからと考えられる．図 2 からもわかる通り，PS-GTR は PS よりも学習効率が良い．そのためエージェント同士の協調的行動の学習が進んでおり，どのエージェントが故障を起こしてもすぐに対応できたのではないかと考えられる．また PS-GTR が故障発生後の学習において PS よりも良い性能を示した理由も，再学習の機能があるからだと考えられる．故障発生後の学習では，故障前に有効であったルールが無効なルールへと変化してしまう場合がある．PS では過去の経験に固執し続けるため，無効となったルールを抑制するまでに時間を要し学習効率が低下する．しかし PS-GTR では，以前より優れた経路を発見すると再学習を行えるため，過去の経験には固執しない．そのため故障発生後に無効となったルールを早めに抑制できる．以上のことから，PS-GTR は環境変化に対応でき、また PS よりも良い性能に収束したと考えられる．

4　まとめと今後の課題

本稿では，PS-GTR がマルチエージェント環境においても適用可能であることを確認し，故障発生後の環境において有効な手法であることを確認した．

今回の実験において，故障発生時期が早いほど故障後のステップ数が増加した(図 3~図 6)．今後の課題としては，この原因の追及が考えられる．さらに故障人数やタスク等を変更することで，故障発生後の学習が困難となる条件を明らかにしたいと考えている．

参考文献

[1] 菊田洋一, 相場亮, “環境変化に対するマルチエージェントのロバスト性に関する研究,” 情報処理学会第 67 回全国大会, pp.265-266, 2005.
[2] 植村渉, 上野敦志, 辰巳 明治, “経験に固執しない Profit Sharing 法,” 人工知能学会論文誌, Vol.21, No.1, pp.81-93, 2006