1L5-4

# 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

## A Study of Reward Design Problem of Multiagent Domain via Inverse Reinforcement Learning in Multiagent's Environment

荒井幸代　堀澤優介　北里勇樹
Sachiyo Arai　Yusuke Horisawa　Yuki Kitazato

千葉大学大学院工学研究科
Graduate School of Engineering, Chiba University

We focus on the designing reward of multiagent reinforcement learning problem to make autonomous agents act desirable. We take the problem, which is known as Sokoban, warehouse management problem. The previous studies, which take Sokoban in the multiagent reinforcement learning context, don't always achieve the optimal policy for a volume of empirical knowledge to share the reward among the agents. In this paper, we propose the method applying Inverse reinforcement learning (IRL), which is a framework to estimate a reward function from state transition trajectories of the expert. We show potential of this method to achieve the optimal or desired cooperative behavior of the agents' through some empirical results.

## 1. はじめに

複数のタスクを複数の自律主体（以下，エージェントと呼ぶ）間で最適に処理するための行動計画問題を対象とする．具体的には，PSPACE 完全問題のクラスに属する倉庫番問題を取り上げる．PSPACE 完全は問題のサイズに対して多項式オーダのメモリを使用して解ける問題の中で最も難しいクラスの問題に属し，一般的な倉庫番パズルを解く効率的なアルゴリズムは存在しない．そのため，背景知識の利用や近似解法など様々な接近法が試みられている．

本研究ではこの問題に対して，マルチエージェントによるアプローチをとる．既存研究ではマルチエージェント強化学習によって解決するための，各エージェントへの適切な報酬配分法が提案されている．しかし，この報酬配分方法は経験的知識に基づいているため，必ずしも最適な行動系列，あるいは所望の行動系列を生成させることを保証しない点が課題である．

そこで本研究では，最適な（所望の）行動系列を所与として，逆強化学習を用いて報酬関数を推定する方法を推定する．また，各エージェントがこの報酬関数に基づいて自律的に学習した結果，最適なタスク配分が実現できることを計算機実験によって確認し，提案手法の有効性を考察する．

## 2. 対象問題

マルコフ決定過程 (Markov Decision Process:MDP) は，$\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, D, R \rangle$ の組みからなり，$\mathcal{S}$ は状態集合，$\mathcal{A}$ は行動集合，$\mathcal{T} = \{P_{sa}\}$ は状態 $s$ において行動 $a$ を選択した時の状態遷移確率の集合を表し，$D$ は初期状態分布，$R$ は報酬関数である．

このとき，$k$ 個の特徴量を持つ特徴ベクトル $\phi : \mathcal{S} \to [0,1]^k$ を考えると，報酬関数 $R(s)$ は $R(s) = w \cdot \phi(s)$（ただし $w \in \mathbb{R}^k$，$w$ は各特徴量の重み）と表せる．ある方策 $\pi$ のもとでの状態価値は，初期状態を $s_0$ として式 (1) で表せる．ここで $\gamma$ は割引率である．方策 $\pi$ のもとでの特徴ベクトルの期待値を，特徴期待値 $\mu(\pi)$ と定義すると，特徴期待値は式 (2)，状態価値は式 (3) となる．

$$\begin{aligned} E_{s_0 \sim D}\left[V^{\pi}(s_0)\right] &= E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \middle| \pi\right] \\ &= E\left[\sum_{t=0}^{\infty} \gamma^t w \cdot \phi(s_t) \middle| \pi\right] \\ &= w \cdot E\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t) \middle| \pi\right] \end{aligned} \tag{1}$$

$$\mu(\pi) = E\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t) \middle| \pi\right] \in \mathbb{R}^k \tag{2}$$

$$E_{s_0 \sim D}\left[V^{\pi}(s_0)\right] = w \cdot \mu(\pi) \tag{3}$$

### 2.1 報酬配分問題

各エージェントが共通の目標状態を実現するためのマルチエージェント系計画問題に対して，強化学習を適用する場合，全エージェントに共通の報酬 $r$ を与えても，最適な行動が獲得できないことが指摘されている [1]．これは，マルチエージェント系特有の他エージェントの行動に対する知覚の不完全性，同時学習に起因して，全体で共通の報酬だけでは，報酬獲得と無関係な行動が強化されるためである．たとえば，目標状態到達に直接関与したエージェントだけが報酬を得る場合が生じ，その結果，報酬獲得を補助したエージェントの行動が評価されないまま強化が進む．

宮崎ら [2] は Profit Sharing を用いた学習において単位行動あたりの報酬期待値が正となるような政策を獲得するための報酬配分指針を示している．また，保知ら [3] は完全知覚エージェント集団についてベイジアンネットワークにより状態遷移確率を同定し報酬配分を行う機構を集団の外部に設置する手法を提案している．本研究では，文献 [3] のマルチエージェント倉庫番の問題設定にもとづき，報酬関数を推定する．

#### マルチエージェント倉庫番

実験環境は文献 [3] にしたがい，$5 \times 6$ の 2 次元の格子状の環境に 3 人のエージェントと 2 つの荷物及びゴールが存在する．図 1 は実験タスクの初期状態を表しており，a0，a1，a2 はエージェントを，b0，b1 は荷物，G はゴールをそれぞれ表

連絡先: 荒井幸代，千葉大学大学院工学研究科，千葉市稲毛区弥生町 1-33，sachiyo@faculty.chiba-u.jp

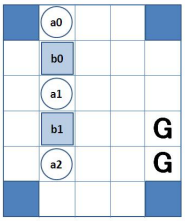



図 1: マルチエージェント倉庫番 (初期状態)

しており，四隅は壁となっている．エージェントは，自身，他のエージェント，荷物，ゴール及び壁の位置を知覚する．行動選択は 4 つあり，隣接する上下左右のマスに移動できる．ただし，各エージェントは他のエージェントがどの行動をとったのかは知覚できない．エージェントは荷物と隣接しているときに荷物を押すことができるが，荷物を引くことはできない．各離散時間ステップにおいて全エージェントが同時に行動を行うものとする．同一マスに複数の物体やエージェントが存在することはできず，エージェントや荷物はその場にとどまる．タスクの目標状態は，すべての荷物がエージェントによってゴール位置 $G$ に移動された状態である．

目標達成 (goal)，デッドロック (deadlock) または 200 ステップを超える (timeover) と 1 エピソードが終了し，初期状態へ戻る．この問題では，荷物が壁の隣に移動したとき deadlock に陥ったと判断する．これは，荷物がゴール位置に到達していないにもかかわらず，それ以上押すことができずにタスク遂行は望めない状態となるからである．

### 2.2 予備実験

強化学習では目標状態に到達した時の報酬 $r_g$ の設定だけで学習できることがその長所でもあるが，マルチエージェント系や不完全知覚が生じている場合，さらに大規模な状態空間では goal 以外の状態で設定する必要がある．

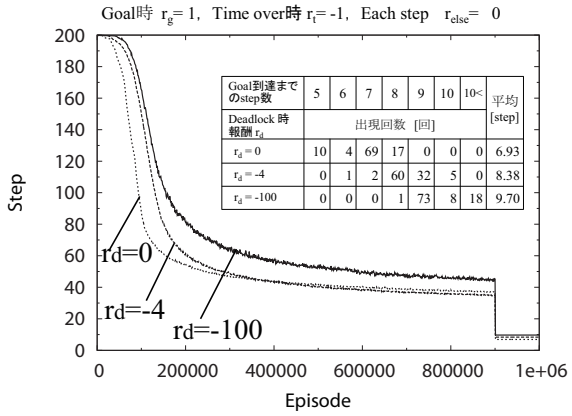
文献 [3] における強化学習では goal 時の報酬 $r_g = 1$ の他に 3 つのタイミングで報酬を設定している．timeover では報酬 $r_t = -1$，deadlock では報酬 $r_d = -4$，その他の場合は報酬 $r_{else} = 0$ である．本節では，文献 [3]の報酬設定，および $r_g = 1$ 以外の deadlock 時，および各ステップでエージェントに与えらえる報酬の影響を観察する予備実験を行った．

各実験は 1 試行 100 万エピソードを繰り返し，各エージェントの学習には Q 学習 [5] を用いた．行動選択は $\epsilon$-greedy 選択とし，$\epsilon$ は 90 万エピソードまでを $\epsilon = 0.3$，それ以降は 0 とした．学習率と割引率はそれぞれ 0.01，0.9 とし，報酬の与え方は，得られた全報酬量を全エージェントに均等配分する．

ここで，deadlock では報酬 $r_d$，その他の場合は報酬 $r_{else}$ の値の影響を観察する実験を行った．図 2 は $r_d$ の値を-100，-4，0 とした場合，図 3 はステップ毎に与える値として $r_{else} = 0$ と $r_{else} = -0.1$ とした場合の実験結果を示している．

それぞれのグラフは 100 試行の実験結果を平均した値をプロットしており，図の横軸はエピソード数，縦軸は 1 エピソードが終了するまでのステップ数を 1,000 エピソードごとに平均した値である．各図中の表は，100 試行中 goal 到達までに要したステップ数とそれぞれの出現回数である．

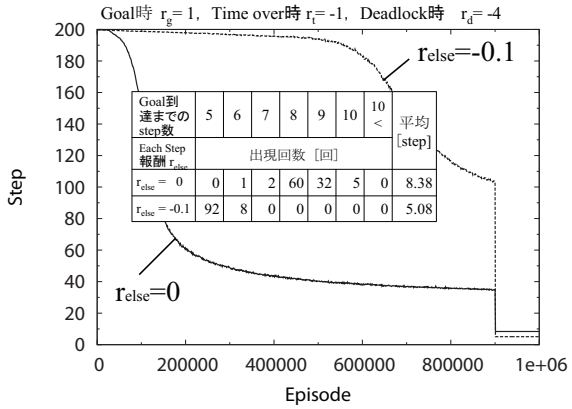
■ $r_d$ **の影響**：図 2 より，目標到達までに要した平均ステップ数は，$r_d = -100$ では 9.70，$r_d = -4$ では 8.38，$r_d = 0$ では 6.93 となっており，この実験の範囲からは，与える負の報酬が大きければよいというわけではなさそうである．これは，deadlock での負の報酬が大きいほど，その状態を避けようとエージェントが動くため，最短手数で目標状態に到達するという視点からは一見無駄な行動を，各エージェントが学んだ結果であると考える．




| Goal到達までのstep数 | 5 | 6 | 7 | 8 | 9 | 10 | 10< | 平均 [step] |
|---|---|---|---|---|---|---|---|---|
| Deadlock 時 報酬 $r_d$ | 出現回数 [回] | | | | | | | |
| $r_d = 0$ | 10 | 4 | 69 | 17 | 0 | 0 | 0 | 6.93 |
| $r_d = -4$ | 0 | 1 | 2 | 60 | 32 | 5 | 0 | 8.38 |
| $r_d = -100$ | 0 | 0 | 0 | 1 | 73 | 8 | 18 | 9.70 |

図 2: デッドロック時の報酬値 $r_d$ が学習性能に与える影響

■ $r_{else}$ **の影響**：図 3 より，$r_{else} = -0.1$ では，100 試行中 92 回の試行で最短手数である 5 ステップに収束していることがわかる．毎ステップ負の報酬を与えることによって学習の高速化が図れることは既存の研究でも示されている．しかし，$r_{else} = 0$ の場合でもマルコフ決定過程であれば最終的には最短ステップに収束してもよいはずである．つまり，これらの実験結果は，各エージェントの報酬を同じ値を，アドホックに設定しても最短手数である 5 ステップで目標状態に到達することはできないことを示している．




| Goal到達までのstep数 | 5 | 6 | 7 | 8 | 9 | 10 | 10< | 平均 [step] |
|---|---|---|---|---|---|---|---|---|
| Each Step 報酬 $r_{else}$ | 出現回数 [回] | | | | | | | |
| $r_{else} = 0$ | 0 | 1 | 2 | 60 | 32 | 5 | 0 | 8.38 |
| $r_{else} = -0.1$ | 92 | 8 | 0 | 0 | 0 | 0 | 0 | 5.08 |

図 3: 毎ステップの報酬値 $r_{else}$ が学習性能に与える影響

## 3. 提案手法

マルチエージェント強化学習において goal 時（所望の状態）だけに報酬を設定しても最適（所望な）方策が得られるとは限らない．マルコフ決定過程を仮定できない以上，得られる方策は初期乱数や行動選択法に依存し，報酬によって制御することは難しいことを予備実験の結果が示している．

そこで本研究では，「所望の行動」が予めわかっている場合に，その行動を自律的に獲得するための報酬設計方法を提案す

る．所望の行動がわかっていればそれらを，各エージェントに組み込めばよいのではないかという考え方もあるが，一般に「所望の行動」が既知であるのは状態空間の一部である．それ以外の状態に陥った場合の行動は与えられない．これらを学習させるためには，一旦報酬を設定し，状態空間全体の行動を学習させる必要が生じる．

本研究では，文献 [4] に示されている逆強化学習をマルチエージェント環境に適用し，得られた報酬関数に基づいて学習した結果，最適なタスク配分が実現できることを示す．

### 3.1 逆強化学習

最適な（所望の）行動系列を知るエージェントをエキスパートと呼ぶ．逆強化学習 (Inverse Reinforcement Learning:IRL)[4] は，真の報酬関数 $R^*(s) = w^* \cdot \phi(s)$ が自明でない環境において，エキスパートのパフォーマンスに近い方策を得ることが目的である．エキスパートの方策を $\pi_E$ とすると，$s_0 \sim D$ からスタートし，$\pi_E$ に従って行動することによって得られたエキスパートの行動系列を観測することによって，真の報酬関数を推定することができる．特に，$m$ 個の行動系列 $\{s_0^i, s_1^i, \ldots\}_{i=1}^m$ が与えられた時の，エキスパートの特徴期待値 $\mu_E = \mu(\pi_E)$ は，式 (4) により求められる．

$$\mu_E = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{\infty}\gamma^t\phi(s_t^i) \tag{4}$$

エキスパートのパフォーマンスに近い方策を得るためには，$||\mu(\tilde{\pi}) - \mu_E||_2 \leq \epsilon$ となる方策 $\tilde{\pi}$ を求める必要がある．$w^* \in \mathbb{R}^k$ であり，かつ $||w^*||_1 \leq 1$ である方策 $\tilde{\pi}$ のもとでは，

$$\left| E\left[\sum_{t=0}^{\infty}\gamma^t R(s_t)\middle|\pi_E\right] - E\left[\sum_{t=0}^{\infty}\gamma^t R(s_t)\middle|\tilde{\pi}\right]\right| \tag{5}$$

$$= |w \cdot \mu(\tilde{\pi}) - w \cdot \mu_E| \tag{6}$$

$$\leq ||w||_2||\mu(\tilde{\pi}) - \mu_E||_2 \tag{7}$$

$$\leq 1 \cdot \epsilon = \epsilon \tag{8}$$

となり，問題は $||\mu(\tilde{\pi}) - \mu_E||_2 \leq \epsilon$ を満たす方策 $\tilde{\pi}$ を見つけることに帰着する．

### 3.2 Projection Method

方策 $\tilde{\pi}$ を得るアルゴリズムの 1 つに，文献 [4] の Projection Method(以後 PM と表記) がある．式 (9) に PM の式を，図 4 に PM を用いた IRL のアルゴリズムを示す．ただし，$\bar{\mu}^{(0)} = \mu^{(0)}$ とする．

$$\bar{\mu}^{(i-1)} = \bar{\mu}^{(i-2)} + \frac{\left(\mu^{(i-1)} - \bar{\mu}^{(i-2)}\right)^T\left(\mu_E - \bar{\mu}^{(i-2)}\right)}{\left(\mu^{(i-1)} - \bar{\mu}^{(i-2)}\right)^T\left(\mu^{(i-1)} - \bar{\mu}^{(i-2)}\right)} \cdot \left(\mu^{(i-1)} - \bar{\mu}^{(i-2)}\right) \tag{9}$$

## 4. 実験

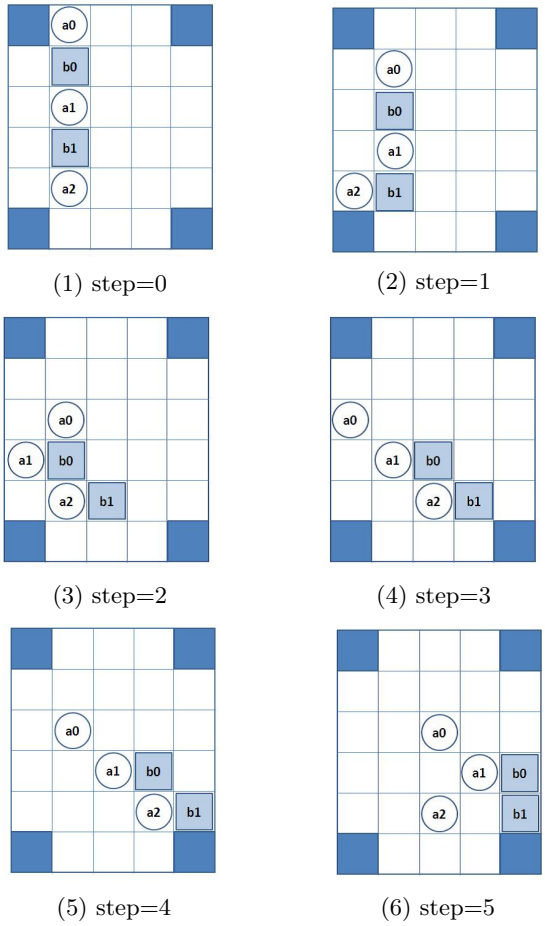
逆強化学習におけるエキスパートの行動を図 5 に示す．これは，最短手数で目標状態に到達する方策のうち，timeover となる回数や deadlock に陥る回数が少ないものを採用した．

### 4.1 実験設定

各エージェントの学習には Q 学習を用い，行動選択は $\epsilon$-greedy 選択とした．また，図 4 に示した方策 $\pi^{(0)}$ は，常にランダムな行動をとる方策を採用し，終了条件は $t^{(i)} = ||\mu_E - \bar{\mu}^{(i-1)}||_2$ が 0.000001 以下とした．

1. 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．
2. 式 (9) より $\bar{\mu}^{(i-1)}$ を求め，$w^{(i)} = \mu_E - \bar{\mu}^{(i-1)}$，$t^{(i)} = ||\mu_E - \bar{\mu}^{(i-1)}||_2$ を計算し，$t^{(i)} \leq \epsilon$ ならばアルゴリズムを終了する．
3. 強化学習を用いて，報酬関数 $R = w^{(i)} \cdot \phi$ のもとでの最適な方策 $\pi^{(i)}$ を求め，$\mu^{(i)} = \mu(\pi^{(i)})$ を計算する．
4. $i = i + 1$ としてステップ 2 に戻る．

図 4: PM を用いた IRL アルゴリズム




(1) step=0　(2) step=1

(3) step=2　(4) step=3

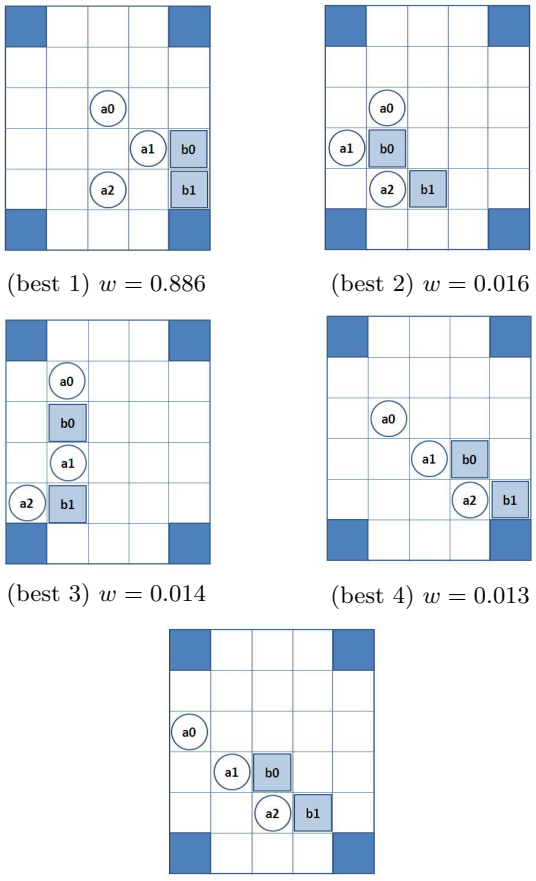
(5) step=4　(6) step=5

図 5: エキスパートの行動系列

### 4.2 実験結果

図 6 に逆強化学習により得られた報酬関数のうち，各特徴量の重み $w$ が 0 より大きくなった上位 5 状態を示す．各状態ともエキスパートと同じ状態の報酬が大きくなっており，特に目標に到達した状態の報酬が最も大きい．この 5 状態以外のいずれの状態も報酬値は 0 か負の値であった．

次に，逆強化学習により求められた報酬関数を用いて強化学習を行った．実験結果を図 7 に示す．また表 1 に，それぞれの収束後の目標到達までのステップ数，1 試行 100 万エピソード中に goal に到達した回数，timeover となった回数，deadlock となった回数を 100 試行平均した値を示す．RL($r_{else} = 0$) では最適方策は得られなかったが，RL($r_{else} = -0.1$)，IRL では最適方策を得ることができた．また，RL($r_{else} = -0.1$) では deadlock に陥る回数が多く，学習途中のステップ数が他の 2 つのグラフに比べ多くなっている．さらに，収束後のステップ数は 100 試行中 8 試行は最短手数になっていない．一方 IRL は，全試行でエキスパートと同じ行動を学習し，5 ステップで目標状態に到達することができ，timeover となった回数や deadlock に陥った回数も RL($r_{else} = -0.1$) より少なかった．




(best 1) $w = 0.886$　(best 2) $w = 0.016$

(best 3) $w = 0.014$　(best 4) $w = 0.013$

(best 5) $w = 0.012$

図 6: 逆強化学習により得られた報酬関数のうち，$w$ の値が大きかった上位 5 状態

## 5. 結論および今後の課題

本研究ではマルチエージェント強化学習の報酬設計問題に対して，最適な行動系列を所与とし，逆強化学習を用いて報酬関数を推定することによって，各エージェントが最適なタスク配分を自律的に学習できることを示した．提案手法は，エージェント毎に適切と思われる報酬関数を獲得することができるため，アドホックな報酬配分が不要であることがシステム設計上の利点といえる．

今後の課題として，逆強化学習アルゴリズムでは複数の解候補が得られるため，効率よく学習可能な報酬関数の選定法，さらにエキスパートの行動が不完全な場合の補修法の検討を挙げる．

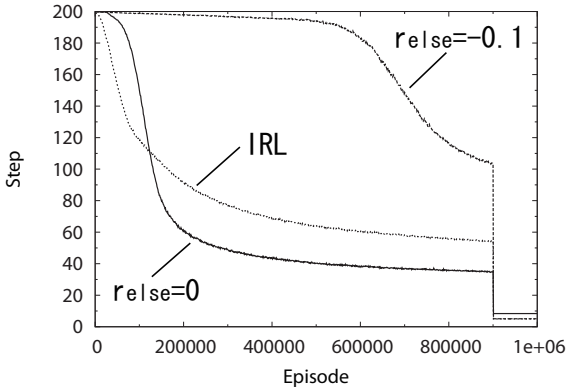



図 7: RL( $r_{else} = 0, -0.1$) と IRL の目標到達までのステップ数比較

表 1: 収束後のステップ数と 100 万エピソード中に goal, timeover，deadlock となった回数 (100 試行平均)

| Reward | step | goal | timeover | deadlock |
|---|---|---|---|---|
| $r_{else} = 0$ | 8.38 | 788,214 | 24,730 | 187,056 |
| $r_{else} = -0.1$ | 5.08 | 228,796 | 21,750 | 749,453 |
| IRL | 5.00 | 686,664 | 1,586 | 311,750 |

## 参考文献

[1] 荒井幸代，宮崎和光，小林重信，“マルチエージェント強化学習の方法論–Q-learning と Profit Sharing による接近–”，人工知能学会誌，Vol. 13，No. 5，pp.609-618 (1998).

[2] 宮崎和光，荒井幸代，小林重信，“Profit Sharing を用いたマルチエージェント強化学習における報酬配分の理論的考察”，人工知能学会誌，Vol. 14，No. 6，pp.1156-1164 (1999).

[3] 保知良暢，新谷虎松，伊藤孝行，大囿 忠親，“外部評価機構を導入したマルチエージェント強化学習における過去の事象に基づく報酬配分”，電子情報通信学会論文誌，vol.J87-D-1，no.12，pp.1119–1127 (2004.12).

[4] P. Abbeel，A. Ng，“Apprenticeship Learning via Inverse Reinforcement Learning”，ICML 21 (2004).

[5] Richard S. Sutton，Andrew G. Barto，“Reinforcement Learning: An Introduction”，*A Bradford Book*，The MIT Press (1998).