4G1-3

# マルチエージェント環境における「ゴールを譲る行動」の強化学習

## Reinforcement learning of "a goal ceding behavior" in a multi-agent environment

長行 康男
Yasuo Nagayuki

大手前大学現代社会学部
Faculty of Modern Social Studies, Otemae University

The application of reinforcement learning to multi-agent systems has attracted recent attention. In multi-agent systems, it is important that the agents have the "sociality". In this article, we propose a reinforcement learning method, which is based on Q-learning, that the agent is able to learn "a goal ceding behavior", that is, "sociality" in a multi-agent environment. In this learning method, the agent learns to ignore the near goal, which is left for the other agent, and go toward the farther goal, if the agent judges that the decision is effective from the social viewpoint, but not from the agent's greedy viewpoint.

## 1. はじめに

マルチエージェント環境におけるエージェントの適応行動の実現は，工学及び認知科学の観点から興味深い研究課題である．その中でも，学習による適応行動の自律的獲得に関する研究が強化学習[Sutton 98]の発展を契機として注目を集めている．

ところで，マルチエージェント環境のなかには，強化学習が得意としないような環境（状況）も存在する．例えば，近くに存在するゴールを他エージェントに譲り，自分は遠くに存在するゴールに向かう行動を選択したほうが，社会性という観点からは良いような場合（状況）である．強化学習は基本的にゴールに向かって貪欲（greedy）な行動を学習（選択）する手法であり，「ゴールを譲る」という行動を強化学習で学習することは難しい．では，我々人間は，「ゴールを譲る」という社会性行動をどうやって学習しているのであろうか．おそらく，「ゴールに到達する」という利己的報酬と「ゴールを譲る」という社会的報酬の 2 つの報酬をうまく使い分け，「ゴールを譲る」という社会的報酬を利用して社会性行動を学習しているのではないかと思われる．

本稿では，「ゴールに到達する」という利己的報酬に加え，「ゴールを譲る」という社会的報酬を強化学習に導入することにより，近くに存在するゴールを他エージェントに譲り，遠くに存在するゴールに向かう社会性行動を獲得できるような Q 学習[Watkins 92]に基づいた新たな強化学習法を提案する．

## 2. 実験タスク

### 2.1 タスクの設定

「ゴールを譲る」という社会性行動を強化学習で学習できるかどうかを調査するため，本研究では，次のような設定の実験タスクを用意した．

- 2 次元（6×8）のグリッド空間中に，2 体のエージェント（図 1 中の A1 と A2）と 2 つの固定されたゴール位置（図 1 中の網掛け部（G1 と G2））が存在する．
- 各エージェントは，1 タイムステップごとに，上下左右のいずれかの隣接するセルに移動する 4 通りの行動を実行できるものとする．ただし，隣接するセルが存在しない方向への移動はできないものとする．例えば，図 1 の状態において，エージェント A1 は上，左への移動はできるが，下，右への移動はできない．

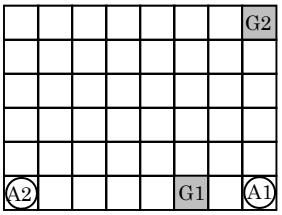
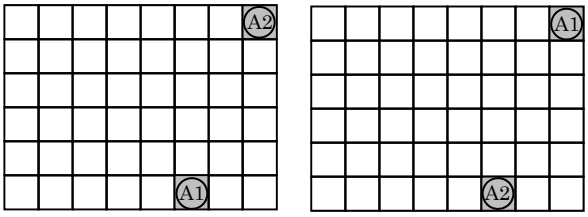



図 1 実験タスク

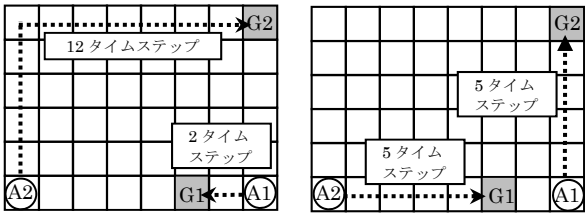
- 各エピソードにおいて，それぞれのエージェントの初期位置は図 1 の位置とする．
- それぞれのエージェントは，各タイムステップごとに同期して行動を実行するものとする．
- エージェントがいずれかのゴール位置に移動したとき，そのエージェントは正の報酬『1.0』を得る．ゴール位置以外のセルに移動したときの報酬は『0.0』とする．
- エージェントがゴール位置に到着すると，そのゴール位置の正の報酬『1.0』は消滅するものとする．すなわち，それぞれのゴール位置において，正の報酬『1.0』は，そのゴール位置に最初に到着したエージェントに 1 度のみ与えられる．
- 両エージェントが同一のセル上に移動してもよいものとする．両エージェントが同一のゴール位置に同時に移動した場合は，一方のエージェント（無作為に選択される）にのみ報酬『1.0』が与えられるものとする．
- ゴール位置に到着し，報酬『1.0』を獲得したエージェントは，そのエピソードでの行動と学習を終了し，もう 1 体のエージェントがもう一方のゴール位置に到着するまで待機するものとする．
- このタスクにおける（グローバルな）目標は，G1，G2 の両方のゴール位置に，それぞれどちらか 1 体のエージェントが到着することとする．すなわち，図 2(a)の状態か，図 2(b)の状態のいずれかになったとき，このタスクの目標は達成されたものとし，タスク（1 エピソード）が終了するものとする．
- 本稿では，図 2(a)，図 2(b)の状態を，それぞれ目標達成状態 1，目標達成状態 2 と呼ぶことにする．

連絡先：長行康男，大手前大学現代社会学部
〒664-0861 兵庫県伊丹市稲野町 2-2-2
Tel：072-770-6334，E-mail：nagayuki@otemae.ac.jp




(a) 目標達成状態 1　　(b) 目標達成状態 2

図 2　実験タスクにおける 2 つの目標達成状態




(a) 目標達成状態 1 の場合　　(b) 目標達成状態 2 の場合

図 3　初期位置から目標達成状態までのそれぞれの
エージェントの最短経路（最短タイムステップ）例

### 2.2 タスクの意味

ここで, それぞれのエージェントが, 初期位置(図 1 中のそれぞれの位置)から各ゴール位置(G1, G2)に最短経路で到着した場合, 図 2(a)の目標達成状態 1 と図 2(b)の目標達成状態 2 のどちらの目標達成状態がより効率的かを考える.

目標達成状態 1(図 2(a))の場合, 図 3(a)に示すように, エージェント A1 は最短で 2 タイムステップでゴール位置 G1 に到着でき, エージェント A2 は最短で 12 タイムステップでゴール位置 G2 に到着できる. したがって, 初期状態(初期配置)から, この目標達成状態 1 に至るまでの最短のタイムステップ数は 12 である(エージェント A1 はエージェント A2 がゴール位置 G2 に到着するまで待たされる).

一方, 目標達成状態 2(図 2(b))の場合, 図 3(b)に示すように, エージェント A1 は最短で 5 タイムステップでゴール位置 G2 に到着でき, エージェント A2 も最短で 5 タイムステップでゴール位置 G1 に到着できる. したがって, 初期状態(初期配置)から, この目標達成状態 2 に至るまでの最短のタイムステップ数は 5 である.

以上より, もしそれぞれのエージェントが最短経路でそれぞれのゴール位置に移動した場合, 目標達成状態 1 より目標達成状態 2 の方が, より速く目標達成状態に到達できるため, より効率的であるといえる. しかしながら, 目標達成状態 2 となるためには, エージェント A1 は近くにあるゴール位置 G1 をエージェント A2 に譲り, より遠くのゴール位置 G2 に向かわなければならない. つまり, エージェント A1 に社会性がなければ, 効率的である目標達成状態 2 は達成できない.

## 3. マルコフ決定過程と Q 学習

本稿で提案する強化学習法は Q 学習[Watkins 92]に基づいたものである. Q 学習はマルコフ決定過程[Puterman 94]を対象として提案された強化学習法である.

### 3.1 マルコフ決定過程

マルコフ決定過程[Puterman94]は, 環境状態の有限集合 S とエージェントの行動の有限集合 A によって定義される. 各離散タイムステップにおいて, エージェントは環境の状態 s ∈S を観測し, 行動 a ∈A を実行する. そして, 環境の状態は s' ∈S に遷移し, エージェントは環境から報酬 r を受け取る. ここで, 環境の状態遷移は, 状態遷移確率関数

$$P_{ss'}^{a} = \Pr(s' \mid s, a) \qquad (1)$$

によって定義される. ここで, Pr(s'| s, a) はエージェントが状態 s で行動 a を実行したときに状態が s' に遷移する確率を表す. エージェントが環境から受け取る報酬 r も確率的で, その確率は状態 s と行動 a のみに依存する.

### 3.2 Q 学習

Q 学習[Watkins92]では, Q 関数と呼ばれる関数 Q(s, a) をもとに行動選択, 行動学習を行う. ここで, Q(s, a) は状態 s で行動 a を実行する価値を表すスカラー関数で, 値が大きいほど, その状態でその行動を実行することが効果的(将来に多くの報酬が得られることが期待できる)ことを表す.

Q 学習では, すべての状態 s ∈S, 行動 a ∈A に対する Q 関数値 Q(s, a) を任意の初期値に設定して学習を開始し, エージェントが試行錯誤の経験を通して Q 関数を更新することにより学習を進行する. Q 学習における学習の流れは次の通りである.

1. 現在の環境の状態 s において, エージェントは, 確率ε ∈ [0,1] で行動集合 A の中から無作為に行動を選んで実行し(実行した行動を a とする), 確率1 − ε で式(2)を満たす行動 a を実行する.

$$a = \arg\max_b Q(s, b) \qquad (2)$$

2. エージェントが手続き 1 で行動 a を実行することにより, 環境の状態は(状態遷移確率関数 $P_{ss'}^{a}$, に従って)s' に遷移し, エージェントは環境から報酬 r を受け取る. そして, エージェントは状態 s, 行動 a に対する Q 関数値 Q(s, a) を式(3)に従って更新(学習)する.

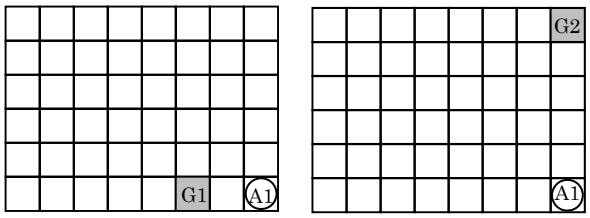
$$Q(s, a) \leftarrow (1-\alpha)Q(s, a) + \alpha(r + \gamma \max_{a'} Q(s', a')) \qquad (3)$$

ここで, α∈(0,1], γ ∈[0,1]は, それぞれ学習率, 割引率と呼ばれるパラメータである.

3. 学習の終了条件を満たしていれば学習終了し, そうでなければ, s' を s に代入して, 手続き 1 に戻る.

ここで, 2.1 で定義した実験タスクにおいて, それぞれのエージェントに Q 学習をそのまま適用した場合, Q 学習の特性より, エージェント A1 はゴール位置 G2 に向かうという社会性行動を学習することができず, より近くに存在するゴール位置 G1 に向かう行動を学習することが予想される. エージェント A2 も環境(近いゴール)に対して貪欲であり, より近いゴールであるゴール位置 G1 に向かおうとするはずであるが, ゴール位置 G1 はエージェント A1 に先に取られてしまい, ゴール位置 G1 での正の報酬は消滅する. それにより, エージェント A2 は, 正の報酬が残っているゴール位置 G2 に向かう行動を学習することになると予想される. つまり, 社会的観点で効率的ではない目標達成状 1 に学習が収束してしまうことが予想される.

実際に実験を行ってみた結果を表 1 の『通常の Q 学習』の列に示す. 100 回のコンピュータシミュレーション実験中 100 回とも, 社会的観点で効率的ではない目標達成状態 1 に学習が収束するという結果になった.




(a) G1 用分割タスク空間　(b) G2 用分割タスク空間

図 4　A1 の観点からのゴールごとに分割されたタスク空間

## 4. 提案手法

エージェント A1 がゴール位置 G1 をエージェント A2 に譲り，より遠くにあるゴール位置 G2 を目指すといった社会性行動を学習しない限り，効率的な目標達成状態 2 に学習は収束しない．

本稿では，近くのゴールを他エージェントに譲るという社会性行動を学習し，2.1 で定義したタスクにおいて，目標達成状態 2 に学習が収束するような Q 学習に基づいた強化学習法を提案する．

まず，それぞれのエージェントにおいて，自分から見たグリッド空間を，ゴールごとのタスク空間に分割する．ここでのタスク空間分割には，Whitehead らがシングルエージェント環境におけるマルチタスクの強化学習においてタスク空間分割を行った手法[Whitehead 93]と同様の手法を採用する．この手法により，例えば，図 1 中のエージェント A1 から見たグリッド空間は，2 つのゴール位置 G1，G2 ごとに，それぞれ図 4(a)，図 4(b)のような 2 つのタスク空間に分割される．

本稿で提案する手法では，エージェントは，分割タスク空間ごと(ゴールごと)に別々の Q 関数を持ち，それぞれの Q 関数で別々に通常の Q 学習を行う．例えば，A1 は，G1 用分割タスク空間(図 4(a))の Q 関数で G1 へ移動する行動を学習し，G2 用分割タスク空間(図 4(b))の Q 関数で G2 へ移動する行動を学習する．ここで重要となってくるのが，行動選択時に，どちらの分割タスク空間の Q 関数を利用するかである．本稿で提案する手法では，学習中の

『エージェント Ai(i=1,2) が Gj(j=1,2) に位置して目標達成状態を達成したときの，初期配置から目標達成状態達成までに費やしたタイムステップ数の最小値($ms_{ij}$とする)』

の値を利用する．エージェント Ai は，$ms_{ix}$ の値と $ms_{iy}$ の値(x=1 のとき y=2, x=2 のとき y=1)を比較し，$ms_{ix}<ms_{iy}$ のとき，ある確率 p で Gy 用分割タスク空間の Q 関数を利用して行動選択をし，確率 $1-p$ で Gx 用分割タスク空間の Q 関数を利用して行動選択をする($ms_{ix}=ms_{iy}$ のときは，確率 1/2 でどちらかの Q 関数を選択)．ここで，確率 p の値は，学習エピソード数に応じて減衰させるものとする．

各エージェントは，各エピソード開始時に上記確率でどちらのゴール用分割タスク空間の Q 関数を利用するか決定し，そのエピソードの間は，決定した方の Q 関数を行動選択時に利用し続けるものとする(行動学習(Q 関数の更新)はすべての行動において，両方の Q 関数で更新を行う)．

## 5. 実験

### 5.1 実験結果

2.1 で定義した実験タスクの両エージェントに対して，「通常の Q 学習」と「提案手法」の 2 つの強化学習法を適用し，コンピュータシミュレーション実験を行った．学習時・行動選択時に使用したパラメータの値は，学習率α=0.2，割引率γ=0.9，ε=0.5×$0.9977^{num\_ep}$，p=0.5×$0.9977^{num\_ep}$ である．ここで，num_ep は学習エピソード数を表し，0.9977 という値は，$0.9977^{1000}$ が約 0.1 となるように選ばれた値である．

表 1　学習による各目標達成状態への収束状況
(シミュレーション実験数：それぞれの学習法で 100 回ずつ)

| | 通常の Q 学習 | 提案手法 |
|---|---|---|
| 目標達成状態 1 | 100 回 | 0 回 |
| 目標達成状態 2 | 0 回 | 100 回 |

表 2　提案手法における，学習後のそれぞれのエージェントのゴール到達までのタイムステップ数とその回数
(シミュレーション実験数：100 回)

| タイムステップ数 | | 回数 |
|---|---|---|
| A1 | A2 | |
| 5 | 5 | 73 回 |
| 5 | 7 | 18 回 |
| 7 | 5 | 8 回 |
| 5 | 9 | 1 回 |

「通常の Q 学習」と「提案手法」のそれぞれで，乱数の種を変えて 100 回のシミュレーション実験を行った．そして，学習結果が十分に安定する 10000 エピソードほど学習させたときに，学習がどちらの目標達成状態に収束したかを表 1 に示す．表 1 より，「通常の Q 学習」では 100 回すべて目標達成状態 1(社会的観点で，効率的ではない状態)に収束し，「提案手法」では 100 回すべて目標達成状態 2(社会的観点で，効率的な状態)に収束していることがわかる．また，目標達成状態 2 に 100 回すべて収束した提案手法において，それぞれのエージェントが，何歩(何タイムステップ)でゴールに到達したかを表 2 に示す．表 2 より，100 回中 73 回は，両エージェントとも最短の 5 タイムステップでゴールに移動していることがわかる．ただ，残り 27 回は目標達成状態 2 に収束しているものの，無駄な行動を行っていることがわかる．学習ログをチェックした結果，学習がある程度安定した後にも，0 でない確率 p で A1 が G1 を目指す場合があり，それ以降，学習が不安定になってしまい，最短経路を見失ってしまっている場合があることがわかった．

## 6. おわりに

本稿では，マルチエージェント環境において，近くに存在するゴールを他エージェントに譲り，遠くに存在するゴールに向かう社会性行動を獲得できるような Q 学習に基づいた新たな強化学習法を提案した．そして提案手法により，エージェントが他エージェントに身近なゴールを譲るという社会性行動を学習できることを実験的に示した．

今後の課題は，本手法をエージェント数やゴール数を増やした環境などに適用し，本手法の汎用性の検証を行うことである．

## 参考文献

[Puterman 94] Puterman, M. L.: Markov Decision Processes, Wiley Interscience (1994)

[Sutton 98] Sutton, R. S. and Barto, A. G.: Reinforcement Learning: An Introduction, MIT Press (1998)

[Watkins 92] Watkin, C. J. C. H. and Dayan, P.: Technical note Q-learning. Machine Learning 8(3):279-292 (1992)

[Whitehead 93] Whitehead, S., Karlsson J. and Tenenberg, J.: Learning Multiple Goal Behavior via Task Decomposition and Dynamic Policy Merging, Robot Learning, pp.45-78, Kluwer Academic Publishers (1993)