状態遷移確率の異なる MDP 環境間で無矛盾な報酬の推定法

Estimating Consistent Reward of Expert in Multiple Dynamics via Linear Programming Inverse Reinforcement Learning

中田 勇介
Yusuke Nakata

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
a1019nakata@gmail.com

荒井 幸代
Sachiyo Arai

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, linear programming

Summary

Reinforcement learning is a powerful framework for decision making and control, but it requires a manually specified reward function. Inverse reinforcement learning (IRL) automatically recovers reward function from policy or demonstrations of experts. Most of the existing IRL algorithms assume that the expert policy or demonstrations in a fixed environment is given, but there are cases these are collected in multiple environments. In this work, we propose an IRL algorithm that is guaranteed to recover reward functions from models of multiple environments and expert policy for each environment. We assume that the expert in multiple environments shares a reward function, and estimate reward functions for which each expert policy is optimal in the corresponding environment. To handle policies in multiple environments, we extend linear programming IRL. Our method solves the linear programming problem of maximizing the sum of the original objective functions of each environment while satisfying conditions of all the given environments. Satisfying conditions of all the given environments is a necessary condition to match with expert reward, and estimated reward by proposed method satisfies this necessary condition. In the experiment, using Windy grid world environments, we demonstrate that our algorithm is able to recover reward functions for which expert policies are optimal for corresponding environments.

1. はじめに

ある環境でのタスク実行に対して、最適な行動系列を生成するための方法の一つに強化学習がある。強化学習は一般に、環境モデルに対する事前知識や、逐次的な教師信号が不要で、人の設計の手間がかからない方法として注目されてきたが、実問題に対しては状態空間や報酬の設計がボトルネックであった。しかし、状態空間の設計に対して深層学習の適用が有望になったことから、再度注目が集まっている。一方、もう一つのボトルネックである報酬の設計に対しては、逆強化学習が注目されている。逆強化学習は、「実問題では多くの場合、環境モデルや、望ましい行動系列が既知である」という考えに基づいている。具体的には、＜状態集合、行動集合、状態遷移確率＞からなる環境モデルと、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）の軌跡や方策などのデータを所与として、エキスパートの報酬を推定する方法である。

逆強化学習によって推定された報酬の用途は大きく二つに分けられる。一つ目は、上述したタスク実行における最適な行動系列を獲得すること。もう一つは、行動のインセンティブとしての報酬を位置づけることによる動物や人の行動解析である。前者の研究には、自動運転 [Kuderer 15] や、ロボットアームの制御 [Finn 16] などがあり、報酬の設計は困難だが、エキスパートのデータが取得可能な問題に用いられている。

後者の研究として、Yamaguchi ら [Yamaguchi 18] は、逆強化学習を用いて空腹時と満腹時の線虫の温度走性（適温に向かって移動する性質）を解析しており、文献 [Kitani 12] は人の歩行解析、文献 [Yang 18] ではソーシャルメディアで各トピックに言及する人々の傾向を解析している。

本論文では後者に属する研究における報酬の考え方に着目する。例えば、Kitani らの研究 [Kitani 12] では、特定の駐車場における人の歩行データから推定した報酬を、異なる駐車場や歩道における人の歩行経路の予測に用いている。実際、異なる駐車場や歩道であっても、歩行者が

持つインセンティブとしての報酬は変わらないと考えるのは自然である。しかし、異なる駐車場や歩道は、状態集合や行動集合が同じであっても、状態遷移確率は同じである保証はない。これに対して、既存の逆強化学習手法の多く [Abbeel 04, Ng 00, Ramachandran 07, Ziebart 08] は、ある単一の状態遷移確率を持つ環境モデルと、エキスパートのデータから報酬を推定する。したがって、状態遷移確率が異なる環境に対しては、たとえ同じ報酬で行動するエキスパートのデータを用いたとしても、二つの環境ごとに報酬を推定する必要があるが、推定される報酬間の無矛盾性は保証していない。後者のタイプの研究のような、複数の異なる状態遷移確率を持つ環境間で、エキスパート（動物や歩行者）のデータを説明するための報酬推定においては、ある特定の環境だけでなく、各環境におけるエキスパート（動物や歩行者）のデータを矛盾なく説明する報酬を推定する必要がある。

そこで本論文では、「状態遷移確率が異なる複数の環境下で得られるエキスパート方策」と、推定報酬との無矛盾性を保証する逆強化学習手法を提案する。具体的には、複数の環境から報酬を推定する問題を線形計画問題として定式化し、「状態遷移確率だけが異なる複数の環境のモデル」と「各環境でのエキスパートの方策」から、各環境におけるエキスパート方策が推定された報酬に対して最適方策であることを保証する。推定報酬に対して、各環境におけるエキスパート方策が最適となることは、推定報酬がエキスパートの報酬と一致するための必要条件である。

実験には、ある一定の確率で風が吹く方向に遷移する Windy grid world 環境を用いる。風向きが異なる複数の Windy grid world 環境と、各環境におけるエキスパート方策を用意し、既存手法の推定報酬が、各環境におけるエキスパート方策と矛盾することを確認する。また、提案法の推定報酬が、各環境におけるエキスパート方策に無矛盾であることを示す。

以下、2 章で本論文を読み進めるために必要な逆強化学習の概要を説明する。3 章で対象問題と関連研究、4 章で提案法について述べる。5 章で計算機実験を行い、6 章で実験結果を考察する。7 章でまとめと今後の課題を整理する。

2. 準備

本章では、マルコフ決定過程、強化学習、提案法の基礎となる線形計画法を用いた逆強化学習 [Ng 00] について述べる。

2.1 マルコフ決定過程 (MDP)

マルコフ決定過程は、エージェントの行動による状態遷移にマルコフ性を仮定した数学モデルである。

本論文の対象とする有限マルコフ決定過程は $\mathcal{M} = \langle \mathcal{S},$

$\mathcal{A}, P, \gamma, R \rangle$ からなる。 $\mathcal{S}$ は有限状態集合、 $\mathcal{A}$ は行動集合である。 $P(s'|s, a)$ は状態 $s \in \mathcal{S}$ で行動 $a \in \mathcal{A}$ をとった時、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 $s \in \mathcal{S}$ の報酬を返す関数である。

エージェントは時刻 t において状態 $s_t \in \mathcal{S}$ を観測し、自身の方策 $\pi: \mathcal{S} \rightarrow \mathcal{A}$ に基づいて行動 $a_t \in \mathcal{A}$ を選択する。その後、時刻 $t+1$ では s_t, a_t によって確率的に次状態 s_{t+1} に遷移し、報酬 $R(s_{t+1})$ を得る。

2.2 強化学習

強化学習は、状態遷移確率などの環境のモデルを所与とせず、最適方策を獲得する方法である。最適方策とは、方策 π の下で得られる累積報酬の期待値が最大の方策である。

方策 π の下で得られる報酬の期待値は、状態価値関数 $V^\pi: \mathcal{S} \rightarrow \mathbb{R}$ と行動価値関数 $Q^\pi: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ で表される。状態価値 $V^\pi(s)$ 、行動価値 $Q^\pi(s, a)$ の定義式をそれぞれ式 (1)、式 (2) に示す。

$$V^\pi(s) = E_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | s_t = s \right] \quad (1)$$

$$Q^\pi(s, a) = E_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | s_t = s, a_t = a \right] \quad (2)$$

ここで、 R_t はエージェントが時刻 t に獲得する報酬を指す。状態価値 $V^\pi(s)$ は、方策 π の下で状態 s から得られる報酬の期待値、行動価値 $Q^\pi(s, a)$ は、方策 π の下で状態 s で行動 a を選択した時に得られる報酬の期待値を示す。

状態価値と行動価値は、それぞれ式 (3)、式 (4) のベルマン方程式を満たす。

$$V^\pi(s) = R(s) + \gamma \sum_{s'} P(s'|s, \pi(s)) V^\pi(s') \quad (3)$$

$$Q^\pi(s, a) = R(s) + \gamma \sum_{s'} P(s'|s, a) V^\pi(s') \quad (4)$$

最適方策 π^* の状態価値 $V^{\pi^*}(s)$ は任意の状態 $\forall s \in \mathcal{S}$ において最大値を取る。最適方策 π^* が満たす条件を式 (5) に示す。

$$V^{\pi^*}(s) = R(s) + \gamma \max_{\pi} \sum_{s'} P(s'|s, \pi(s)) V^{\pi^*}(s') \quad \forall s \in \mathcal{S} \quad (5)$$

2.3 線形計画法を用いた逆強化学習 (LPIRL)

逆強化学習は、環境 $\mathcal{M}$ から報酬 R を除いた環境のモデル $\mathcal{M} \setminus R$ と、エキスパートの方策 π_E や軌跡を所与として、エキスパート方策 π_E が最適となる報酬 R を推定する。

Ng らは、エキスパートの報酬を推定する線形計画問題を三つ定式化した [Ng 00]。これら三つの定式化は、状態行動空間とエキスパートのデータに対する前提が異なる。一つ目は、離散状態行動空間の環境のモデル $\mathcal{M} \setminus R$ と、エキスパートの方策 π_E を所与とする定式化、二つ目は、連続状態行動空間の環境のモデル $\mathcal{M} \setminus R$ と、エキスパート方策 π_E を所与とする定式化、三つ目は、エキスパートの軌跡が所与の場合の定式化である。三つ目の定式化は、状態行動空間が離散、連続のどちらの場合にも適用できる。また、任意の報酬に対する最適方策が学習できる環境である場合には環境のモデル $\mathcal{M} \setminus R$ は未知でもよい。一つ目の定式化では、各状態の報酬を独立に推定するものの、二つ目と三つ目の定式化は、報酬関数を線形関数で近似する。

提案法は、Ng らが提案した一つ目の定式化、離散状態行動空間の環境のモデル $\mathcal{M} \setminus R$ と、エキスパートの方策 π_E を所与とする定式化に基づく。以下で、離散状態行動空間の環境のモデル $\mathcal{M} \setminus R$ と、エキスパートの方策 π_E を所与として、エキスパート方策 π_E が最適となる報酬 R を推定する線形計画問題の定式化 (LPIRL) を示す。

§1 状態に依存する報酬 $R(s)$ の推定

Ng ら [Ng 00] は、環境のモデル $\mathcal{M} \setminus R$ とエキスパート方策 π_E を所与とし、エキスパートの報酬 R を推定する問題を、線形計画問題として定式化した。

Ng らの定式化を式 (6), (7) に示す。

$$\begin{aligned} \text{maximize : } & \sum_{i=1}^{|\mathcal{S}|} \min_{a \in \mathcal{A} \setminus a_1} \{ (P_{a_1}(i) - P_a(i)) \\ & (I - \gamma P_{a_1})^{-1} R \} - \lambda \|R\|_1 \quad (6) \end{aligned}$$

$$\begin{aligned} \text{subject to : } & (P_{a_1} - P_a)(I - \gamma P_{a_1})^{-1} R \succeq 0 \\ & \forall a \in \mathcal{A} \setminus a_1 \quad (7) \end{aligned}$$

$$|R_i| \leq R_{\max}, \quad i = 1, \dots, |\mathcal{S}| \quad (8)$$

ここで、 a_1 は状態 s におけるエキスパートの行動 $\pi_E(s)$ を、 $|\mathcal{S}|$ は状態数を指す。状態遷移行列 P_a は状態遷移確率 $P(s_j|a, s_i)$ を (i, j) 成分とする $|\mathcal{S}| \times |\mathcal{S}|$ 行列、状態遷移ベクトル $P_a(i)$ は P_a の第 i 行ベクトルを表す。 λ は報酬の総量を調節するペナルティ係数である。

式 (6) に示す目的関数の第一項はエキスパートの行動と2番目により行動の行動価値の差の総和である。式 (6) の第二項は、報酬の総量に比例するペナルティで、報酬の総量を調節する役割を果たす。

式 (7) は、エキスパート方策 π_E が報酬 R に対して最適方策となることを保証する制約条件であり、式 (9) と表すことができる。

$$Q^{\pi_E}(s, a_1) \geq Q^{\pi_E}(s, a) \quad \forall a \in \mathcal{A} \setminus a_1, \forall s \in \mathcal{S} \quad (9)$$

式 (9) を満たす報酬の下で、エキスパート方策 π_E は最適方策である。なぜなら、式 (9) を満たす報酬の下では、任意の状態におけるエキスパートの行動 a_1 の行動価値 $Q(s, a_1)$ は、それ以外の行動 $a \in \mathcal{A} \setminus a_1$ の行動価値 $Q(s, a)$ 以上の値をとるからである。式 (9) と式 (7) は等しいため、式 (7) を満たす報酬の下でエキスパート方策 π_E は最適方策である。

一方、式 (7) を満たさない報酬の下では、エキスパート方策 π_E は最適方策でない。そのため、式 (7) を満たすことは、推定報酬 $\hat{R}$ がエキスパートの報酬と一致する必要条件である。

§2 状態と行動に依存する報酬 $R(s, a)$ の推定

2.3 節 §1 で示した式 (6), (7), (8) からなる線形計画問題は、状態に依存し、行動に依存しない報酬 $R: \mathcal{S} \rightarrow \mathbb{R}$ を推定する定式化である。

状態と行動の両方に依存する報酬 $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ を推定する線形計画問題も、エキスパート方策 π_E が報酬 R に対して最適方策となることを保証する制約条件を導出することによって定式化できる。以下に、制約条件の導出を示す。はじめに、状態と行動の両方に依存する報酬の下での行動価値は式 (10) と定義される。

$$Q^{\pi}(s, a) = R(s, a) + \gamma \sum_{s'} P(s'|s, a) V^{\pi}(s') \quad (10)$$

各状態で行動 a をとった時に得られる報酬 $R(s, a)$ を要素とするベクトルを R^a とおけば、式 (9), (10) から、エキスパート方策 π_E が報酬 $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ に対して最適方策となることを保証する制約条件、式 (11) が導かれる。

$$\begin{aligned} R(s, a_1) + \sum_{s'} P(s'|s, a_1) V^{\pi_E}(s') \\ \geq R(s, a) + \sum_{s'} P(s'|s, a) V^{\pi_E}(s') \\ \forall a \in \mathcal{A} \setminus a_1, \forall s \in \mathcal{S} \\ \Leftrightarrow R^{a_1} + \gamma P_{a_1} V^{\pi_E} \succeq R^a + \gamma P_a V^{\pi_E} \quad \forall a \in \mathcal{A} \setminus a_1 \\ \Leftrightarrow R^{a_1} + \gamma P_{a_1} (I - \gamma P_{a_1})^{-1} R^{a_1} \\ \succeq R^a + \gamma P_a (I - \gamma P_{a_1})^{-1} R^{a_1} \quad \forall a \in \mathcal{A} \setminus a_1 \\ \Leftrightarrow R^{a_1} - R^a + \gamma (P_{a_1} - P_a) (I - \gamma P_{a_1})^{-1} R^{a_1} \succeq 0 \\ \forall a \in \mathcal{A} \setminus a_1 \quad (11) \end{aligned}$$

式 (11) はエキスパート方策 π_E が報酬 R に対して最適方策となることを保証する制約条件、式 (7) と対応する。

表 1: 環境と推定報酬の数に基づく逆強化学習手法の分類

		推定報酬	
		単一	複数
環境のモデル	単一	LPIRL [Ng 00], AL [Abbeel 04], BIRL [Ramachandran 07], MaxEnt IRL [Ziebart 08]	ALMI [Babes 11], BNIRL sMDP [Surana 14]
	複数	Repeated IRL [Amin 17], Proposed method	Meta IRL [Li 17]

式 (11) に基づいて、目的関数、式 (7) と対応する制約条件、各状態行動対の報酬の最大値を定める制約条件、からなる線形計画問題を解くことによって、状態と行動の両方に依存する報酬を推定できる。

3. 問題設定

3.1 対象問題

はじめに、本論文が対象とする「複数の環境間の状態遷移確率の差異」が生じる要因を具体例を用いて説明する。次に、状態遷移確率が異なる複数の環境のモデルと、各環境におけるエキスパートの方策から報酬を推定する対象問題を、マルコフ決定過程の枠組みで定式化する。

「複数の環境間の状態遷移確率の差異」が生じる二つの要因を、車のブレーキの制御を例に説明する。状態遷移確率が異なる一つの要因はエージェントの個体差である。車においては、その重量、ブレーキの摩耗量などがエージェントの個体差に該当する。車の個体差があった時、複数の車のブレーキを同じ力で制御しても各車の減速度が、その個体差（例：重量、ブレーキの摩耗の仕方）によって異なることが分かる。

もう一つの要因は、エージェントが観測できない環境の差異である。例えば、車が走行する路面の摩擦係数は、塗装の種類や天候によって異なり、エージェントは路面の摩擦係数を観測できない。路面の摩擦係数が異なれば、特定の車のブレーキを同じ力で制御した時に生じる減速度も異なる。これら、二つの要因、エージェントの個体差、エージェントが観測できない環境の差異、によって生じる車の減速度の違いは、マルコフ決定過程における状態遷移確率の差異として等価に扱うことができる。これら状態遷移確率が異なる複数の環境においてブレーキを制御するエキスパートが、ある一つの報酬に従うと仮定すると、ある特定の環境だけでなく、各環境におけるエキスパートのデータを矛盾なく説明する報酬を推定する必要がある。

本論文では、状態遷移確率が異なる複数の環境におけるエキスパートが、ある一つの報酬 R に従う仮定の下で、エキスパートの報酬 R を推定する。具体的には、状態遷移確率のみ異なる環境のモデルの集合族 $\mathfrak{M} = \{\mathcal{M}_j \setminus R\}_{j=1}^{|\mathfrak{M}|}$ と、各環境 $\mathcal{M}_j$ におけるエキスパート方策 π_E^j の集合

$\Pi_E = \{\pi_E^j\}_{j=1}^{|\mathfrak{M}|}$ を所与として、各環境 $\mathcal{M}_j \in \mathfrak{M}$ におけるエキスパート方策 $\pi_E^j \in \Pi_E$ が最適となる報酬 R を推定する。

式 (5) より、状態遷移確率 P^j の環境のモデル $\mathcal{M}_j \setminus R$ においてエキスパート方策 π_E^j が最適となる報酬 R は、式 (12) を満たす。

$$V^{\pi_E^j}(s) = R(s) + \gamma \max_{\pi} \sum_{s'} P^j(s'|s, \pi(s)) V^{\pi}(s') \quad \forall s \in \mathcal{S} \quad (12)$$

以上のことから、本論文の対象問題は、所与の $\mathcal{M}_j \setminus R$ と π_E^j のすべての組 $(\mathcal{M}_j \setminus R, \pi_E^j)$ が、式 (12) を満たす報酬 R の推定である。

3.2 関連研究

提案法は、LPIRL に基づいて、複数の環境のモデルと、各環境におけるエキスパートのデータを扱うモデルベースの逆強化学習である。近年、[Finn 16] をはじめとした、状態遷移確率が未知の環境に適用可能なモデルフリー逆強化学習手法も提案されている。しかし、著者らが知る限り、複数の環境のモデルや、複数のエキスパートのデータをモデルフリー逆強化学習で扱う方法はない。したがって、以下では、本論文と同様、複数の環境のモデルや、複数のエキスパートのデータを扱うモデルベース逆強化学習手法に限定し、提案法との違いを述べる。

表 1 に、環境モデルの数と、推定報酬の数に基づいた逆強化学習の分類を示す。

ある環境モデルの下で、一つの推定報酬を得るアルゴリズムのうち、LPIRL[Ng 00] は、線形計画問題を解くことによって報酬を推定するため、動的計画法や強化学習を複数回解くことを要する AL[Abbeel 04], BIRL[Ramachandran 07], MaxEnt IRL[Ziebart 08] と比較して計算コストに優れているためである。

ある環境で、異なる報酬に従う複数のエキスパートのデータが混在するデータセットから、各エキスパートの報酬を推定する方法には、EM アルゴリズムを用いた逆強化学習 (ALMI) [Babes 11], ノンパラメトリックベイズを用いた逆強化学習 (BNIRL sMDP)[Surana 14] がある。これらの手法は、状態遷移確率が単一の環境で、(異なる

報酬関数を持つ) 複数のエキスパートが生成したデータを扱う。これに対して、提案法は、状態遷移確率が異なる複数の環境を扱う点が、ALMI [Babes 11], BNIRL sMDP [Surana 14] と異なる。

複数の環境のモデルを扱い、複数の報酬を推定する逆強化学習手法に Meta IRL [Li 17] がある。Li らは、各エキスパートの推定報酬の差を小さくするペナルティ項を Maximum Entropy IRL(MaxEnt IRL) [Ziebart 08] の目的関数に追加し、各エキスパート間で類似した複数の報酬を推定する方法を提案している。各エキスパート間で類似した報酬を推定することによって、各環境におけるエキスパートの軌跡のデータ量が少ない時に、推定報酬のエキスパートの軌跡への過学習を抑制している。

Meta IRL は、複数の環境のモデルを扱う点で提案法と類似しているが、各環境におけるエキスパートの報酬に対する仮定が異なる。Meta IRL は、各環境におけるエキスパートが異なる報酬を持つという仮定の下で、複数の報酬を推定する。そのため、Meta IRL によって複数の環境のデータと無矛盾な報酬が推定されることはない。一方、提案法は、各エキスパートが同じ報酬を持つと仮定し、推定報酬が複数の環境のデータと無矛盾であることが理論的に保証される。

複数の環境から単一の報酬を推定する手法に Repeated IRL [Amin 17] がある。Repeated IRL は、複数の環境に対する情報が逐次的に与えられる場合の報酬更新方法として提案されている。したがって、その推定報酬は環境が与えられる順番と報酬更新の閾値（ハイパーパラメータ）に依存する。一方、提案法は逐次的な更新ではなく、複数の環境におけるエキスパートのデータがバッチで与えられる場合を対象とし、複数の環境間の無矛盾性を保証することを目的としている点で立場が異なる。

4. アプローチ

提案法は、LPIRL[Ng 00] を拡張し、状態遷移確率が異なる複数の環境のモデルと、各環境におけるエキスパート方策から報酬を推定する。

提案法の定式化を以下に示す。

$$\begin{aligned} \text{maximize : } & \sum_{j=1}^{|\mathfrak{M}|} \sum_{i=1}^{|\mathcal{S}|} \min_{a^j \in \mathcal{A} \setminus a_1^j} \{ (P_{a_1}^j(i) - P_a^j(i)) \\ & (I - \gamma P_{a_1}^j)^{-1} R \} - \lambda \|R\|_1 \quad (13) \end{aligned}$$

$$\begin{aligned} \text{subject to : } & (P_{a_1}^1 - P_a^1)(I - \gamma P_{a_1}^1)^{-1} R \succeq 0 \\ & \forall a \in \mathcal{A} \setminus a_1^1 \\ & \vdots \\ & (P_{a_1}^{|\mathfrak{M}|} - P_a^{|\mathfrak{M}|})(I - \gamma P_{a_1}^{|\mathfrak{M}|})^{-1} R \succeq 0 \\ & \forall a \in \mathcal{A} \setminus a_1^{|\mathfrak{M}|} \quad (14) \end{aligned}$$

$$|R_i| \leq R_{\max}, \quad i = 1, \dots, |\mathcal{S}| \quad (15)$$

式 (13) の第一項は各環境における LPIRL の目的関数、式 (6) の総和である。式 (13) の第二項は、報酬の総量に対するペナルティを指す。式 (14) は各環境のモデル $\mathcal{M}_j \setminus R \in \mathfrak{M}$ の下で、エキスパート方策 $\pi_E^j \in \Pi_E$ が最適となる制約条件をすべて列挙したものであり、各制約条件式 (14) は、式 (7) と等しい。状態遷移行列 P_a は状態遷移確率 $P(s_j|a, s_i)$ を (i, j) 成分とする $|\mathcal{S}| \times |\mathcal{S}|$ 行列、 a_1^j は環境 $\mathcal{M}_j$ の各状態 s におけるエキスパートの行動 $\pi_E^j(s)$ を示す。

提案法の定式化は、推定報酬 $\hat{R}$ がエキスパートの報酬と一致するための必要条件から導かれる。推定報酬 $\hat{R}$ がエキスパートの報酬と一致する必要条件とは、各環境のモデル $\mathcal{M}_j \setminus R \in \mathfrak{M}$ におけるエキスパート方策 $\pi_E^j \in \Pi_E$ が、推定報酬 $\hat{R}$ に対して最適となることである。なぜなら、任意の状態遷移確率の環境におけるエキスパート方策は、エキスパートの報酬に対して最適方策となるからである。

提案法の推定報酬 $\hat{R}$ が、エキスパートの報酬と一致する必要条件を満たす理由を以下で述べる。ある環境 $\mathcal{M}$ において、報酬 R に対してエキスパート方策 π_E が最適となる条件は、式 (7) で表される。よって、各環境 $\mathcal{M}_j \setminus R \in \mathfrak{M}$ におけるエキスパート方策 $\pi_E^j \in \Pi_E$ が最適となるすべての制約条件式 (14) を、報酬 R が満たすとき、各環境のモデル $\mathcal{M}_j \setminus R \in \mathfrak{M}$ における各エキスパート方策 $\pi_E^j \in \Pi_E$ は報酬 R に対して最適方策である。よって、提案法の推定報酬は、エキスパートの報酬と一致する必要条件「各環境のモデル $\mathcal{M}_j \setminus R \in \mathfrak{M}$ におけるエキスパート方策 $\pi_E^j \in \Pi_E$ が、推定報酬 $\hat{R}$ に対して最適となる」を満たす。

提案法の定式化の実行可能領域を、LPIRL の実行可能領域と比較し説明する。各環境 $\mathcal{M}_j \in \mathfrak{M}$ において、エキスパート方策 $\pi_E^j \in \Pi_E$ が最適となる報酬 R の集合は、各環境の制約条件式 (7) を満たす実行可能領域の共通集合と等価である。提案法の定式化の制約条件式 (14) は、各環境の制約条件式 (7) から構成されるため、提案法の定式化の実行可能領域もまたこれと等しい。

提案法の実行可能領域は、各環境のモデル $\mathcal{M}_j \setminus R \in \mathfrak{M}$ と、それに対応するエキスパート方策 $\pi_E^j \in \Pi_E$ から構成される LPIRL[Ng 00] の実行可能領域の共通集合と一致する。そのため、提案法の定式化の下で、報酬を推定できれば、推定報酬 $\hat{R}$ と各環境のモデル $\mathcal{M}_j \setminus R \in \mathfrak{M}$ に

図 1: 風向きが異なる Windy grid world 環境

におけるエキスパート方策 $\pi_E^j \in \Pi_E$ との無矛盾性を保証できる。

本論文で定式化した問題は、線形計画問題であるため、シンプレックス法などの線形計画問題を解く標準的なアルゴリズムを用いて解が求まる。ここでは、提案法の定式化におけるシンプレックス法の計算量を LPIRL の定式化と比較する。シンプレックス法の計算量は、実行可能基底解の数に依存する。線形計画問題の実行可能基底解の数の上界は、その標準形における変数の数 j 、制約条件の数 k を用いて、

$$j \left[k \frac{\beta}{\alpha} \log k \frac{\beta}{\alpha} \right] \quad (16)$$

と表される [Kitahara 13]。ここで α, β はそれぞれ、全ての実行可能基底解の正の要素の最小値、最大値である。線形計画問題が非退化の時、実行可能基底解の数の上界はシンプレックス法の反復回数の上界となる。

提案法の定式化の標準形における変数の数 j と、制約条件の数 k はそれぞれ、

$$j = |\mathcal{S}|(|\mathcal{M}| + 2) \quad (17)$$

$$k = 2|\mathcal{M}||\mathcal{S}|(|\mathcal{A}| - 1) + 4|\mathcal{S}| \quad (18)$$

である。式 (16), (17), (18) から、提案法の定式化における実行可能基底解の数の上界は、環境の数 $|\mathcal{M}|$ 、状態数 $|\mathcal{S}|$ 、行動数 $|\mathcal{A}|$ 、に対して $\mathcal{O}(|\mathcal{A}|(|\mathcal{S}||\mathcal{M}|)^2 \log |\mathcal{A}||\mathcal{S}||\mathcal{M}|)$ のオーダーで増加することがわかる。

本章で提案したアプローチは、式 (11) を用いることによって、状態と行動に依存する報酬 $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ の推定にも適用きる。

5. 実験

提案法の推定報酬 $\hat{R}$ が、各環境のモデル $\mathcal{M}_j \setminus R \in \mathcal{M}$ の下で、エキスパート方策 $\pi_E^j \in \Pi_E$ に無矛盾であることを示す。

実験には、ある一定の確率で風が吹く方向に遷移する Windy grid world 環境を用いる。はじめに、実験設定を示し、実験結果について述べる。

5.1 実験設定

風向きが異なる二つの Windy grid world 環境を用いて、LPIRL [Ng 00] と提案法を比較する。Windy grid world 環境とは、風が吹く迷路問題で、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。エージェントの行動は五つで、上下左右へ移動する四つの行動とその場に留まる行動 x からなる。

図 1 に風向きが異なる二つの実験環境を示す。図 1 の灰色の矢印は、矢印のある状態で吹く風の方向を示す。本実験では風が吹く方向に遷移する確率を 30% とした。各マス目の網の濃淡は報酬の大きさに比例し、右側のバーは網の濃淡に対応する報酬の値を示す。各マス目の数値は各状態 $s \in \mathcal{S}$ の報酬 $R(s)$ であり、バーが示す報酬の値に対応する。最適方策 π^* は報酬の期待値を最大化する方策であり、各環境 $\mathcal{M}_j$ における最適方策 π_j^* はその風向きによって異なる。

図 2 に、図 1 の各環境における最適方策を示す。図 2 の各状態の矢印は、各状態における最適行動を指す。最適行動の計算には割引率 $\gamma = 0.3$ の価値反復 [Sutton 98] を用いた。これら二つの環境における、風が吹く状態の最適行動は、風が吹いていない状態へと最短経路で移動する行動である。具体的には、風が吹く状態の最適行動は、風向きが左の状態の下、風向きが上の状態で右である。

本論文では、これら二つの環境それぞれの最適方策を、各環境におけるエキスパート方策とし、エキスパートの報酬を推定する。報酬を推定する際、既存手法、提案法の報酬の最大値 $R_{\max}$ を定める必要がある。ここでは、制約条件、式 (8), (15) に示す報酬の最大値 $R_{\max}$ をエキスパートの報酬の最大値と同じ値 10 として実験を行った。

報酬のスパース性を調整するパラメータ λ は、最もス

図 2: 各 Windy grid world 環境におけるエキスパート方策

パースな解（全状態の報酬が 0 の解を除く）が得られるパラメータ λ の最小値を、0.1 刻みのグリッドサーチで探索し選択した。ここでは、既存手法の λ を 1.1, 提案法の λ を 0.1 とした。

5.2 実験結果

既存手法と提案法で報酬を推定し、各環境におけるエキスパート方策と、推定報酬に対する最適方策を比較する。

表 2 に各手法による推定報酬と、推定報酬に対する図 1 の各環境の最適方策を示す。表 2 の 1 行目は、各列の図の内容を示す。具体的には、1 列目の各図は報酬、2 列目の各図は風が左へ吹く環境における最適方策、3 列目の各図は風が上へ吹く環境における最適方策である。

1 列目の各図は上から、エキスパートの報酬、風が左へ吹く環境のモデルとエキスパート方策に基づく推定報酬（既存手法）、風が上へ吹く環境のモデルとエキスパート方策に基づく推定報酬（既存手法）、二つの環境のモデルと各環境におけるエキスパート方策に基づく推定報酬（提案法）である。各状態の値は、その状態 s の報酬 $R(s)$ を表す。2 列目の各図は、風が左に吹く環境における、各行の 1 列目の報酬に対する最適方策で、3 列目の各図は、風が上に吹く環境における各行の 1 列目の報酬に対する最適方策である。

ここで、右の二列の各図中の各状態における矢印は、その状態における最適行動を指す。各状態の最適行動を表す矢印の色は、エキスパート方策との一致、不一致を示す。黒色の矢印は、最適行動とエキスパートの行動との一致を示し、赤色の矢印は、最適行動とエキスパートの行動との不一致を示す。

右の二列の各図中の各状態の網の濃淡は状態価値を表し、網の濃淡は状態価値に比例する。各図の右側のバーは網の濃淡に対応する状態価値の値を示す。

同じ報酬に対する最適方策から報酬を推定しても、その最適方策と、状態遷移確率が異なれば、異なる報酬が

推定される。例えば、表 2 の 3 行 1 列目の図は、風向きが左の環境のモデルと最適方策から推定した報酬、4 行 1 列目の図は、風向きが左の環境のモデルと最適方策から推定した報酬を示す。これら二つの推定報酬の値が異なることから、同じ報酬に対する最適方策から報酬を推定しても、その最適方策と、状態遷移確率が異なれば、推定報酬も異なる場合があることが確認できる。

既存手法では、風向きが異なる環境において、推定報酬に対する最適方策はエキスパート方策と一致しない。例えば、表 2 の 3 行 3 列目の図は、風向きが左の環境で推定した報酬に対する、風向きが上の環境における最適方策であり、図の座標 (2,2) における最適行動は「下」である。しかし、表 2 の 2 行 3 列目の図からわかるように、エキスパートの報酬に対する最適行動は「右」である。このように、同じ報酬に対するエキスパート方策を用いて報酬を推定しても、エキスパート方策とは異なる方策が最適方策となる。

表 2 の 5 行目は、二つの環境における既存手法の推定報酬を平均した結果である。複数の環境における推定報酬の平均値に対する最適方策はエキスパート方策と一致しないことが確認できる。

提案法は、既存手法と異なり、風向きが異なる各環境におけるエキスパート方策が最適となる報酬を推定した。これは、環境の風向きにかかわらず、各環境におけるエキスパート方策（表 2, 2 行目）と推定報酬に対する最適方策（表 2, 最下行）が一致したことから確認できる。

次に、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 は、推定報酬に対する最適方策とエキスパートの方策の近さを表す指標で、EVD の値が小さいほど推定報酬もエキスパートの報酬と類似している。推定報酬の EVD は、エキスパート方策で得られるエキスパートの報酬の期待値から、推定報酬に対する最適方策で得られるエキスパートの報酬の期待値を引くことで求まる。ここでは、

表 2: 既存手法と提案法の比較. 1 列目の各図は報酬, 2 列目の各図は風が左へ吹く環境における最適方策, 3 列目の各図は風が上へ吹く環境における最適方策である.

表 3: Expected Value Difference の比較

Method	Environment	EVD (Mean $\pm$ Std)
Ng	Left	$6.97 \times 10^{-3} \pm 8.55 \times 10^{-3}$
Ng	Up	$6.51 \times 10^{-3} \pm 8.23 \times 10^{-3}$
Ng (Mean)	Both	$5.89 \times 10^{-4} \pm 8.09 \times 10^{-4}$
Proposed	Both	$3.91 \times 10^{-4} \pm 7.97 \times 10^{-4}$

EVD の評価には各状態の風向きをランダムに定めた 500 個の環境を用いる。

表 3 に、各手法による推定報酬の EVD を示す。提案法の EVD が最も小さいことから、提案法によって、よりエキスパートに近い報酬が推定されたことが確認できる。

6. 考 察

本章では、5.2 節の実験結果に基づき以下の五点を考察する。

6.1 状態遷移確率が異なれば、同じ報酬のエキスパート方策から報酬を推定しても、異なる報酬が推定される要因

表 2 の 3 行 1 列目、4 行 1 列目は、異なる状態遷移確率の環境の下での、既存手法の推定報酬である。各図を比較すると、各環境間で、推定報酬が異なることが確認できる。ここでは、同じ報酬に対するエキスパート方策を用いても、状態遷移確率の異なる環境間で、推定報酬が異なる原因を考察する。

状態遷移確率の異なる環境間で、推定報酬が異なる原因は二つある。一つ目は、状態遷移確率が異なる環境間での、線形計画問題の実行可能領域の違いである。線形計画問題の制約条件の式は、状態遷移確率を含むため、状態遷移確率が異なれば、線形計画問題は異なる制約条件を持つ。そのため、状態遷移確率が異なる環境間では、実行可能領域が異なり、異なる報酬が推定される場合がある。

二つ目の原因は、状態遷移確率が異なる環境間の線形計画問題の目的関数の違いである。一つ目の要因である、実行可能領域の違いに関する議論が、目的関数に対しても同様に成り立つ。異なる目的関数を持つ問題の解は異なる場合があるため、状態遷移確率が異なる環境間で、異なる報酬が推定される場合がある。これら二つが、同じ報酬に対するエキスパート方策を用いても、状態遷移確率の異なる環境間で推定報酬が異なる原因である。

6.2 報酬を推定した環境と異なる状態遷移確率の環境において、既存手法の推定報酬に対する最適方策が、エキスパート方策と一致しない要因

表 2 の 3 行目と、4 行目の図中に赤い矢印がある。赤い矢印の状態 s は、推定報酬に対する最適方策 $\pi^*(s)$ とエ

キスパート方策 $\pi_E(s)$ が異なる。ここでは、状態遷移確率が異なる環境において、既存手法の推定報酬に対する最適方策と、エキスパート方策が異なる要因を考察する。

状態遷移確率が異なる環境において、既存手法の推定報酬に対する最適方策が、エキスパート方策と一致しない要因は、推定報酬が複数の環境の線形計画問題の制約条件を、同時に満たしていないことにある。各環境の状態遷移確率、エキスパート方策から構成される線形計画問題の制約条件は、「各環境においてエキスパート方策が最適となる」という制約である。表 2 の 3 行目と、4 行目の図に赤い矢印が存在することから、既存手法による推定報酬は、所与の環境の制約を満たすが、他の環境の制約条件を満たさない。つまり、既存手法による推定報酬は、二つの環境の線形計画問題の実行可能領域の差集合の要素であると言える。

6.3 複数の環境における推定報酬の平均値に対する最適方策が、エキスパートの方策と一致しない要因

表 2 の 5 行目は、複数の環境における推定報酬の平均値を用いた結果である。図中に赤い矢印があることから、推定報酬の平均値に対する最適方策 $\pi^*(s)$ とエキスパート方策 $\pi_E(s)$ が異なることがわかる。これは、既存手法による推定報酬の平均値が、報酬の推定に用いた環境の下での制約条件を満たしていないことを示す。一方、提案法は複数の環境の下での制約条件を満たす報酬を推定している。

6.4 提案法の推定報酬がエキスパートの報酬と一致するための必要条件を満たす要因

表 2 の 6 行目の図が示すように、提案法は、複数の環境のエキスパート方策が最適となる報酬を推定している。ここでは、提案法の推定報酬がエキスパートの報酬と一致するための必要条件を満たす理由を述べる。

4 章で述べたように、任意の状態遷移確率の環境において、エキスパートの報酬に対する最適方策は、エキスパート方策と一致する。そのため、複数の環境のエキスパート方策が推定報酬に対して最適となることは、推定報酬がエキスパートの報酬と一致する必要条件である。提案法は、各環境の線形計画問題の制約条件式 (7) を満たす実行可能領域の共通集合から報酬を推定する。したがって、提案法の線形計画問題に実行可能領域が存在するとき、推定報酬はエキスパートの報酬と一致する必要

条件を満たす。

6.5 提案法の推定報酬がスパースでない要因

表 2 の 1 列目の各手法の推定報酬を比較してわかるように、既存手法ではスパースな報酬が得られている一方、提案法の推定報酬はスパースでない。ここでは、提案手法の推定報酬がスパースでない要因について考察する。

提案手法の推定報酬がスパースでない要因は、L1 正則化のパラメータ λ の値、または提案法の定式化にあると考えられる。提案法、既存手法ともに L1 正則化のパラメータ λ の値が特定の値以上になると、最適解の報酬 R のノルム $\|R\|_1$ は 0 となる。提案手法の推定報酬がスパースでない要因が、L1 正則化のパラメータ λ の値の影響であるか明らかにするため、パラメータ λ の値を変化させ推定報酬のスパース性を確認した。提案法の L1 正則化のパラメータ λ の値を 0 から、 $\|R\|_1 = 0$ となるまで 0.01 刻みで変化させたが、スパースな報酬は得られなかった。よって、今回の実験設定において提案法の推定報酬がスパースでない要因は、提案法の定式化にあると言える。

本論文の実験においては、提案法の推定報酬は、既存手法の推定報酬と比較してスパースでなかった。しかし、任意の環境において、提案法と比較して既存手法の推定報酬が高いスパース性を備えるかは明らかでない。今後の課題として、理論的な考察や、様々な環境での実験を通した、提案法と既存手法の定式化の差異が推定報酬のスパース性に与える影響の解明が挙げられる。

7. ま と め

本論文では、「状態遷移確率の異なる複数の環境下で得られるエキスパート方策」と、推定報酬との無矛盾性を保証する逆強化学習手法を提案した。具体的には、状態遷移確率の異なる複数の環境下のエキスパートが持つ報酬を推定する問題を線形計画問題として定式化し、「状態遷移確率のみ異なる複数の環境のモデル」と「各環境で同じ報酬に従うエキスパート方策」から、各環境におけるエキスパート方策が最適となる報酬を推定する方法を示した。

提案法の貢献として、「状態遷移確率の異なる複数の環境下で得られるエキスパート方策」と、推定報酬との無矛盾性を保証したことを挙げる。風向き（状態遷移確率）が異なる複数の Windy grid world 環境を用いた実験を通じて、既存手法の推定報酬は、状態遷移確率が異なる環境におけるエキスパート方策と矛盾するが、提案法の推定報酬は、各環境におけるエキスパート方策に無矛盾であることを確認した。

提案法は、線形計画法を用いた逆強化学習 [Ng 00] に基づいた方法であるが、今後の課題として、他の代表的な逆強化学習の方法である、ペイジアン逆強化学習

(BIRL)[Ramachandran 07] や Maximum Entropy 逆強化学習 (MaxEnt IRL)[Ziebart 08] に基づいた「状態遷移確率が異なる環境間で無矛盾な報酬の推定法」を定式化することを挙げる。

◇ 参 考 文 献 ◇

- [Abbeel 04] Abbeel, P. and Ng, A. Y.: Apprenticeship learning via inverse reinforcement learning, in Proceedings of the 21th International Conference on Machine Learning, p. 1, ACM (2004)
- [Amin 17] Amin, K., Jiang, N., and Singh, S.: Repeated inverse reinforcement learning, in Advances in Neural Information Processing Systems, pp. 1815-1824 (2017)
- [Babes 11] Babes, M., Marivate, V., Subramanian, K., and Littman, M. L.: Apprenticeship learning about multiple intentions, in Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 897-904 (2011)
- [Finn 16] Finn, C., Levine, S., and Abbeel, P.: Guided cost learning: Deep inverse optimal control via policy optimization, in International Conference on Machine Learning, pp. 49-58 (2016)
- [Kitahara 13] Kitahara, T. and Mizuno, S.: A bound for the number of different basic solutions generated by the simplex method, Mathematical Programming, Vol. 137, No. 1-2, pp. 579-586 (2013)
- [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)
- [Kudrner 15] Kudrner, M., Gulati, S., and Burgard, W.: Learning driving styles for autonomous vehicles from demonstration, in Robotics and Automation (ICRA), pp. 2641-2646, IEEE (2015)
- [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)
- [Li 17] Li, K. and Burdick, J. W.: Meta inverse reinforcement learning via maximum reward sharing for human motion analysis, CoRR, Vol. abs/1710.03592, (2017)
- [Ng 00] Ng, A. Y., Russell, S. J., et al.: Algorithms for inverse reinforcement learning., in Proceedings of the 17th International Conference on Machine Learning, pp. 663-670 (2000)
- [Ramachandran 07] Ramachandran, D. and Amir, E.: Bayesian inverse reinforcement learning, Urbana, Vol. 51, No. 61801, pp. 1-4 (2007)
- [Surana 14] Surana, A. and Srivastava, K.: Bayesian nonparametric inverse reinforcement learning for switched Markov decision processes, in Machine Learning and Applications (ICMLA), 2014 13th International Conference on, pp. 47-54, IEEE (2014)
- [Sutton 98] Sutton, R. S. and Barto, A. G.: Reinforcement Learning: An Introduction, Vol. 1, MIT Press, Cambridge (1998)
- [Yamaguchi 18] Yamaguchi, S., Naoki, H., Ikeda, M., Tsukada, Y., Nakano, S., Mori, I., and Ishii, S.: Identification of Animal Behavioral Strategies by Inverse Reinforcement Learning, bioRxiv (2018)
- [Yang 18] Yang, J., Ye, X., Trivedi, R., Xu, H., and Zha, H.: Deep mean field games for learning optimal behavior policy of large populations, in International Conference on Learning Representations (2018)
- [Ziebart 08] Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K.: Maximum entropy inverse reinforcement learning, in Proceedings of the 23th AAAI Conference on Artificial Intelligence, Vol. 8, pp. 1433-1438, Chicago, IL, USA (2008)

〔担当委員：篠崎 隆志〕

2019 年 2 月 27 日 受理

著 者 紹 介

中田 勇介(正会員)

千葉大学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，模倣学習の研究に従事している。

荒井 幸代(正会員)

慶應義塾大学理工学部卒業後，ソニー（株），東京工業大学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。