CF-004

非通信マルチエージェント強化学習における 獲得報酬値の変動を用いたエージェント数の動的変化への追従

Non-Communicative and Cooperative Multi-Agent Reinforcement Learning with Reward Fluctuation to Adapt to Dynamic Number of Agents

上野 史† 髙玉 圭樹†
Fumito Uwano Keiki Takadama

1. はじめに

マルチエージェント強化学習はエージェントと呼ばれる行動の主体が複数集まり，お互いの適切な相互作用を学習することで，1 個の計算機では解決できない困難な問題を解く手法である．近年では，ゲーム AI，航路プランニング，ロボット制御からデータマイニングに至るまで幅広く様々な分野で活用されており，重要な技術であるといえる．しかしエージェントの数が多ければ多いほどそれを制御することは難しく，従来では通信や観測情報を利用し，他エージェントの振舞いを知ったうえで自身の行動を学習する[1,2,3]．しかしながら，これらの手法はエージェント数を固定したうえで初めて機能するものであり，現実問題を解決する上での大きな制約となる．例えば最適な経路を示すカーナビシステムをマルチエージェント強化学習で実現しようと思えば，エージェントである車の数は変更しないという前提を置く必要がある．しかしそれは適切ではない．そこで，本研究はエージェント数が動的変化する状況を想定し，その変化に追従して適切な協調行動を学習可能な方法を提案することを目的とする．具体的には，強化学習における報酬，特に動的変化におけるその変動を利用して，各エージェントが協調行動を学習するために適切な報酬値を設定する手法を提案する．

本論文の構成は以下のようになる．まず 2 章で本研究が扱う問題を定義し，3 章では本研究で用いる強化学習である Q 学習と利益最小型強化学習（PMRL）を紹介する．次に 4 章にて本研究の提案手法の説明をする．その後 5 章にて実験とその結果について議論を行い，最後に 6 章で本論文をまとめる．

2. 問題設定

以下の式は本研究で扱う環境の状態遷移を示す式である．S, A, τはそれぞれ環境の状態，状態遷移のきっかけとなるエージェントの行動，そして状態遷移関数を示す．環境状態Sはそれぞれエージェントに観測される状態$s_0, \ldots, s_D$から成り立ち，行動Aは各エージェントの行動$a^1, \ldots, a^n$の積で計算され，環境は状態遷移関数τに従って，次の状態へ遷移する．この時右式は 1 であり，これは次状態へ確率 1 で遷移することを示す．また，式(4)は報酬関数をであり，各エージェントがこの式に従い報酬を獲得する．式(4)は任意のエージェントiの報酬関数を示す．s^iは任意のエージェントiの現在状態であり，S_{goal}は目的の環境状態の集合である．つまりこの式はエージェントiが目的状態へ遷移したとき報酬値 10 を獲得するという意味である．この式は全エージェント共通でもっている．

$$S \coloneqq \{s_0, s_1, \ldots, s_D\} \quad (1)$$

$$A = a^1 \times a^2 \times \ldots \times a^n \quad (2)$$

$$\tau: S \times A \times S \to 1 \quad (3)$$

$$r^i = \{10 \mid s^i \in S_{goal}\} \quad (4)$$

2.1 グリッドワールド

本研究では実験で用いる問題としてグリッドワールドを採用する．図 1 は 2 体エージェントによるグリッドワールドの例である．「Agent A」，「Agent B」と示されたマスはエージェントの行動開始地点を示し，「Goal X」，「Goal Y」は目的地点を示す．この時，各マスが状態sを示し，上下左右の行動が行動aとなる．そしてゴール状態がS_{goai}の集合に格納されている．ここでは各エージェントが同じゴールへ到達することなく，ゴールに到達すれば報酬 10 を獲得する．

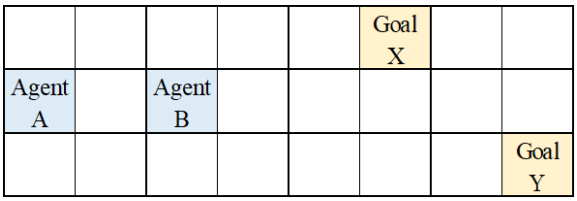

図 1 グリッドワールド

2.2 エージェントの動的変化

本研究では，エージェント数の動的変化を扱う．エージェント数の動的変化とはここでは，エージェント数とゴール数が等しく変化する環境である．つまり本研究において動的変化前も後もエージェント数とゴール数は等しい想定を置く．式(1)~(4)では，式(2)において，エージェント数の最大値nが増え，さらに報酬獲得の条件は変化しないが，ゴールが増えるなど，目的状態が増えるため，式(4)のS_{goal}が変化する．グリッドワールドで言えば，スタート地点およびゴールが図 1 よりも増えるということになる．

3. 従来手法

本研究では学習法として Q 学習，報酬設計方法及び学習方針として利益最小型強化学習(PMRL)を採用する．以下にその説明をする．

3.1 Q 学習

Q 学習[4]は強化学習における代表的手法である．Q 学習においてエージェントは，環境から自分の状態を観測しそれに対する最も適した行動を実行する．そして，その行動

†電気通信大学 The University of Electro-Communications

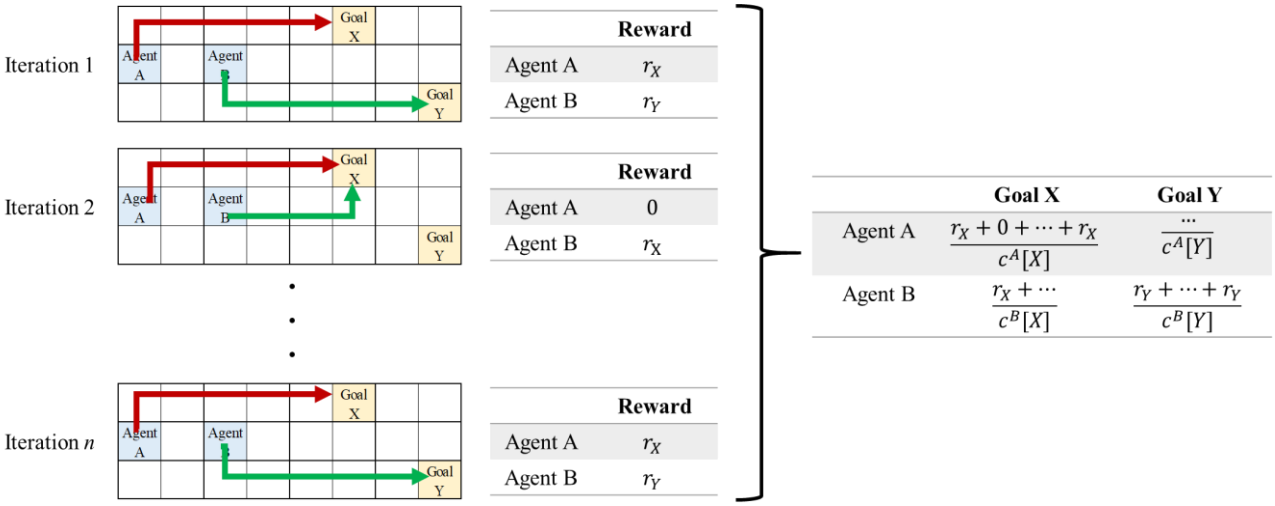

図 2　獲得報酬期待値

によって得られた報酬を基にその状態で選択した行動の価値を推定する．Q 学習ではこの価値を Q 値と呼び，エージェントは単位時間あたりに得られる報酬が最大となるように Q 値を設定して各状態に適切な行動を学習する．学習後は選択した行動によって変化した環境の次の状態を観測し，同じ流れを繰り返す．なお本研究ではエージェントの状態観測，行動選択，報酬獲得，学習までの流れをステップと呼び，各ステップ式(5)で Q 値を更新する．

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max Q(s',a') - Q(s,a)] \quad (5)$$

式(5)では，s,aが観測した状態と実行した行動を示し，その時の Q 値を$Q(s,a)$と表現する．rは次状態へ遷移した際獲得する報酬，$\max Q(s',a')$はその次状態における最大の Q 値を表す．また，γは割引率と呼ばれる次状態の Q 値に掛ける重みを表した 0 から 1 までの実数値である．また，学習率αは，1 回の更新で得られた価値を何割利用するかを決める．式(5)から Q 値は割引期待報酬和に収束する．割引期待報酬和とはその行動をとった時に得られる報酬の重み付き和の期待値である．

3.2 利益最小型強化学習

利益最小型強化学習（PMRL）[5]は，すべてのエージェントが譲歩行動をとるように報酬を設計することによって，システム全体として協調的に振舞えるように学習する手法である．PMRL は以下に示す内部報酬と目的価値の 2 個のメカニズムを導入し，目的価値に基づき達成する目的を選択し，内部報酬を変化させることでその目的を達成する．

3.2.1 内部報酬

PMRL では報酬を獲得した際に，その報酬で学習するのではなく，内部報酬から Q 値を学習する．内部報酬は，獲得した報酬に基づき式(6)に従い計算する．式(6)において，ir_gは任意のゴールgにおける内部報酬を示しています．また，γは Q 学習における割引率を示し，g'は報酬を獲得したg以外の任意のゴールを示す．そして$r_{g'}$はg'における報酬を示し，t_gはゴールgに到達するまでの最短のステップ数を示しています．この式のように内部報酬を設定することで，エージェントはゴールgへ到達できるように学習できる．その理由は，Q 学習が距離に応じて Q 値をγだけ等しく割り引くため，報酬値の大きさとゴールまでの距離によってエージェントがどのゴールへ到達するかが分かるためである．つまり，初期状態での目的のゴールへ向かう行動の Q 値が最大となるように報酬を設定すればその状態へ到達することが可能となる．

$$ir_g = \max_{g' \in G, g' \neq g} r_{g'} \gamma^{t_{g'} - t_g} + \delta \quad (6)$$

3.2.2 目的価値

目的価値は協調行動を学習するためにそれぞれのゴールがどれだけふさわしいかを示す価値である．目的価値はその値自体に大きな意味はなく，その大小関係に意味がある．そして，目的価値の最も大きいゴールが最適なゴールとなる．式(7)は目的価値の更新式を示している．式(7)において，bid_g^iが任意のエージェントiの任意のゴールgにおける目的価値であり，t_g^iはエージェントiがゴールgへ到達するまでの最短ステップ数を示す．この値は学習中にゴールgに到達した際のステップ数から求める．そしてn_g^iはエージェントiがゴールgの目的価値を更新した回数である．$\phi(g)$は one-hot 表現の関数であり，報酬を獲得した場合は 1，そうでない場合は 0 を返す．各エージェントは学習毎に目的価値を式(7)に従って更新し，次の学習から目的価値の最も大きなゴールgに到達するように内部報酬を設定する．式(7)は更新を繰り返すと最短ステップ数t_g^iに収束するため，自身から遠いゴールほど評価される．また関数$\phi(g)$の効果により，自身が報酬獲得可能なゴールの中から，最も遠くのゴールを選択することになる．

$$bid_g^i \leftarrow \frac{n_g^i - 1}{n_g^i} bid_g^i + \frac{t_g^i \phi(g)}{n_g^i} \quad (7)$$

4. 期待値に基づく目的選択法

エージェントやその目的が動的に変化することで，各エージェントに必要な協調行動も変化するため，その振る舞いを変える必要がある．特に従来手法では，常にすべてのゴールの中から遠くのものを目指すため，例えばエージェントが増えたことにより遠くのゴールへ向かう必要がなくなったとしても，そのエージェントはそれに気づかずに遠くを目指し続けてしまう．本研究では環境変化を察知し，それに合わせて目指す目的を制限することでその問題に対処する．具体的には従来手法 PMRL に対して報酬の期待値を導入する．報酬の期待値とは今までの学習において各ゴールへ到達した際の報酬獲得の確率に報酬値を掛けた値で

ある．図 2 は獲得報酬期待値の計算方法を示す．左側の各行は各学習の結果を示し，その左はグリッドワールドと各エージェントの辿った経路を示し，右は獲得した報酬値を示す．また図の右側は左側の経験から計算した獲得報酬期待値を示す．図の式では，r_X, r_Yはゴール X, Y へ到達した際に獲得可能な報酬値（ここでは両方 10）であり，$c^A[X], c^A[Y], c^B[X], c^B[Y]$は各エージェントがそれぞれのゴールに到達した回数である．図のように獲得報酬値を到達回数で割ることにより獲得報酬期待値を求める．そして本研究ではこの値を用いて各エージェントの到達ゴールを制限する．具体的には各エージェント獲得報酬期待値を計算し，この期待値が閾値を超えないゴールに関しては目的価値推定の際に選択しないという方法をとる．

図 3 はエージェント数が 1 体増えた際の振舞いの変化を示す例である．図は上段から下段にかけて環境変化し，左

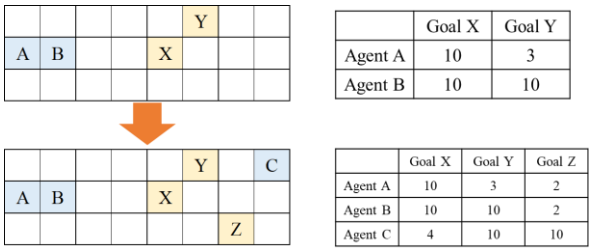

図 3　環境変化と獲得報酬期待値の変動

側はグリッドワールドの変化，右側は獲得報酬期待値の変化を表している．環境変化前はエージェント A, B が共にゴール Y へ到達しようと試み，エージェント A が報酬を獲得できずにゴール X へ到達するようになるため，票のようにエージェント A のゴール Y に対する期待値が小さくなる．環境変化後においては，エージェント C とゴール Z が新たに増えるが，ゴール Z においてはエージェント C が最も近く，他のエージェントにとっては到達し辛いため期待値が下がり，エージェント C にとってはゴール X が遠くにあり，環境変化前にエージェント A, B がゴール X へ到達するように学習しているため，ゴール X への期待値は小さくなる．この時閾値を 5 に設定すれば環境変化前と後で協調行動を獲得可能となる．つまり，環境変化後では，エージェント A がそれぞれゴール X, エージェント B がゴール X と Y，そしてエージェント C がゴール Y と Z のみを到達目標として選択するようになるため，エージェント A, B, C がそれぞれゴール X, Y, Z へ到達するように学習する．これはこの環境における最適解である．なお従来法ではすべてのゴールを考慮するため，エージェント C はゴール X へ到達するように学習し，その影響でエージェント B がゴール Z へ到達するように学習するため，余計に移動距離がかかる．以上が提案法の効果である．

5. 実験

5.1 実験設定

本研究における提案法の効果を検証するため，図 3 に示すグリッドワールドにおいて従来法 PMRL と比較実験を行う．本実験はすべてのエージェントが全ゴールに到達するまでのステップ数を評価する．各エージェントはそれぞれのゴール到達までに費やしたステップ数の最大値が小さくなる時もっとも良い状況として評価される．なお，学習と評価は別々に行い，各アルゴリズムで学習が終了した後，学習を抜いた状態で同じように学習反復を行い，その時のステップ数を評価する．そして，実験は疑似乱数の出力パターンが異なる（ランダムシードの異なる）30 試行を行い，その平均値で評価する．加えて，もし複数のエージェントが同じゴールへ到達してしまった場合，良い評価値となることを避けるためステップ数は 100 とする．

実験パラメータは，学習回数が 100000 回，環境変化は 50000 回時点で起こるものとする．また 1 学習における最大のステップ数を 100 に設定する．学習において初期 Q 値は 0，学習のパラメータは学習率αが 0.1，割引率γは 0.9 に設定する．内部報酬を設定するためのパラメータδは 10 に設定する．提案法において獲得報酬期待値の閾値は 5 に設定する．

5.2 実験結果

実験結果を図 4 に示す．図の縦軸は全エージェントがゴ

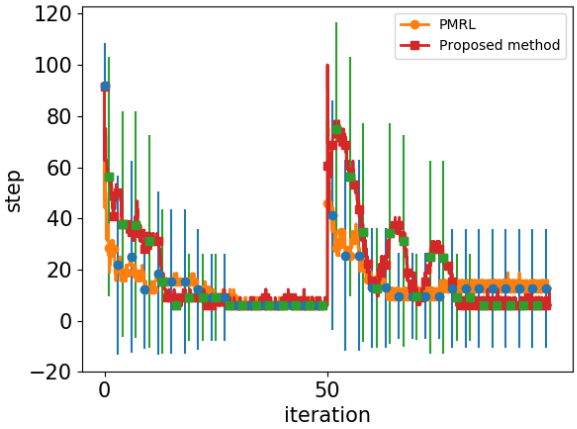

図 4　実験結果

ールへ到達するまでのステップ数，横軸は学習回数（目盛りは 1000 分の 1 の値）を示す．橙線が PMRL の結果，赤線が提案法を導入した結果である．ここでは目盛り 50 で環境変化が起こっている．図から PMRL は環境変化後にエラーバーが増え，ステップ数が多くなってしまっているのに対し，提案法では環境変化前も後も一定のステップ数でゴールに到達するように学習できていることがわかる．また提案法のステップ数は学習前も後もこの問題における最短ステップ数である 6 を示しており，最適方策を学習させることに成功している．

5.3 考察

図 4 から提案法により，エージェントの動的変化へ対応可能となっていることがいえる．また，図 5 は提案法を導入したことによる目的価値を示している．図 5 において，縦軸は目的価値，横軸は各ゴールを示す．青, 緑, 橙の棒グラフはそれぞれエージェント A, B, C の目的価値を示している．このグラフからわかる通り，エージェントは目的価値の最も大きなゴールを目指して学習するため，各エージェントは適切なゴールを選択できていることがわかる．特に PMRL であればエージェント C はゴール X の目的価値

が最も大きくなるはずであるが，それが抑えられており，提案法が正しく機能していることがわかる．

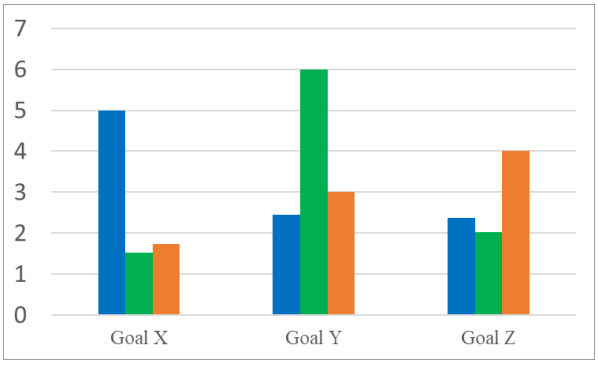

図 6　提案法の目的価値（環境変化後）

また，図 6 は提案法を導入したエージェントの学習した Q 値を示している．上段中段下段はそれぞれエージェント A, B, C の Q 値を示している．白, 紫, 緑のマスはそれぞれスタート，ゴール，通行可能な場所を示しており，矢印と数字がそれぞれ行動とその Q 値である．なお図では Q 値が大きければ大きいほど矢印も大きくなる．図を見ると，各エージェントについて，目的価値の最も大きなゴールへ向かう行動の Q 値が最も高く，スタート地点から最大の Q 値をたどれば適切なゴールに到達する．これは適切なゴールを選択後，それに到達できるように学習できていることを示している．以上から提案法によりエージェントが動的変化する状況においても適切な学習が可能となる．

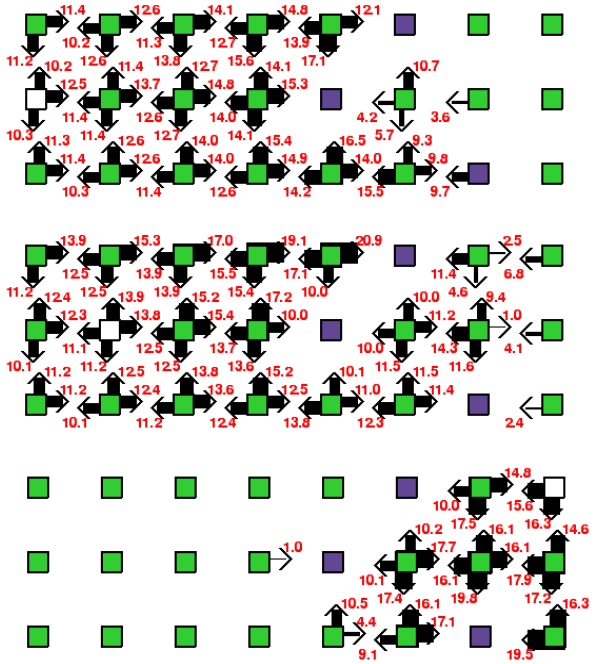

図 5　提案法の Q 値（上中下はエージェント A, B, C）

6. 終わりに

本研究は，現実問題におけるマルチエージェント強化学習の可能性を広げるため，エージェント数が動的変化する際にそれに適応し適切な協調行動が学習可能な手法を提案した．具体的には，従来手法 PMRL を拡張し，各エージェントの獲得報酬の期待値が閾値を超えた目的のみを考慮した協調行動の学習を行う手法を提案した．

実験ではエージェント数が 2 体から 3 体へ増えるグリッドワールドの問題を取り上げ，従来法として PMRL と性能比較を行った．結果として PMRL は環境変化前に協調行動を獲得していたが，環境変化後は学習が不安定になり，最短ステップ数でエージェントをゴールに到達させることができなかった．それに対し，提案手法は環境変化前も後も最短ステップ数でエージェントをゴールさせ，疑似乱数の出力パターンに影響されることなく安定して協調行動が学習可能であることがわかった．加えて，提案法では環境変化後も到達目的を適切に設定でき，学習結果である Q 値も適切なゴールへ到達可能なように設定できていることがわかった．これにより提案法はエージェントの動的変化に適応し協調行動の学習が可能であることが示された．

今回は簡単な問題において，従来法と提案法の性能を比較した．今後の予定としてはゴールへ到達するだけではなく，より現実に即したマルチタスクの問題への適用を検討している．

謝辞

本研究は JSPS 科研費 JP17J08724 の助成を受けたものです．

参考文献

[1] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. *In proceedings of the Tenth International Conference on Machine Learning*, pages 330-337. Morgan Kaufmann, 1993.

[2] Mohamed Elidrisi, Nicholas Johnson, Maria Gini, and Jacob Crandall. 2014. Fast adaptive learning in repeated stochastic games by game abstraction. In Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems (AAMAS '14). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1141-1148.

[3] Roberta Raileanu, Emily Denton, Arthur Szlam, and Rob Fergus. Modeling Others using Oneself in Multi-Agent Reinforcement Learning. Technical report, 2018.

[4] Sutton, R.S., Barto, A.G.: Introduction to Reinforcement Learning. MIT Press, Cambridge, MA, USA, 1st edn. (1998)

[5] Fumito Uwano, Naoki Tatebe, Yusuke Tajima, Masaya Nakata, Tim Kovacs, and Keiki Takadama. Multi-agent cooperation based on reinforcement learning with internal reward in maze problem. SICE Journal of Control, Measurement, and System Integration, 11(4):321–330, 2018.