4C2-1

# 協調行動の獲得に向けた逆強化学習の導入

Acquisition of Cooperative Behavior via Introducing Inverse Reinforcement Learning

本木 雄斗 *1　荒井 幸代 *1
Yuto Motoki　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 multiagent reinforcement learning problem, it is difficult to design reward function which make agents to learn optimal policy. Inverse Reinforcement Learning(IRL) is the framework to estimate the reward function from optimal policy. We take a multiagent path planning problem and introduce Joint-state learners, which receive the location of the agents as state information but choose their actions independently. Joint-state learners sometimes falls into local optimal solution, because of the two problems, concurrent learning and perceptual aliasing. We propose two methods to avoid these problems under estimated reward function. First, each agent learns in the environment that the other agents act optimally. Second, agents learn at the same time after make agents to act optimally in some episodes.

## 1. はじめに

本論文では，各エージェントが目標状態への最適性を追求すると競合が生じる問題を対象とする．この競合が生じる状態を「干渉状態」とよび,「干渉状態」において各エージェントが折り合いをつけ，全体として最適な行動を学習させる方法を提案する．

提案法では，逆強化学習 [Abbeel 04] によって得られる報酬関数に着目する．はじめに，干渉状態を検出する．次に，マルチエージェント系の同時学習の影響を抑制し，局所解への収束を回避する．

## 2. マルチエージェント系のモデリング

### 2.1 マルコフゲーム

マルコフゲームは，マルコフ決定過程 (Markov decision process; MDP) を，$n$ 人のエージェントが存在するゲーム状況に拡張したモデルで，確率ゲームともよばれる．

マルコフゲームは，一般にタプル $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}_1, ..., \mathcal{A}_n, \mathcal{P}, \mathcal{R}_1, ..., \mathcal{R}_n \rangle$ で表現される．$\mathcal{N}$ はエージェント集合（エージェント数 $n = |\mathcal{N}|$），$\mathcal{S}$ はエージェントが観測する状態集合，$\mathcal{A}_1, ..., \mathcal{A}_n$ の各項 $\mathcal{A}_i$ はエージェント $i$ の行動集合を示す．$\mathcal{P}$ は遷移先の確率分布の集合，$\mathcal{R}_1, ..., \mathcal{R}_n$ の各項 $\mathcal{R}_i$ は状態遷移後にエージェント $i$ が獲得した報酬の集合を示す．なお，$n = |\mathcal{N}| = 1$ のマルコフゲームは，MDP$\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$ と等価である．

### 2.2 協力型問題と非協力型問題

ゲーム理論におけるゲームの分類の一つであるゼロ和ゲームの枠組みに従って協力型問題と非協力型問題を定義する．ゼロ和ゲームでは，意思決定主体であるプレイヤーの行動による利得の総和がゼロになる，すなわちエージェント間の利害が相反している．本論文では，この状況を競合とよぶ．マルコフゲームにおいては，初期状態からタスク達成までの状態遷移のうち，競合を含む問題を非協力型問題と定義する．一方，初期状態からタスク達成までの状態遷移のうち，競合を含まない問題を協力型問題と定義する．

マルチエージェント系の強化学習のテストベッドとして用いられる問題のうち，追跡問題やマルチエージェント倉庫番は協力型問題に，狭路すれ違い問題や経路計画問題は非協力型問題に分類できる．

### 2.3 マルチエージェント強化学習

マルチエージェント強化学習は，複数のエージェントが存在する環境で，各エージェントの適切な行動を学習する枠組みである．各エージェントは環境の局所的な情報に基づいて，自身に与えられた目標を達成する行動ルールを獲得する一方，系全体の最適化が達成されることが望ましい．これに対して，タスクを達成した際の報酬さえ設定できれば適切な方策の獲得が見込まれるマルチエージェント強化学習は，マルチエージェント系への応用が期待されるアルゴリズムである．

一方，マルチエージェント強化学習では，状態空間の爆発問題，同時学習問題，不完全知覚問題といったマルチエージェント系特有の問題が生じることが知られている [Arai 01]．

## 3. 対象問題

### 3.1 関連研究

**■報酬設計**

逆強化学習は，最適な行動軌跡や状態遷移確率を所与として，報酬関数を求める問題として定義される [Russell 98]．[Ng 00] は，有限状態空間をもつ環境に対しては線形計画法，無限の状態空間をもつ環境に対してはモンテカルロ法を用いて，報酬関数を推定する手法を提案している．マルチエージェント系に対しては，[Natarajan 10] が，Ng らによる報酬関数を，複数エージェントの報酬関数によって構成されたベクトルを用いて推定し，系全体の挙動を制御する手法を提案している．

**■協力型問題における収束性**

[Arai 15] は，追跡問題やマルチエージェント倉庫番を対象として，各エージェントの報酬関数の設計に逆強化学習 [Abbeel 04] を導入し，得られた報酬関数に基づいた逐次的な報酬を与えることによって局所解に陥ることを回避している．具体的には，最適な行動軌跡から各エージェントの報酬関

連絡先: 本木雄斗，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，千葉市稲毛区弥生町 1-33，adaa2063@chiba-u.jp

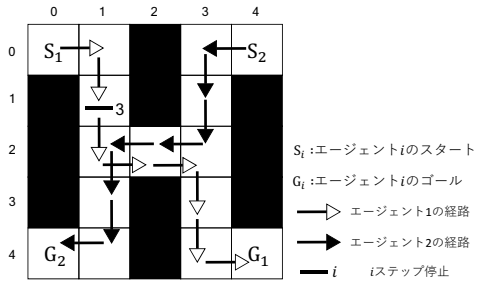


図 1: 実験環境とエキスパートの行動軌跡

数を求めた後，各エージェントはそれぞれの報酬関数にしたがって「同時」に学習する設定の下で，最適解に収束することを実験的に示している．

■非協力型問題における収束性

[Kyo 14] は Independent learners や Joint-state learners を用いるだけでは最適解に収束しない非協力型問題における改善法を示している．Independent learners とは，各エージェントが他エージェントの状態を考慮せず，自らの状態だけを知覚し，独立に学習する手法である．一方，Joint-state learners とは，他エージェントの状態を観測し，他エージェントの状態と自らの状態との組み合わせを状態入力とする手法である．ただし，他エージェントの行動は把握できない．

提案法は，他エージェントの行動が系全体の最適解獲得に対して影響する状態を「干渉状態」と定義し，これを抽出する方法と，干渉状態と非干渉状態における学習方法の 2 段階からなる．干渉状態を抽出する際，この状態では方策が不安定になることに着目し，情報理論のエントロピーを用いて Q 値の変動を計算している．また，非干渉状態では自らの状態だけを知覚する一方，干渉状態では他エージェントの状態も知覚することにより，最適解を獲得している．

これらに対して本論文では，非協力型問題において，逆強化学習により得られた報酬関数を用いて協調行動を獲得させることによって，系全体の最適解への収束を目指す．

## 3.2 予備実験

### 3.2.1 実験環境と設定

■実験環境

本論文では，マルチエージェント経路計画問題を用いる．マルチエージェント経路計画問題とは，各エージェントがそれぞれのスタートからゴールまでの最短経路を見つける問題である．しかし，各エージェントがゴールまでの最短経路をとろうとするとエージェント同士が衝突する．衝突を回避するためには，一方が前進し，もう一方がこれを譲るという協調行動が必要となる．

図 1 に実験環境を示す．5×5 の格子環境にエージェント $i$ のスタート $S_i$，エージェント $i$ のゴール $G_i$ がそれぞれ存在する．黒い部分は壁を表し，通過できない．格子周辺の数字は，各軸に対する位置を表す．以後，横軸の位置 $x$，縦軸の位置 $y$ を座標 $(x, y)$ と表記する．本実験環境における最短ステップ数は 11 ステップである．各エージェントの初期状態は，それぞれのスタートに初期配置され，タスク達成までを 1 エピソードとする．またエピソード内にステップ数の上限を設け，上限を超えた場合も 1 エピソードの終了とする．

■エージェントの設定

エージェントは，{ 上，下，右，左，停止 } の 5 つから行動を選択し，実行する．なお，各エージェントへの状態入力やエージェントの行動はそれぞれ 1 ステップに一度，全エージェント同時に行われる．また行動を実行した結果，あるエージェントが他エージェントや壁に衝突した場合，そのエージェントを 1 ステップ前の位置に戻す．

表 1: Joint-state learners，Independent learners の収束回数 [回]

| 手法 | 11step | 13step | 14∼18step | 発散 |
|---|---|---|---|---|
| Independent learners | 0 | 0 | 0 | 100 |
| Joint-state learners | 0 | 76 | 7 | 17 |

### 3.2.2 不完全知覚問題の影響

3.1 節の Independent learners，Joint-state learners の両者で不完全知覚が生じることを示す．

■実験設定

1 回の実験 (以下，各エージェントが同時に学習する 1 回の実験を 1 試行とよぶ) ではエピソード数を 2000，ステップ数の上限を 400 ステップとしてエージェントに学習させる．1 試行の最終エピソードにおいて，タスク達成までのステップ数が上限の 400 ステップに達した場合，学習は発散したとする．

各エージェントの行動選択には $\epsilon$-greedy を用いる．学習開始から 2000 エピソードまでは，$\epsilon$ を段階的に小さくし，最終エピソードである 2000 エピソードでは $\epsilon$ を 0.0 に固定する．また，学習率 $\alpha$=0.0001，割引率 $\gamma$=0.9 とする．

エージェントへの報酬は，両エージェントがゴールに到達した場合，両エージェントに正の報酬 100 を，壁に衝突した場合，衝突したエージェントだけに負の報酬-10 を，他エージェントと衝突した場合，各エージェントに負の報酬-4 を与える．

■実験結果と考察

表 1 に Joint-state learners と Independent learners 各場合の 100 試行の学習結果を示す．

Independent learners では，全試行で発散しており，各エージェントは初期状態で停止行動をとっていた．これは，壁や他エージェントと衝突時の負の報酬の影響であると考えられる．特に，Independent learners では他エージェントの状態を考慮していないため，例えば状態 (2, 2) では，他エージェントと衝突した場合と，していない場合が区別できないにも関わらず，負の報酬を受けるときと受けないときが存在する．このとき各エージェントは移動すると負の報酬を受ける可能性が高くなるため，各エージェントは初期状態で停止行動をとり続けると考えられる．

一方，Joint-state learners では，76 試行で最短から 2 ステップ多い 13step に収束した．また，学習が発散する場合も 17 試行あった．発散したときエージェント 1 は状態 (1, 2) で，エージェント 2 は状態 (3, 2) で停止行動をとっていた．これは 2 エージェントのうち，一方が前進し，もう一方が譲ることを学習する必要のある実験環境において，両エージェントの譲り合う状況が生じているためと考えられる．最短である 11 ステップに収束しない，あるいは学習が発散する原因は他エージェントと衝突時の負の報酬の影響であると考えられる．

この収束性の問題は，他エージェントの状態と行動を観測できれば回避可能であると考えられる．しかし，他エージェントの行動まで観測することは現実的に困難である上，エージェント数が増加すると状態空間の爆発が生じる可能性が高まる．そこで本論文では，以後 Joint-state learners を用いる．

表 2: 逆強化学習パラメータ

| パラメータ | 数値 |
|---|---|
| IRL 割引率 | 0.90 |
| RL 割引率 | 0.90 |
| RL 学習率 | 0.0001 |
| RL 探索率 | 1.0 から段階的に下げる |
| RL エピソード数 | 10000 |
| 終了条件 | 0.000001 |

表 3: 荒井らの手法の収束回数 [回]

| 11step | | 12~13step | 発散 |
|---|---|---|---|
| 軌跡一致 | 軌跡不一致 | | |
| 7 | 48 | 43 | 2 |

#### 3.2.3 同時学習問題の影響

逆強化学習を用いて推定した報酬関数に基づき，Joint-state learners を用いて同時に学習 [Arai 15] しても，同時学習問題が生じることを示す．

■実験設定

図 1 に逆強化学習において所与であるエキスパートの行動軌跡を示す．また，表 2 に逆強化学習におけるパラメータをまとめる．エキスパートの行動軌跡とこれらのパラメータは，5 章の計算機実験でも用いる．

Joint-state learners を用いて同時に学習する際，設定したエピソード数を終えるまでを 1 試行とし，1 回の実験では，逆強化学習を用いて報酬関数を推定し，得られた報酬関数に基づき各エージェントは 1 試行同時に学習する．同時に学習する際，エピソード数 100000，ステップ数の上限を 400 ステップとしてエージェントに学習させる．1 試行の最終エピソードにおいて，タスク達成までのステップ数が上限の 400 ステップに達した場合，学習は発散したとする．

各エージェントの行動選択には $\epsilon$-greedy を用いる．学習開始から 100000 エピソードまでは，$\epsilon$ を 0.3 から段階的に小さくし，最終エピソードである 100000 エピソードでは $\epsilon$ を 0.0 に固定する．学習率 $\alpha$=0.0001，割引率 $\gamma$=0.9 とする．これらのパラメータは実験的に設定した．

■実験結果と考察

表 3 に 100 回の実験したときの学習結果を示す．表 3 より最適な行動軌跡と一致しない場合が多いだけでなく，発散する場合もあることが確認できた．この原因として二つが影響し，報酬獲得と無関係な行動が強化されたためと考えられる．一つ目は，Abbeel の逆強化学習をマルチエージェント系に適用する際，他エージェントに最適行動をとらせ，MDP を仮定する一方，得られた報酬関数に基づき同時に学習する際は非 MDP となる．この環境の違いのため，エキスパートの軌跡から外れる回数が多くなる点である．二つ目は，Abbeel の逆強化学習ではエキスパートの軌跡上の報酬関数だけしか推定できない点である．特に，この一つ目において各エージェントが同時に学習する際，同時学習問題が影響していると考えられる．

### 3.3 問題設定

予備実験より，Joint-state learners を用いても，同時に学習すると不完全知覚問題や同時学習問題が生じ，最短である 11 ステップに収束しないことが確認できた．そこで本論文では，Abbeel の逆強化学習を用いて推定した報酬関数に基づき，Joint-state learners を用いて同時に学習するとき，同時学習の影響を抑制し，協調行動を獲得させる方法を提案する．

## 4. 提案法

提案は二つあり，それぞれ二段階からなる．いずれの提案も，はじめに，非協力型問題へ逆強化学習を適用し，報酬関数を推定する．次に，得られた報酬関数に基づき Joint-state learners を用いて学習する際，エージェントごとに学習させる順次学習法と，各エージェントが同時に学習する際の非 MDP 性を緩和する最適解吸着法を提案する．

### 4.1 逆強化学習のマルチエージェント系への適用

Abbeell の逆強化学習は，MDP 環境で，最適なあるいは所望な行動軌跡を所与とし，単一エージェントの報酬関数を推定する手法である．しかし非協力型問題では，各エージェントが同時に学習することにより状態遷移が動的に変化し，非 MDP 環境となる．

そこで他エージェントに最適行動をとらせることで MDP 環境とし，エージェントごとに報酬関数を推定する．Algorithm 1 に，非協力型問題に Abbeel の逆強化学習を適用する場合のアルゴリズムを示す．ただし，エキスパートの最適行動軌跡を数値化した特徴期待値は，各エージェントの状態の組み合わせを用いる．また Algorithm 1 のステップ 5 の強化学習において，エージェント同士が衝突すると，報酬関数を推定しているエージェント以外が最適行動をとれなくなるためエピソードを終了する．これに関連して他エージェントと衝突後に適切な学習ができない場合があるため，終了条件に推定した報酬関数に基づき学習させると最適方策を獲得できることを加える．

**Algorithm 1** 非協力型問題における見習い学習

1: 各エージェントに対して繰り返す：
　//各エージェントの最適な行動軌跡を所与とする
　//エージェント $j$ の報酬関数を推定する
2: ランダムに選んだ方策 $\pi^{(0)}$ のもとで $\mu^{(0)} = \mu(\pi^0)$ を計算し，$i = 1$ とする
　($j$ 以外のエージェントは最適行動をとる)
3: 以下の式により $\overline{\mu}^{(i-1)}$，$w^{(i)}$，$t^{(i)}$ を計算する
　- $\overline{\mu}^{(i-1)} = \overline{\mu}^{(i-2)} + \frac{(\mu^{(i-1)} - \overline{\mu}^{(i-2)})^T(\mu_E - \overline{\mu}^{(i-2)})}{(\mu^{(i-1)} - \overline{\mu}^{(i-2)})^T(\mu^{(i-1)} - \overline{\mu}^{(i-2)})}(\mu^{(i-1)} - \overline{\mu}^{(i-2)})$
　- $w^{(i)} = \mu_E - \overline{\mu}^{(i-1)}$
　- $t^{(i)} = ||\mu_E - \overline{\mu}^{(i-1)}||_2$
　(1 回目の繰返しでは，$w^{(0)} = \mu_e - \mu^{(0)}$，$\overline{\mu}^{(0)} = \mu^{(0)}$ とする)
4: $t^{(i)} < \epsilon$ かつ現在の報酬関数で最適方策を獲得できるとき，繰り返しを終了する
5: 報酬関数 $R = (w^{(i)})^T\phi$ のもとでの最適方策 $\pi^{(i)}$ を強化学習により求める
　($j$ 以外のエージェントは最適行動をとる．エージェント同士が衝突した場合，そのエピソードを終了する)
6: 方策 $\pi^{(i)}$ のもとで $\mu^{(i)} = \mu(\pi^{(i)})$ を計算する
　($j$ 以外のエージェントは最適行動をとる)
7: $i = i + 1$ としてステップ 3 へ戻る

### 4.2 提案 1：順次学習によるマルコフ性維持

提案 1 は，他エージェントに最適行動をとらせる環境を，エージェントごとに順次学習する手法である．

非協力型問題では，各エージェントが同時に学習するため環境が動的に変化する．この非マルコフ性が原因で適切に学習できない場合がある．そこでエージェントごとに順次学習することを考える．このとき学習中のエージェント以外に最適行動をとらせることで，環境が MDP となり適切に学習できると考

表 4: 提案法の収束回数 [回]

| 手法 | | 11step | | 12~13step | 発散 |
|---|---|---|---|---|---|
| | | 軌跡一致 | 軌跡不一致 | | |
| 順次学習法 | | 30 | 0 | 0 | 0 |
| 最適解吸着法 | $i$=1 | 71 | 26 | 3 | 0 |
| | $i$=2 | 91 | 9 | 0 | 0 |
| | $i$=3 | 97 | 3 | 0 | 0 |
| | $i$=4 | 100 | 0 | 0 | 0 |

えられる．全エージェント学習後，各エージェントは獲得した Q 値を用いて同時に行動する．

マルチエージェント系では各エージェントが同時に学習することが多い．一方，提案 1 では，各エージェントは同時に学習していないことに注意が必要である．

### 4.3 提案 2：最適方策吸着による非マルコフ性緩和

提案 2 は，学習の最初のエピソード，各エージェントに最適行動をとらせてから，同時に学習する手法である．

非協力型問題では，非マルコフ性の影響で適切に学習できない場合がある．一方で，学習が収束するにつれて方策が固定されていくため MDP 環境に近づく．そこで各エージェントに同時に学習する際の最初の $i$ エピソード，最適行動をとらせてから，各エージェントは推定した報酬関数に基づき同時に学習する．最初の $i$ エピソードにおいて軌跡と一致する行動が強化されるため，非マルコフ性を緩和できる．これにより報酬関数が機能しやすくなり，適切に学習できると考えられる．また，未知の環境を学習できる強化学習と異なり，逆強化学習では最適方策が所与であるため，この緩和法は不自然でないと考えられる．

## 5. 計算機実験

### 5.1 実験 1：順次学習法

■実験設定

提案法の一つ目である順次学習法を用いる．ただし 1 回の実験では，逆強化学習を用いて報酬関数を推定し，得られた報酬関数に基づき各エージェントごとに学習後，獲得した Q 値に基づき行動するとして，30 回実験する．ただし，逆強化学習において所与であるエキスパートの行動軌跡，パラメータは 3.2.3 項と同様とする．

■実験結果と考察

表 4 に順次学習法の結果を示す．実験結果より，推定した報酬に基づき他方に最適行動をとらせてエージェントごとに学習すれば，両エージェント学習後は最適行動をとることが確認できた．また，Abbeel の逆強化学習を用いて推定した報酬関数は妥当であり，得られた報酬関数に基づき各エージェントごとに学習する順次学習法は，非協力型問題への逆強化学習の導入法として有効であるといえる．

### 5.2 実験 2：最適解吸着法

■実験設定

提案法の二つ目である最適解吸着法を用いる．ただし 3.2.3 項と同様，Joint-state learners を用いて同時に学習する際の設定したエピソード数を終えるまでを 1 試行とし，1 回の実験では，逆強化学習を用いて報酬関数を推定し，両エージェントに最初の $i$ エピソード，最適行動をとらせた後，得られた報酬関数に基づき各エージェントは 1 試行同時に学習する．ただし，逆強化学習において所与であるエキスパートの行動軌跡，パラメータは 3.2.3 項と同様とする．

■実験結果と考察

表 4 は各エピソード $i$ について 100 回実験を行った結果である．荒井らの手法では発散する場合があったが，$i = 1$ のときでも発散することがなくなった．また $i$ を増やしていくと，全試行で最適行動を獲得できた．つまり最適解吸着法により，非協力型問題に Abbeel の逆強化学習を適用し，同時に学習する際に生じる問題を抑制できているといえる．

## 6. 結論及び今後の課題

本論文では，タスク達成中に競合が生じる非協力型問題を対象とし，協調行動を獲得させることによって最適解へ収束させる方法を提案した．具体的には，はじめに，Abbeel の逆強化学習をマルチエージェント系へ適用して報酬関数を推定した．次に，得られた報酬関数に用いて，学習させるエージェント以外に最適行動をとらせた環境をエージェントごとに順次学習をさせ，獲得した Q 値を用いて行動させる順次学習法と，学習の最初のエピソードにおいて，各エージェントに最適行動をとらせた後，得られた報酬関数に基づき同時に学習する最適解吸着法を提案した．また，その有効性を計算機実験によって示した．

今後の課題として，エージェント数が増加した際の状態空間の爆発問題や同時学習問題，不完全知覚問題といったマルチエージェント系特有の問題に対して，報酬の設計だけでなく状態空間の設計も考える必要がある．

## 参考文献


[Abbeel 04] Pieter Abbeel, and Andrew Y. Ng: Appresnticeship Learning via Inverse Reinforcement Learning, Proceedings of the 21st International Conference on Machine Learning, ICML'04, pp.7381-7406, 2004.

[Arai 01] 荒井幸代: マルチエージェント強化学習-実用化に向けての課題・理論・諸技術との融合-，人工知能誌，vol.16，no.4，pp.476–481，2001.

[Russell 98] S. Russell: Learning agents for uncertain environment (extended abstract), Proceeding of the 16th International Conference on Machine Learning, pp. 278–287, 1998.

[Ng 00] Andrew Y. Ng, and Stuart J. Russell: Algorithms for Inverse Reinforcement Learning, In Proceedings of the 17th International Conference on Machine Learning, pp. 663–670, 2000.

[Natarajan 10] Natarajan, Sriraam and Kunapuli, Gautam and Judah, Kshitij and Tadepalli, Prasad and Kersting, Kristian and Shavlik, Jude: Multi-agent inverse reinforcement learning, In ICMLA2010, pp.395–400, IEEE, 2010.

[Arai 15] 荒井幸代 and 堀澤優介 and 北里勇樹: マルチエージェント逆強化学習による報酬設計問題の考察, 人工知能学会全国大会論文集, vol.29, pp.1–4, 2015.

[Kyo 14] 許海遅 and 荒井幸代: 行動干渉状態の検出によるマルチエージェント強化学習法の改善, 電気学会論文誌. C, vol.134, no.9, pp.1310–1317, 2014.