# 二個体間協調に基づく重みづけ行動評価によるマルチエージェント逆強化学習

発表者: I 類 メディア情報学プログラム 学籍番号 1910094 植木駿介
指導教員: 髙玉圭樹 教授

## 1 はじめに

複数のエージェントが相互作用するマルチエージェントシステムにおいて，マルチエージェント強化学習は有効な機械学習手法の一つである．最適な振る舞いの獲得は，報酬関数に大きく依存するが，意図した行動を導くような報酬関数の設計は困難である．この問題に対して，最適な振る舞いをするエキスパート行動を再現するような報酬関数を推定することで解決を図るマルチエージェント逆強化学習がある．従来手法の多くはエキスパート行動が最適であると仮定し，エージェントが目指す共通の解概念をナッシュ均衡解などとしている．しかし，最適なエキスパート行動の獲得や，ナッシュ均衡解のような全エージェントを考慮することは環境の複雑化，エージェント数の増加により困難を極める．

そこで，本研究では，各エージェントの観点では最適だが，全体の観点では干渉するために最適ではない非最適なエキスパートを行動入力として，最適な報酬関数の獲得を目指す．この目的達成に向け，二個体間の協調行動に着目し，行動系列をアーカイブ・評価し，その中で評価値が最大の行動系列でエキスパート行動を更新することにより，最適な報酬関数の獲得を目指す Two- individuals Cooperative Multi-Agent Inverse Reinforcement Learning（TC-MAIRL）を提案する．また，TC-MAIRL の改良手法として，行動系列を重みづけて評価する Weighted TC-MAIRL（WTC-MAIRL）を提案する．

## 2 逆強化学習

逆強化学習の代表的な手法である Maximum Entropy Inverse Reinforcement Learning（MaxEntIRL）[1] は，逆強化学習に最大エントロピーの原理を適用することで，観測したエキスパートの行動軌跡を導く確率を最大化し，未観測の不確実性が高い状態は一様な確率を与える報酬関数を推定する．報酬関数 $R(s)$ は $R(s) = \theta^T \phi(s)$ で定義され，$\theta$ は報酬関数のパラメータであり，$\phi(s)$ は one-hot ベクトルである．訪れた一連の状態を行動系列 $\zeta$ と呼ぶ．強化学習（Inner loop）を実行して獲得した方策 $\pi_\theta(a|s)$ からエージェントがある行動系列を実行する確率 $P(\zeta|\theta)$ を求める．エキスパートの行動系列群 $Z$ を実行するエントロピー最大化は $\max \sum_{\zeta \in Z} P(\zeta|\theta) log P(\zeta|\theta)$ であり，ラグランジュ未定乗数法により求めた解は $P(\zeta|\theta) = \exp(\theta^T f_\zeta) / \sum_{\zeta \in Z} \exp(\theta^T f_\zeta)$ である．これを尤度関数とし，最適なパラメータ $\theta^*$ は対数尤度の勾配を最大化することで求める．勾配は $\nabla L(\theta) = \frac{1}{M} \sum_{\zeta \in Z}^{M} f_\zeta - \sum_{s \in \zeta} P(s_i|\theta) f_{s_i}$ で求められ，$f_\zeta$ は行動系列 $\zeta$ の特徴量であり，$f_\zeta = \sum_{s \in \zeta} \phi(s)$ で求められる．

## 3 提案手法

提案手法のアーキテクチャを図 1 に示す．TC-MAIRL は，従来手法の MaxEntIRL をベースとし，行動系列のアーカイブ・行動系列の評価・エキスパート行動の更新の操作を追加する．WTC-MARIL はさらに，関連度の計算の操作を追加し，関連度に重みづけて行動系列を評価する．

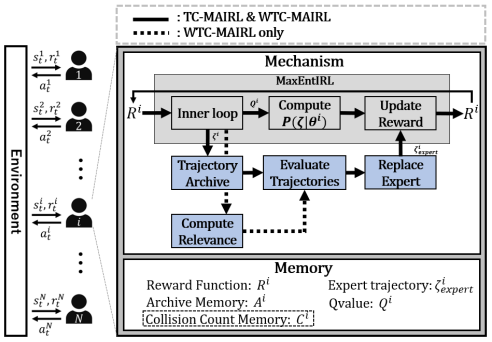


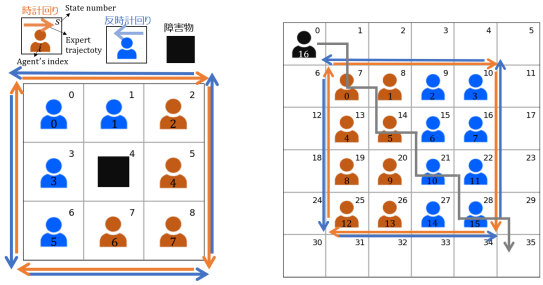

**図 1:** TC-MAIRL と WTC-MAIRL のアーキテクチャ

### 3.1 行動系列のアーカイブ

エージェントは行動系列や，その情報を保存できるメモリ $A$ を持つ．Inner loop において，実行した行動系列が有用な行動系列の場合，メモリ $A$ にアーカイブする．また，獲得した行動系列がすでにメモリ $A$ 内にある場合は，その行動系列を実行した回数と，各エージェントの衝突回数，または非衝突回数を更新する．

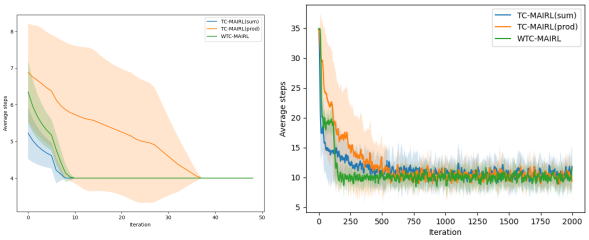
### 3.2 関連度の計算

エージェントは他の各エージェントと衝突した回数を記録するメモリ $C$ を持つ．そして，Inner loop で獲得した Q 値に基づき，全エージェントが greedy に行動した際の各




**(a)** 迷路問題 1　　**(b)** 迷路問題 2

**図 2:** 実験環境




**(a)** 迷路問題 1　　**(b)** 迷路問題 2

**図 3:** 平均ステップ数

エージェントの衝突回数を更新する．関連度はメモリ $C$ に記録された衝突回数のうち各エージェントの割合と定義し，Agent$i$ の Agent$j$ との関連度 $c^{i,j}_{agent}$ は，$c^{i,j}_{agent} = \frac{C^{i,j}}{\sum_{k=0}^{N_{agent}-1} C^{i,k}}$ で求められる．

### 3.3　行動系列の評価・エキスパート行動の更新

行動系列の評価について，TC-MAIRL では，1．評価値を全エージェントの非衝突率の平均値とする方法（式（2））と 2．全エージェントの非衝突率の総積とする方法（式（2））がある．WTC-MAIRL では関連度で重みづけした評価値（式（3））で計算する．

$$Eval_{sum}(\zeta_i^k) = \frac{1}{N_{agent}} \sum_{j=0,j\neq i}^{N_{agent}-1} \frac{N_{non-col}^{\zeta_i^k,j}}{N_{col}^{\zeta_i^k,j} + N_{non-col}^{\zeta_i^k,j}} \tag{1}$$

$$Eval_{prod}(\zeta_i^k) = \Pi_{j=0,j\neq i}^{N_{agent}-1} \frac{N_{non-col}^{\zeta_i^k,j}}{N_{col}^{\zeta_i^k,j} + N_{non-col}^{\zeta_i^k,j}} \tag{2}$$

$$Eval_{rel}(\zeta_i^k) = \sum_{j=0,j\neq i}^{N_{agent}-1} c_{agent}^{i,j} \frac{N_{non-col}^{\zeta_i^k,j}}{N_{col}^{\zeta_i^k,j} + N_{non-col}^{\zeta_i^k,j}} \tag{3}$$

ここで，$N_{agent}$ はエージェント数，$\zeta_i^k$ は Agent$i$ がアーカイブした $k$ 番目の行動系列であり，$N_{col}^{\zeta_i^k,j}, N_{non-col}^{\zeta_i^k,j}$ は，それぞれ $\zeta_i^k$ の行動をしたとき，Agent$j$ と衝突・非衝突した回数である．最後に，エキスパート行動の更新では，評価値が最も高い行動系列をエキスパート行動に更新する．

## 4　実験

### 4.1　実験内容

提案法の有効性を検証するために，2 つの迷路問題（図 2a，図 2b）を用いる．与えるエキスパート行動は図中の矢印で示す．全エージェントが最適な行動を取った場合，各迷路問題の平均の最短ステップ数はそれぞれ 4steps，5.29steps である．

### 4.2　結果と考察

図 3 に各迷路問題の全エージェントの平均ステップ数の変化を示す．行動系列は TC-MAIRL（sum）では評価方法 1，TC-MAIRL（prod）では評価方法 2 で評価している．TC-MAIRL（sum）は行動系列の評価を評価方法 1，TC-MAIRL（prod）は評価方法 2 である．迷路問題 1 では最短ステップ数に到達し最適な報酬関数が獲得できている．迷路問題 2 は最適なエキスパート行動に更新できていたが，Q 学習が衝突回避する最適行動よりステップ数が少ない局所解に陥りやすいため，安定して最適行動に収束していない．TC-MAIRL（sum）は，関係の薄いエージェントが多くなると，少数の関係の深いエージェントがほとんど考慮されない評価値となってしまうため迷路問題 2 では収束速度が WTC-MAIRL より遅くなったと考えられる．TC-MAIRL（prod）の収束速度が遅い原因は，偶発的に衝突しない限りエージェントとの非衝突率が 0.0 になり，評価値が 0.0 になりやすいためである．WTC-MAIRL は迷路問題 1 では，関連度の変化が TC-MAIRL（sum）より収束が遅れた原因であり，迷路問題 2 では関連度による重みづけ評価により適切に評価でき，収束速度が他手法よりも早くなったと考えられる．

## 5　おわりに

本研究では，二個体間の関係に着目して，行動系列のアーカイブ・関連度に重みづけて行動系列を評価し，エキスパート行動を更新することで非最適な行動系列から最適な報酬関数を獲得した．2 つの迷路問題で実験し，提案手法が有効であることを示した．今後の課題は，連続環境に適応することである．

## 参考文献

[1] B.D.Ziebart et al. Maximum entropy inverse reinforcement learning. In *23rd AAAI Conf on Artificial Intelligence*, pp. 1433–1438, 2008.