複数環境におけるエキスパート軌跡を用いた ベイジアン逆強化学習

転移可能な報酬の推定に向けたアプローチ

Bayesian Inverse Reinforcement Learning for Demonstrations of an Expert in Multiple Dynamics

Toward Estimation of Transferable Reward

中田 勇介
Nakata Yusuke

千葉大学大学院 融合理工学府 都市環境システムコース
Department of Urban Environment Systems, Graduate School of Science and Engineering, Chiba University
al019nakata@gmail.com

荒井 幸代
Arai Sachiyo

千葉大学大学院 工学研究院 都市環境システムコース
Department of Urban Environment Systems, Graduate School of Engineering, Chiba University
sachiyo@faculty.chiba-u.jp

keywords: inverse reinforcement learning, reinforcement learning, Bayesian inference, Markov decision processes

Summary

Though a reinforcement learning framework has numerous achievements, it requires a careful shaping of a reward function that represents the objective of a task. There is a class of task in which an expert could demonstrate the optimal way of doing, but it is difficult to design a proper reward function. For these tasks, an inverse reinforcement learning approach seems useful because it makes it possible to estimate a reward function from expert's demonstrations. Most existing inverse reinforcement learning algorithms assume that an expert gives demonstrations in a unique environment. However, an expert also could provide demonstrations of tasks within other environments of which have a specific objective function. For example, though it is hard to represent objective explicitly for a driving task, the driver could give demonstrations under multiple situations. In such cases, it is natural to utilize these demonstrations in multiple environments to estimate expert's reward functions. We formulate this problem as *Bayesian Inverse Reinforcement Learning problem* and propose a Markov Chain Monte Carlo method for the problem. Experimental results show that the proposed method quantitatively overperforms existing methods.

1. はじめに

ある環境におけるタスクに対して最適な方策を獲得する方法の一つに強化学習がある [Sutton 18]. 強化学習は一般に、タスクの目的を反映した報酬を与えることによって、環境のモデルに対する事前知識や、逐次的な教師信号なしに、タスクの解法を学習できる方法として注目されてきた [Mnih 15, Levine 16, Silver 17]. しかし、実問題においては、タスクの目的を適切に反映した報酬の設計が困難なことが多い [Amodi 16, Everitt 16]. この報酬設計を回避する方法に逆強化学習 [Ng 00, Abbeel 04, Ramachandran 07, Ziebart 08] がある.

逆強化学習は、あるタスクの解法を知るエージェント (以下、エキスパート) が生成した軌跡などの行動履歴データから、「エキスパートの方策が最適となる報酬」を推定する方法である. そのため、逆強化学習で推定した報酬 (以下、推定報酬) に対する最適方策を求めること

によってエキスパートの方策を学習できる. また、エキスパートの報酬が正しく推定できれば、これを他の環境に転移することによって、転移先の環境におけるエキスパートの方策を学習することも可能になる [Fu 18]. すなわち、推定報酬がエキスパートの報酬と等しいならば、その推定報酬は、任意の状態遷移確率の環境に対して適用可能である. これを本研究では「報酬の転移」と呼ぶことにする. 報酬の転移が可能になれば、転移先の環境におけるエキスパートの軌跡を必要とせずに、転移先の環境におけるエキスパートの方策が学習できる.

報酬の転移が有用な問題としては、特定のゲームにおける複数のステージ間での転移 [Nichol 18], シミュレータから実環境への転移 [Pinto 17] が挙げられる. また、実データを用いた例に、特定の駐車場における歩行者の軌跡から推定した報酬を用いて、異なる駐車場や歩道における歩行者の方策を学習した研究がある [Kitani 12].

報酬が転移可能である条件は、推定報酬がエキスパー

トの報酬と定数倍の違いを除いて一致していることである [Ng 00]. しかし、従来の逆強化学習では、ある環境において生成したデータが十分に存在しても、転移先でのエキスパートの方策を学習できる報酬が推定される保証はない。これは、「エキスパートの方策が最適となる報酬」を推定する問題の解が複数存在し、逆強化学習の推定報酬がエキスパートの報酬と必ずしも一致しないからである。そこで、本研究では、報酬の転移が要請される場合には、しばしば状態遷移確率の異なる複数の環境（例：ゲームの複数のステージ）におけるエキスパートの軌跡が得られることに着目する。そして、エキスパートが複数の環境で生成した軌跡から報酬を推定する方法を提案する。

提案法は、代表的な逆強化学習の一つである Bayesian Inverse Reinforcement Learning (BIRL) [Ramachandran 07] に基づく。複数の環境のエキスパートの軌跡から報酬の事後分布を推定する逆強化学習問題を定式化し、報酬の事後分布からのサンプリングを実現するマルコフ連鎖モンテカルロアルゴリズムを提案する。そして、提案したアルゴリズムが報酬の事後分布の実現に要するステップ数が、環境の数に依存しないことを示す。

実験では、二つの実験設定で BIRL と提案法を比較する。逆強化学習手法の比較に一般的に用いられる指標である Expected Value Difference (EVD) [Levine 11] を用いて比較した結果、どちらの実験設定においても、提案法の推定報酬が既存手法の推定報酬と比較してエキスパートの報酬に近いことが確認された。また、ハイパーパラメータの変化に対しても頑健であることが実験的に示された。

以下、2 章では、提案法の説明に必要なマルコフ決定過程と BIRL [Ramachandran 07] について述べる。3 章では対象問題と関連研究について述べ、4 章では提案法を説明する。5 章で計算機実験の設定と結果を述べ、6 章で実験結果を考察する。そして、7 章でまとめと今後の課題について述べる。

2. 準備

2.1 マルコフ決定過程 (MDP)

本研究の対象とするマルコフ決定過程 [Barto 89] $\mathcal{M} = (E, R)$ は、環境 $E = \langle \mathcal{S}, \mathcal{A}, T, \gamma \rangle$ と報酬 R からなる。ここで、 $\mathcal{S}$ は状態集合、 $\mathcal{A}$ は行動集合、 $T(s'|s, a)$ は状態 $s \in \mathcal{S}$ で行動 $a \in \mathcal{A}$ を取ったとき次状態 $s' \in \mathcal{S}$ に遷移する状態遷移確率、 $\gamma \in [0, 1]$ は割引率、 $R: \mathcal{S} \rightarrow \mathbb{R}$ は状態 $s \in \mathcal{S}$ の報酬 r を返す報酬関数を表す。エージェントは、状態 s で行動 a を取る確率である方策 $\pi: \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ に基づいて行動を生成する。報酬 R 、方策 π の下での状態 s の状態価値、行動価値をそれぞれ式 (1), (2) に示す。

$$V^\pi(s, R) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t) | S_0 = s \right] \quad (1)$$

$$Q^\pi(s, a, R) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t) | S_0 = s, A_0 = a \right] \quad (2)$$

報酬 R 、最適方策 π^* の下での状態価値 V^* は式 (3) を満たす。

$$V^*(s, R) = R(s) + \gamma \max_a \sum_{s'} T(s'|s, a) V^*(s', R) \quad (3)$$

そして、最適方策の行動価値 Q^* は、最適方策の下での状態価値 V^* を用いて式 (4) と表される。

$$Q^*(s, a, R) = R(s) + \gamma \sum_{s'} T(s'|s, a) V^*(s', R) \quad (4)$$

報酬 R が $P(R)$ からサンプリングされる環境における方策 π の損失は、式 (5) で定義される [Ramachandran 07]。

$$L_{policy}^p(P(R), \pi) = \mathbb{E}_{P(R)} \left[\|V^*(R) - V^\pi(R)\|_p \right] \quad (5)$$

ここで、 $V^\pi(R)$ は報酬 R 、方策 π の下での各状態の状態価値 $V^*(s, R)$ をベクトルとして並べ上げたもので、 p は任意のノルムを指す。式 (5) を最小化する方策 π は、マルコフ決定過程 $\mathcal{M} = (E, \mathbb{E}[R])$ に対する最適方策である。

2.2 Bayesian Inverse Reinforcement Learning (BIRL)

BIRL は、ある報酬に従うエキスパートが生成した状態行動対 (s, a) のデータセット $D = \{(s_i, a_i)\}_{i=1}^N$ と、報酬に対する事前知識を反映した報酬の事前分布 $P(R)$ を所与として、報酬の事後分布 $P(R|D)$ を推定する枠組みである [Ramachandran 07]。報酬の事後分布は、ベイズの定理を用いて、

$$P(R|D) = \frac{P(D|R)}{P(D)} P(R), \quad (6)$$

と表される。尤度 $P(D|R)$ は、報酬 R に対する最適方策の行動価値である $Q^*(s, a, R)$ 値を用いてモデル化される。尤度 $P(D|R)$ の定式化を式 (7) に示す。

$$P(D|R) = \frac{1}{Z} \exp \left(\frac{1}{\kappa} \sum_{(s,a) \in D} Q^*(s, a, R) \right) \quad (7)$$

ここで、 $\kappa > 0$ はエキスパートが最適行動を取る確率を調節するパラメータで、ボルツマン方策 [Sutton 90] における温度パラメータに相当する。報酬の事後分布 $P(R|D)$ は、式 (6) と式 (7) を用いて式 (8) と表すことができる。

$$P(R|D) = \frac{1}{Z'} \exp \left(\frac{1}{\kappa} \sum_{(s,a) \in D} Q^*(s, a, R) \right) P(R) \quad (8)$$

式 (8) の分配関数 Z' は計算困難だが、指数は報酬 R に対する最適方策の行動価値 Q^* を求めることによって計算できる。また、

$$P(R|D) \propto \exp \left(\frac{1}{\kappa} \sum_{(s,a) \in D} Q^*(s, a, R) \right) P(R), \quad (9)$$

図 1 Graphical models for IRL Problems

が成り立つため、マルコフ連鎖モンテカルロ法 (MCMC) を用いて事後分布 $P(R|D)$ から報酬をサンプルできる。BIRL では PolicyWalk と呼ばれる MCMC アルゴリズムを用いて事後分布 $P(R|D)$ から報酬 R をサンプルする。PolicyWalk とは、代表的な MCMC アルゴリズムである GridWalk [Vempala 05] を、不要な強化学習を省くことによって効率化した MCMC アルゴリズムである。

3. 対象問題

本章では、対象問題の定式化を示したのちに関連研究を整理する。

3.1 Bayesian Inverse Reinforcement Learning for Multiple Dynamics (BIRL-MD)

本論文では、ある報酬に従うエキスパートが、複数の環境で生成した軌跡からエキスパートの報酬の分布を推定する問題を定式化する。BIRL と対象問題のグラフィカルモデルの比較を図 1 に示す。

図 1 の影がついたノードは観測可能なデータを、矢印はデータ間の依存関係を指す。BIRL は、エキスパートがある環境 E において生成したデータセット D を所与として、報酬の分布 $P(R|D) = P(R|D, E)$ を推定する問題を定式化した。一方、本論文が定式化する BIRL for Multiple Dynamics (BIRL-MD) は、状態遷移確率が異なる環境 $E_m = \langle \mathcal{S}, \mathcal{A}, T_m, \gamma \rangle$ と、エキスパートが各環境 E_m で生成したデータセット D_m の組からなる集合 $\{(E_m, D_m)\}_{m=1}^M$ を所与として、報酬の事後分布 $P(R|\{(D_m, E_m)\}_{m=1}^M)$ を推定する定式化である。また、図 1 の矢印から分かるように、報酬 R は環境 E と独立、つまりエキスパートの報酬は環境の状態遷移確率に依存しないものとする。

複数の環境におけるエキスパートのデータセット $\{(E_m, D_m)\}_{m=1}^M$ の下での報酬の事後分布を式 (10) に示す。

$$P(R|\{(D_m, E_m)\}_{m=1}^M) = \frac{\prod_{m=1}^M P(D_m|R, E_m)}{\prod_{m=1}^M P(D_m|E_m)} P(R), \quad (10)$$

式 (7) に示した BIRL のエキスパートのモデルを、環境 E 、報酬 R の下での最適方策の行動価値 $Q^*(s, a, R, E)$

へと拡張すれば、報酬の事後分布は

$$P(R|\{(D_m, E_m)\}_{m=1}^M) = \frac{1}{Z} \exp \left(\frac{1}{\kappa} \sum_{m=1}^M \sum_{(s,a) \in D_m} Q^*(s, a, R, E_m) \right) P(R), \quad (11)$$

と表される。

3.2 関連研究

関連研究と本研究が提案する定式化の比較を表 1 に示す。BIRL は、準最適な方策や確率の方策から生成された軌跡と、エキスパートの報酬の事前分布を扱うことができる。しかし、複数の環境で得られた軌跡を扱うことができない。Repeated IRL は、複数の環境で得られた軌跡を扱うことができるが、準最適な方策や確率の方策から生成された軌跡や報酬の事前分布を扱うことができない。Meta IRL [Li 17] は複数の環境において確率の方策から生成された軌跡を扱うことができるが、報酬の事前分布を扱う枠組みにはなっていない。提案法はベイズのアプローチであるため、Maximum Entropy IRL [Ziebart 10] に基づく Meta IRL [Li 17] と比較して多くの計算量を要するが、複数の環境において確率の方策から生成された軌跡と報酬の事前分布を扱うことができる。提案法のように複数の環境を扱う場合には、環境の組合せが重要であることが知られている [Amin 17]。この問題は教師あり学習においても議論されている [Zamir 18]。

提案法を含む上記の逆強化学習手法は、エキスパートの方策や軌跡から報酬を推定し、推定報酬を転移することによって、転移先の環境におけるエキスパート方策を学習する。これに対して、推定報酬ではなく方策を転移するアプローチも考えられる。具体的には、Behavioral Cloning (BC) [Pomerleau 89] や Dagger [Ross 11] を用いて方策を学習し、Meta-Learning [Thrun 98, Finn 17, Duan 17, Clavera 19] や Domain Adaptation [Daume III 06, Jiang 08, Pan 17], Fine-tuning [Lamblin 10, Bengio 12, Yosinski 14] などの手法を用いて方策を転移するアプローチである。しかし、これらのアプローチは転移先の環境におけるエキスパートの軌跡を要するという欠点がある。一方、逆強化学習は推定報酬を転移する際、転移先の環境におけるエキスパートの軌跡は不要である。

文献 [Yan 17] では、Dagger と Domain Randomization [Tobin 17, Peng 18, Andrychowicz 18] を組み合わせることによって、転移先のエキスパートの軌跡を用いることなく転移可能な方策の学習法を提案している。ただし、Domain Randomization は学習に用いる環境と転移先の環境の状態遷移確率が類似しているという仮定が必要となる。これに対して、「報酬の転移」では環境の状態遷移確率の類似性を仮定する必要はない。エキスパートの報

表 1 Related Works

Methods	BIRL [Ramachandran 07]	Repeated IRL [Amin 17]	Meta IRL [Li 17]	BIRL-MD (Ours)
Multiple Dynamics	-	✓	✓	✓
Sub-optimal Expert	✓	-	✓	✓
Incorporates prior distribution on rewards	✓	-	-	✓

Algorithm 1 PolicyWalk for Multiple Dynamics**INPUT:** Environments $\{E_m\}_{m=1}^M$, Demonstrations $\{D_m\}_{m=1}^M$, Prior $P(R)$, Step Size δ **OUTPUT:** Sampled Rewards $\{R_i\}_{i=1}^N$

```

1: Pick a random vector  $R_0 \in \mathbb{R}^{|S|}/\delta$ 
2:  $\{\pi_m\}_{m=1}^M \leftarrow \{\text{PolicyIteration}(E_m, R_0)\}_{m=1}^M$ 
3: for  $i = 1$  do  $N$ 
4:   Pick a reward vector  $\tilde{R}$  uniformly at random from the neighbours of  $R_{i-1} \in \mathbb{R}^{|S|}/\delta$ 
5:   Compute  $Q^\pi(s, a, \tilde{R}, E) \quad \forall \{s, a, (E_m, \pi_m)\} \in \mathcal{S} \times \mathcal{A} \times \{(E_m, \pi_m)\}_{m=1}^M$ 
6:   if  $\exists \{s, a, (E, \pi)\} \in \mathcal{S} \times \mathcal{A} \times \{(E_m, \pi_m)\}_{m=1}^M, Q^\pi(s, \pi(s), \tilde{R}, E) < Q^\pi(s, a, \tilde{R}, E)$  then ▷ If any policy is not optimal
7:      $\{\tilde{\pi}_m\}_{m=1}^M \leftarrow \{\text{PolicyIteration}(E_m, \tilde{R})\}_{m=1}^M$ 
8:      $R_i \leftarrow \tilde{R}$  and  $\{\pi_m\}_{m=1}^M \leftarrow \{\tilde{\pi}_m\}_{m=1}^M$  with probability  $\min \left\{ 1, \frac{P(\tilde{R}, \{(D_m, E_m)\}_{m=1}^M)}{P(R_{i-1}, \{(D_m, E_m)\}_{m=1}^M)} \right\}$ 
9:   else
10:     $R_i \leftarrow \tilde{R}$  with probability  $\min \left\{ 1, \frac{P(\tilde{R}, \{(D_m, E_m)\}_{m=1}^M)}{P(R_{i-1}, \{(D_m, E_m)\}_{m=1}^M)} \right\}$ 
11:   end if
12: end for

```

酬が推定できれば、任意の状態遷移確率の環境におけるエキスパートの方策を学習できるからである。

4. ア プ ロ ー チ

提案法は、環境 $E_m = \langle \mathcal{S}, \mathcal{A}, T_m, \gamma \rangle$ と、エキスパートが各環境において生成したデータセット D_m の組からなる集合 $\{(E_m, D_m)\}_{m=1}^M$ を所与として、報酬の事後分布 $P(R|\{(D_m, E_m)\}_{m=1}^M)$ からの報酬 R のサンプリングを実現する。式 (11) の指数は、各環境 E における報酬 R に対する最適方策の行動価値 $Q^*(s, a, R, E)$ を計算することで求める。そのため、BIRL-MD においても BIRL と同様に、MCMC を用いて事後分布 $P(R|\{(D_m, E_m)\}_{m=1}^M)$ から報酬をサンプルできる。提案法の MCMC アルゴリズム、PolicyWalk for Multiple Dynamics (PolicyWalk-MD) を Algorithm 1 に示す。ここで、Algorithm 1 の 7 行目の PolicyIteration は方策反復と呼ばれる動的計画法の一種である [Sutton 18]。

BIRL の MCMC アルゴリズムである PolicyWalk は、 $P(R|D, E)$ から誤差 ϵ 以下のサンプリングを $O(|S|^2 \log \frac{1}{\epsilon})$ のステップ数で実現する [Ramachandran 07]。PolicyWalk-MD のサンプリングの計算量に関して次の定理が成り立ち、提案法は PolicyWalk と同じオーダーで事後分布からのサンプリングを実現する。

補題 1. [Ramachandran 07]

F を、定義域が $\{x \in \mathbb{R}^n | -d \leq x_i \leq d\}$ の正の実数関数とする。ここで、 d は任意の正の実数である。また、任意の $\lambda \in [0, 1]$ の下で、ある α, β が存在し、以下の二つの式

$$|f(x) - f(y)| \leq \alpha \|x - y\|_\infty, \quad (12)$$

$$f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y) - \beta, \quad (13)$$

が成り立つと仮定する。ここで、 $f(\cdot) = \log F(\cdot)$ である。この時、関数 F に対する GridWalk (PolicyWalk, PolicyWalk-MD も同様に) は誤差 ϵ 以下のサンプリングを、 $O(n^2 d^2 \alpha^2 e^{2\beta} \log \frac{1}{\epsilon})$ ステップで実現する [Apple-gate 91]。

定理 1. 環境 $E = \langle \mathcal{S}, \mathcal{A}, T, \gamma \rangle$ 、事前分布 $P(R)$ が一様分布 $\mathcal{U}(-R_{\max}, R_{\max})$ で、事後分布 $P(R|\{(D_m, E_m)\}_{m=1}^M)$ が式 (11) に示す式で表されるとする。 $R_{\max} = O(\frac{1}{\sum_{m=1}^M |D_m|})$ と仮定すると、PolicyWalk-MD は $P(R|\{(D_m, E_m)\}_{m=1}^M)$ からの誤差 ϵ 以下のサンプリングを、 $O(|S|^2 \log \frac{1}{\epsilon})$ のステップ数で実現する。

Proof.

$$f(R) = \frac{1}{\kappa} \sum_{m=1}^M \sum_{(s,a) \in D_m} Q^*(s, a, R, E_m) \quad (14)$$

$$f_\pi(R) = \frac{1}{\kappa} \sum_{m=1}^M \sum_{(s,a) \in D_m} Q^\pi(s, a, R, E_m) \quad (15)$$

ここで, f_π はベクトル化した R について線形で, $f(R) \geq f_\pi(R)$ が成り立つ. また, Q 値について,

$$\max_{s,a} Q^*(s, a) \leq \sum_{t=0}^{\infty} \gamma^t R_{\max} = \frac{R_{\max}}{1-\gamma} \quad (16)$$

$$\min_{s,a} Q^*(s, a) \geq -\frac{R_{\max}}{1-\gamma} \quad (17)$$

が成り立つ. Q 値の下界を与える不等式 (17) から, $f_\pi(R)$ の下界

$$f_\pi(R) \geq -\sum_{m=1}^M |D_m| \frac{R_{\max}}{\kappa(1-\gamma)}, \quad (18)$$

が得られる. また, 式 (16) から,

$$\frac{\alpha \sum_{m=1}^M |D_m| R_{\max}}{1-\gamma} \geq f(R), \quad (19)$$

$$0 \geq f(R) - \sum_{m=1}^M |D_m| \frac{R_{\max}}{\kappa(1-\gamma)}, \quad (20)$$

と言える. 式 (18) と (20) を足すことによって $f(R)$ と $f_\pi(R)$ の関係式,

$$f_\pi(R) \geq f(R) - 2 \sum_{m=1}^M |D_m| \frac{R_{\max}}{\kappa(1-\gamma)}. \quad (21)$$

が得られる. 式 (21) を用いれば

$$f(\lambda R_1 + (1-\lambda)R_2) \geq f_\pi(\lambda R_1 + (1-\lambda)R_2) \quad (22)$$

$$= \lambda f_\pi(R_1) + (1-\lambda)f_\pi(R_2) \quad (23)$$

$$\geq \lambda f(R_1) + (1-\lambda)f(R_2) \quad (24)$$

$$- \frac{2 \sum_{m=1}^M |D_m| R_{\max}}{\kappa(1-\gamma)}$$

となる. ここで, f_π はベクトル化した R について線形であることを用いた. 補題の変数 α, β はそれぞれ,

$$\begin{aligned} \alpha &= \frac{|f(R_1) - f(R_2)|}{\|R_1 - R_2\|_\infty} \\ &\leq \frac{2 \sum_{m=1}^M |D_m| R_{\max}}{\kappa(1-\gamma) O\left(\frac{1}{\sum_{m=1}^M |D_m|}\right)} = O\left(\sum_{m=1}^M |D_m|\right) \end{aligned} \quad (25)$$

$$\begin{aligned} \beta &= \frac{2 \frac{1}{\kappa} \sum_{m=1}^M |D_m| R_{\max}}{1-\gamma} \\ &= 2 \sum_{m=1}^M |D_m| \frac{O\left(\frac{1}{\sum_{m=1}^M |D_m|}\right)}{\kappa(1-\gamma)} = O(1) \end{aligned} \quad (26)$$

である. よって PolicyWalk-MD は $P(R|\{(D_m, E_m)\}_{m=1}^M)$ からの誤差 ϵ 以下のサンプリングを,

$$\begin{aligned} O\left(|S|^2 \frac{1}{\left(\sum_{m=1}^M |D_m|\right)^2} \left(\sum_{m=1}^M |D_m|\right)^2 \log \frac{1}{\epsilon}\right) \\ = O\left(|S|^2 \log \frac{1}{\epsilon}\right) \end{aligned} \quad (27)$$

のステップ数で実現する. $\square$

報酬のスケールに関する仮定 $R_{\max} = O\left(\frac{1}{\sum_{m=1}^M |D_m|}\right)$ を設けることによって推定報酬の有用性が低下することはない. なぜなら, 推定報酬を定数倍しても, その最適方策は変わらないからである. PolicyWalk-MD が要するステップ数のオーダーは既存手法と等しいが, 各ステップで要する方策反復の数は環境の数に依存する. 提案法の計算量は環境の数に対してたかだか線形である.

2.1 節では, 報酬を状態に依存する関数 $R: \mathcal{S} \rightarrow \mathbb{R}$ として定義したが, 提案法は状態と行動に依存する報酬関数 $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ や, 報酬関数が特徴量 $\phi: \mathcal{S} \rightarrow \mathbb{R}^n$ の線形和 $R(s) = w \cdot \phi(s)$ で表される場合にも適用できる. 状態と行動に依存する報酬関数 $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ を用いた場合, PolicyWalk-MD は $P(R|\{(D_m, E_m)\}_{m=1}^M)$ からの誤差 ϵ 以下のサンプリングを, $O\left((|S||\mathcal{A}|)^2 \log \frac{1}{\epsilon}\right)$ のステップ数で実現する. また, 報酬関数が特徴量 $\phi: \mathcal{S} \rightarrow \mathbb{R}^n$ の線形和 $R(s) = w \cdot \phi(s)$ で表される場合, PolicyWalk-MD は $P(R|\{(D_m, E_m)\}_{m=1}^M)$ からの誤差 ϵ 以下のサンプリングを, $O\left(n^2 \log \frac{1}{\epsilon}\right)$ のステップ数で実現する. 転移可能な報酬を推定するためには, 報酬を状態に依存する関数 $R: \mathcal{S} \rightarrow \mathbb{R}$ として定義すると良いと知られている [Fu 18]. そのため, 本研究では状態に依存する報酬関数 $R(s)$ を用いる.

5. 実験

実験では二つの環境で実験を行う. 一つ目は状態数と行動数が少なく, エキスパートが最適となる報酬が複数存在する環境で, 複数の環境の軌跡を用いることの有用性を確認する. もう一つは, 逆強化学習のベンチマークとして用いられる Grid world 環境で, 一つ目の環境と比較して状態数, 行動数が多い環境における提案法の有用性を確認する.

本実験で用いる評価指標について述べる. 本実験では, 提案法, BIRL の推定報酬と, エキスパートの報酬の近さを Expected Value Difference (EVD) [Levine 11] を用いて評価する. EVD の定義式を (28) に示す.

EVD =

$$E_{\pi_{\text{exp}}} \left[\sum_{t=0}^{\infty} \gamma^t R_{\text{exp}}(s_t) \right] - E_{\hat{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t R_{\text{exp}}(s_t) \right]$$

図 2 An example of IRL problem with multiple environments. Bold lines represent the optimal (expert's) action a_1 , and broken lines represent some other action a_2 .

(28)

ここで、 R_{exp} はエキスパートの報酬、 π_{exp} はエキスパートの方策、 $\hat{\pi}$ は推定報酬 $\hat{R}$ に対する最適方策を指す。EVD の値が小さいほど推定報酬がエキスパートの報酬と類似している。エキスパートの方策 π_{exp} はエキスパートの報酬 R_{exp} に対する最適方策であるため EVD は非負で、その最小値は 0 である。以下で、各実験の設定と、その結果について述べる。

5.1 Small Environments with Uncertainty in Reward Estimation

ここでは、エキスパートが最適となる報酬が複数存在することが明らかな環境を用いて、提案法と既存手法を比較する。図 2 に、報酬の推定に用いる二つの環境を示す。ここで、エキスパートの報酬を $R(s_0) = R(s_1) = 0$, $R(s_2) = R(s_3) = 0.7$ とした。これらの環境は、エキスパートが最適となる報酬が複数存在する。なぜなら、各環境におけるエキスパートの方策（最適方策）は、全状態のうち三つの状態を巡回する方策であり、エキスパートによって巡回されるいずれか、または複数の状態の報酬が大きい場合にエキスパートが最適となるからである。例えば、環境 A では、 s_1, s_3, s_2 という順に状態を巡回する方策がエキスパートの方策で、 s_1, s_3, s_2 のいずれか、または複数の状態の報酬が大きい場合に、エキスパートの方策が最適となる。一方、環境 B では、 s_0, s_3, s_2 という順に状態を巡回する方策がエキスパートの方策で、 s_0, s_3, s_2 のいずれかの状態、または複数の状態の報酬が大きい場合に、エキスパートの方策が最適となる。

EVD の評価には、状態遷移確率を一様分布によって定めた 100 個の環境を用いる。EVD の評価に用いる環境の状態数は 4、行動数は 2 で、報酬の推定に用いた環境と同じである。次に、エキスパートの軌跡について説明する。報酬の推定には、エキスパートの方策から生成された 5 ステップの軌跡を合計 8 個用いる。具体的には、BIRL で報酬を推定する際には、一つの環境でエキスパートの軌跡を 8 個、BIRL-MD で報酬を推定する際には、環

表 2 Evaluation of BIRL and BIRL-MD. BIRL uses only one environment for rewards estimation, and BIRL-MD uses 2 environments.

Method	BIRL	BIRL-MD
Number of environments	1	2
Total number of trajectories	8	8
EVD(Mean)	0.559	0.172
EVD(Max)	4.540	1.188

境 A と環境 B からそれぞれ 4 個、計 8 個の軌跡を用いて報酬を推定する。エキスパートの軌跡を生成する方策には、 $\gamma = 0.95$ の最適行動価値 Q^* に基づく $\epsilon = 0.1$ の ϵ -greedy 方策を用いた。

BIRL と BIRL-MD で報酬を推定する際には、式 (7), (11) のパラメータ κ , 事前分布 $P(R)$, MCMC の各種パラメータの値を定める必要がある。ここでは、式 (11) のハイパーパラメータ κ を 1.0, 事前分布 $P(R)$ を一様分布 $\mathcal{U}(-1, 1)$ とした。また、MCMC のステップ数を 1000, バーンインを 200, ステップサイズ δ を 0.05 とし、収束判定は標本系列の時系列プロットによる可視化に基づいて行なった。また、EVD の計算にはサンプルした報酬の平均値を用いた。

既存手法、提案法の推定報酬の EVD の平均値と最大値を表 2 に示す。表 2 の一行目は報酬の推定に用いた手法を、二行目は報酬の推定に用いた環境の数を、三行目は報酬の推定に用いた軌跡の合計の数を示す。また、四行目は転移先の 100 個の環境における EVD の平均値を、五行目は EVD の最大値を示す。表 2 の二列目に示す BIRL の EVD の平均値は、図 2 の各環境 A, B で推定した二つの報酬の EVD の平均値で、三列目に示す BIRL-MD の EVD は、二つの環境におけるエキスパートの軌跡から推定した報酬の EVD である。表 2 から分かるように、エキスパートの方策が最適となる報酬が複数存在することが明らかな環境において、複数の環境におけるエキスパートの軌跡を用いることによって、EVD の小さい報酬が推定された。EVD が小さいほど推定報酬がエキスパートの報酬に近いことから、ある報酬に従うエキスパート

図3 An example of windy grid world environments.

が複数の環境で生成した軌跡を用いて報酬を推定できる BIRL-MD が有用であると言える。

5.2 Windy Grid World Environments

本実験では、風が吹く Grid world 環境である Windy grid world 環境を用いる [Sutton 18]. Windy grid world 環境におけるエージェントは、風が吹く状態ではエージェントの行動に関わらず、ある一定の確率で風が吹く方向に遷移する。そのため、各状態の風向きを変えることによって、状態遷移確率が異なる環境を生成できる。本実験では、エージェントの行動に関わらず風が吹く方向に遷移する確率を 30% とする。風が吹く方向は上下左右の 4 方向で、無風の状態も存在し、各状態の風向きは独立とする。

本実験で用いる 5×5 マスの Windy grid world の例を図 3 に示す。図 3 の各マス目の矢印は、矢印のある状態で吹く風の方向を示す。矢印がない状態は無風で、エージェントが選択した行動に従い確率 1 で遷移する。風向きの数は 4 方向と無風を合わせた 5 パターンであることから、環境の数は $5^{|S|} = 5^{25}$ パターン存在する。本節では、Windy grid world 環境を用いて、二つの実験を行う。一つ目はエキスパートのデータ数を変える実験で、データ数に対する BIRL と BIRL-MD の性能を比較する。二つ目はモデルのパラメータ κ を変える実験で、異なる κ を推定に用いた際の既存法と提案法の性能を比較する。

エキスパートの報酬を図 4 に示す。エキスパートの報酬は最も右下の状態、座標 (4,4) の報酬が 1.0 で、それ以外の状態における報酬が 0 の一般的な Grid world 環境の報酬設計である。この報酬の下での最適方策は、右下のゴールと反対向きに風が吹く状態を避けつつゴールに向かう方策である。そのため、エキスパートの方策は、その環境の風向きによって異なる。EVD の評価には、各状態の風向きを一様分布で生成した 100 個の環境を用いる、また、各実験の評価値には 10 試行の平均値を用いた。エキスパートの軌跡は、割引率 $\gamma = 0.7$ 、 $\epsilon = 0.1$ の ϵ -greedy 方策を用いて生成した。各軌跡のステップ数は

図4 Expert's rewards in the windy grid world environments. Expert's rewards are shared among all environments.

図5 Evaluation of BIRL and BIRL-MD varying a total number of trajectories for reward estimation. Where M in the legend is a number of environments used for rewards estimation. Plotted values are mean of 10 trials.

15 ステップとした。MCMC の各パラメータは、ステップ数を 20000、バーンインを 5000、ステップサイズ δ を 0.05 とし、収束判定は標本系列の時系列プロットによる可視化に基づいて行なった。EVD の計算には、サンプルした報酬の平均値を用いた。BIRL と BIRL-MD で報酬の推定に用いる事前分布 $P(R)$ は一様分布 $\mathcal{U}(-1, 1)$ とした。なお、報酬の事前分布を標準正規分布にして実験を行なったが、一様分布の場合と同様の結果が得られた。

報酬に用いるエキスパートの軌跡の合計数を変更した計算機実験の結果を図 5 に示す。ここで、式 (11) の κ は 1.0 とし、エキスパートの軌跡の数は、報酬の推定に用いる各環境で均等に得られるものとした。図 5 のプロットの値は、EVD の平均値で、プロットと同じ色の領域は標準偏差を示す。BIRL、BIRL-MD のどちらもエキスパートのデータ数の増加に従い EVD が減少している。軌跡の数が 8 個の場合、報酬の推定に用いた環境の数が 2 個や 4 個の時の EVD の平均値が、環境の数が 1 個の時の EVD よりも大きな値を取っている。しかし、軌跡の数が

図6 Evaluation of BIRL and BIRL-MD varying a parameter κ for reward estimation. Plotted values are mean of 10 trials.

16 個以上の場合には、複数の環境を用いた方が EVD が小さく、複数の環境のエキスパートのデータを用いることによって、よりエキスパートに近い報酬が推定されていることが確認できる。このように、エキスパートの軌跡の数が一定数より多い場合は、複数の環境におけるエキスパートの軌跡を用いることによって、より EVD が小さい報酬が推定できると言える。

次に、式 (7), (11) のエキスパートが最適行動を取る確率を調節するパラメータ κ の値を変更した実験の結果を図 6 に示す。ここで、報酬の推定にはエキスパートの軌跡を合計 128 個用いた。図 6 の結果から、推定に用いるモデルのパラメータの値に依らず BIRL-MD が BIRL よりも小さい EVD の報酬を推定している。また、報酬の推定に用いる環境の数の増加に従い EVD が減少していることも確認できる。よって、複数の環境におけるエキスパートの軌跡から報酬を推定する提案法は、報酬推定のモデルのパラメータの変化に対して頑健だと言える。

6. 考察

複数の環境で得られたエキスパートの軌跡データを用いることによって、EVD が小さい報酬が推定された理由は、一つの環境から得られるエキスパートの報酬に関する情報には限りがあり、複数の環境におけるエキスパートの軌跡を用いることによって、エキスパートの報酬に関する情報をより多く得られるからだと考えられる。例えば、図 2 に示す二つの環境では、各環境においてエキスパートの報酬に関して得られる情報が異なる。具体的には、図 2 のうち、どちらか片方の環境におけるエキスパートの軌跡から得られる情報は、全 4 状態のうち 3 状態（環境 A では s_1, s_3, s_2 、環境 B では s_0, s_3, s_2 ）のいずれかの状態の報酬が高いという情報である。二つの環境の情報を統合することによって、 s_2, s_3 のいずれかの状態、または両方の状態の報酬が高いことは明らかである。

実際、提案法は複数の環境で得られたエキスパートの

軌跡データを用いて報酬を推定することによって、一つの環境を対象とする既存法よりも EVD の小さい報酬を推定している。これは Windy grid world 環境においても同様である。ある一つの環境における軌跡だけでは、エキスパートがある状態を訪れた理由が報酬が高いからなのか、風向きの影響なのか判別することは困難である。しかし、複数の環境のデータを用いることができれば、任意の状態遷移確率の環境において、エキスパートが高い頻度で訪れる状態が明確になり、エキスパートが訪れる頻度の高い状態には大きな報酬が存在すると推測できる。式 (11) で得られる事後分布は、多くの環境でエキスパートに共通する状態行動対の Q 値が高くなる報酬に高い確率を置くようモデル化されているため、エキスパートが複数の環境で生成した軌跡を用いることによって、より EVD が小さい、つまりエキスパートに近い報酬が推定されたと考えられる。

7. まとめと今後の課題

本研究では、「あるタスクの解法を知るエキスパートが、状態遷移確率の異なる複数の環境において生成した軌跡データ」を所与として、そのエキスパートの報酬分布を推定する問題を、BIRL の枠組みで定式化し、マルコフ連鎖モンテカルロ法 (MCMC) を用いた解法 PolicyWalk-MD を提案した。

実験では、逆強化学習手法の性能の評価指標 EVD [Levine 11] を用いて提案法と既存手法を比較した。その結果、二種類の実験設定において、提案法の推定報酬が、既存手法の推定報酬よりもエキスパートの報酬に近いことが確認された。

今後の課題を二つ挙げる。一つは BIRL-MD の異なる問題設定への拡張である。BIRL は、そのモデル化の自由度の高さから、様々な拡張が提案されている。その代表的な例としては、報酬だけでなく、その基底関数 [Choi 13] や、階層構造 [Michini 12]、複数種類の報酬 [Choi 12]、EVD の上界 [Brown 18] の推定法が挙げられる。本論文に基づいて、これら既存手法の問題設定を扱えるよう拡張することは今後の課題である。

もう一つの課題は、計算コストの削減である。提案法の適用範囲は、MCMC に要する計算量の大きさから小規模な離散状態行動空間の環境に限られる。そのため、連続状態行動空間を対象として、複数の環境から報酬を推定可能なベイジアン逆強化学習手法の構築を考えている。

◇ 参考文献 ◇

- [Abbeel 04] Abbeel, P. and Ng, A. Y.: Apprenticeship Learning via Inverse Reinforcement Learning, in *Proceedings of The Twenty-First International Conference on Machine Learning*, p. 1ACM (2004)
- [Amin 17] Amin, K., Jiang, N., and Singh, S.: Repeated Inverse Reinforcement Learning, in *Advances in Neural Information Processing Systems 30*, pp. 1815–1824, Curran Associates, Inc. (2017)

- [Amodei 16] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D.: Concrete Problems in AI Safety, *arXiv preprint arXiv:1606.06565* (2016)
- [Andrychowicz 18] Andrychowicz, M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., et al.: Learning Dexterous In-Hand Manipulation, *arXiv preprint arXiv:1808.00177* (2018)
- [Applegate 91] Applegate, D. and Kannan, R.: Sampling and Integration of Near Log-Concave Functions, in *Proceedings of The Twenty-Third Annual ACM Symposium on Theory of Computing*, pp. 156–163 ACM (1991)
- [Barto 89] Barto, A. G., Sutton, R. S., and Watkins, C. J.: Learning and Sequential Decision Making, in *Learning and Computational Neuroscience* CiteSeer (1989)
- [Bengio 12] Bengio, Y.: Practical Recommendations for Gradient-Based Training of Deep Architectures, in *Neural Networks: Tricks of The Trade*, pp. 437–478, Springer (2012)
- [Brown 18] Brown, D. S. and Niekum, S.: Efficient Probabilistic Performance Bounds for Inverse Reinforcement Learning, in *Thirty-Second AAAI Conference on Artificial Intelligence* (2018)
- [Choi 12] Choi, J. and Kim, K.-E.: Nonparametric Bayesian Inverse Reinforcement Learning for Multiple Reward Functions, in *Advances in Neural Information Processing Systems*, pp. 305–313 (2012)
- [Choi 13] Choi, J. and Kim, K.-E.: Bayesian Nonparametric Feature Construction for Inverse Reinforcement Learning, in *Twenty-Third International Joint Conference on Artificial Intelligence* (2013)
- [Clavera 19] Clavera, I., Nagabandi, A., Liu, S., Fearing, R. S., Abbeel, P., Levine, S., and Finn, C.: Learning To Adapt in Dynamic, Real-World Environments Through Meta-Reinforcement Learning, in *International Conference on Learning Representations* (2019)
- [Daume III 06] Daume III, H. and Marcu, D.: Domain Adaptation for Statistical Classifiers, *Journal of artificial intelligence research*, Vol. 26, pp. 101–126 (2006)
- [Duan 17] Duan, Y., Andrychowicz, M., Stadie, B., Ho, O. J., Schneider, J., Sutskever, I., Abbeel, P., and Zaremba, W.: One-Shot Imitation Learning, in *Advances in Neural Information Processing Systems*, pp. 1087–1098 (2017)
- [Everitt 16] Everitt, T. and Hutter, M.: Avoiding Wireheading with Value Reinforcement Learning, in *International Conference on Artificial General Intelligence*, pp. 12–22 Springer (2016)
- [Finn 17] Finn, C., Abbeel, P., and Levine, S.: Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks, in *Proceedings of The 34th International Conference on Machine Learning-Volume 70*, pp. 1126–1135 JMLR. org (2017)
- [Fu 18] Fu, J., Luo, K., and Levine, S.: Learning Robust Rewards with Adversarial Inverse Reinforcement Learning, in *International Conference on Learning Representations* (2018)
- [Jiang 08] Jiang, J.: A Literature Survey on Domain Adaptation of Statistical Classifiers, URL: <http://sifaka.cs.uiuc.edu/jiang4/domainadaptation/survey>, Vol. 3, pp. 1–12 (2008)
- [Kitani 12] Kitani, K. M., Ziebart, B. D., Bagnell, J. A., and Hebert, M.: Activity Forecasting, in *European Conference on Computer Vision*, pp. 201–214 Springer (2012)
- [Lamblin 10] Lamblin, P. and Bengio, Y.: Important Gains From Supervised Fine-Tuning of Deep Architectures on Large Labeled Sets, in *NIPS* 2010 Deep Learning and Unsupervised Feature Learning Workshop*, pp. 1–8 (2010)
- [Levine 11] Levine, S., Popovic, Z., and Koltun, V.: Nonlinear Inverse Reinforcement Learning with Gaussian Processes, in *Advances in Neural Information Processing Systems*, pp. 19–27 (2011)
- [Levine 16] Levine, S., Finn, C., Darrell, T., and Abbeel, P.: End-To-end Training of Deep Visuomotor Policies, *The Journal of Machine Learning Research*, Vol. 17, No. 1, pp. 1334–1373 (2016)
- [Li 17] Li, K. and Burdick, J. W.: Meta Inverse Reinforcement Learning via Maximum Reward Sharing for Human Motion Analysis, *arXiv preprint arXiv:1710.03592* (2017)
- [Michini 12] Michini, B. and How, J. P.: Bayesian Nonparametric Inverse Reinforcement Learning, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 148–163 Springer (2012)
- [Mnih 15] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Hiedmiller, M., Fidjeland, A. K., Ostrovski, G., et al.: Human-Level Control Through Deep Reinforcement Learning, *Nature*, Vol. 518, No. 7540, p. 529 (2015)
- [Ng 00] Ng, A. Y., Russell, S. J., et al.: Algorithms for Inverse Reinforcement Learning, in *Proceedings of The Seventeenth International Conference on Machine Learning*, pp. 663–670 (2000)
- [Nichol 18] Nichol, A., Pfau, V., Hesse, C., Klimov, O., and Schulman, J.: Gotta Learn Fast: A New Benchmark for Generalization in RL, *arXiv preprint arXiv:1804.03720* (2018)
- [Pan 17] Pan, X., You, Y., Wang, Z., and Lu, C.: Virtual To Real Reinforcement Learning for Autonomous Driving, *arXiv preprint arXiv:1704.03952* (2017)
- [Peng 18] Peng, X. B., Andrychowicz, M., Zaremba, W., and Abbeel, P.: Sim-To-Real Transfer of Robotic Control with Dynamics Randomization, in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8 IEEE (2018)
- [Pinto 17] Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A.: Robust Adversarial Reinforcement Learning, *arXiv preprint arXiv:1703.02702* (2017)
- [Pomerleau 89] Pomerleau, D. A.: Alvin: An Autonomous Land Vehicle in A Neural Network, in *Advances in Neural Information Processing Systems*, pp. 305–313 (1989)
- [Ramachandran 07] Ramachandran, D. and Amir, E.: Bayesian Inverse Reinforcement Learning, *IJCAI International Joint Conference on Artificial Intelligence*, pp. 2586–2591 (2007)
- [Ross 11] Ross, S., Gordon, G., and Bagnell, D.: A Reduction of Imitation Learning and Structured Prediction To No-Regret Online Learning, in *Proceedings of The Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627–635 (2011)
- [Silver 17] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al.: Mastering The Game of Go Without Human Knowledge, *Nature*, Vol. 550, No. 7676, p. 354 (2017)
- [Sutton 90] Sutton, R. S.: Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming, in *Machine Learning Proceedings 1990*, pp. 216–224, Elsevier (1990)
- [Sutton 18] Sutton, R. S. and Barto, A. G.: *Reinforcement Learning: An Introduction*, MIT press (2018)
- [Thrun 98] Thrun, S. and Pratt, L.: Learning To Learn: Introduction and Overview, in *Learning To Learn*, pp. 3–17, Springer (1998)
- [Tobin 17] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P.: Domain Randomization for Transferring Deep Neural Networks From Simulation To The Real World, in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30 IEEE (2017)
- [Vempala 05] Vempala, S.: Geometric Random Walks: A Survey, *Combinatorial and computational geometry*, Vol. 52, No. 573–612, p. 2 (2005)
- [Yan 17] Yan, M., Frosio, I., Tyree, S., and Kautz, J.: Sim-To-Real Transfer of Accurate Grasping with Eye-In-Hand Observations and Continuous Control, *arXiv preprint arXiv:1712.03303* (2017)
- [Yosinski 14] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H.: How Transferable Are Features in Deep Neural Networks?, in *Advances in Neural Information Processing Systems*, pp. 3320–3328 (2014)
- [Zamir 18] Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., and Savarese, S.: Taskonomy: Disentangling Task Transfer Learning, in *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722 (2018)
- [Ziebart 08] Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K.: Maximum Entropy Inverse Reinforcement Learning, in *AAAI*, Vol. 8, pp. 1433–1438 Chicago, IL, USA (2008)
- [Ziebart 10] Ziebart, B. D.: *Modeling Purposeful Adaptive Behavior with The Principle of Maximum Causal Entropy*, Carnegie Mellon University (2010)

〔担当委員：村田 真悟〕

2019 年 7 月 11 日 受理

著 者 紹 介

中田 勇介

2017 年千葉大学工学部都市環境システム学科卒業，2018 年同大学院融合理工学府博士前期課程修了．現在，同大学院融合理工学府博士後期課程在学．機械学習，特に逆強化学習の研究を行なっている．

荒井 幸代(正会員)

慶應義塾大学大学院理工学部卒業後，ソニー（株），東京工業大学院理工学研究科制御工学専攻修了．1998 年博士（工学）．Carnegie Mellon University, Robotics Institute, Fraunhofer AIS senior researcher を経て現在，千葉大学大学院工学研究院教授．マルチエージェントシステム研究に従事．計測自動制御学会，電気学会，電子情報通信学会，日本OR学会他，AAAI，ACM 各会員．