逆強化学習によるヘテロジニアスなシステム間の報酬転移

○増山岳人 (中央大学)　梅田和昇 (中央大学)

1. はじめに

所望のタスクを達成するロボットの制御方策を，経験との相互作用から自律的に獲得する行動学習手法の一つに強化学習が挙げられる [1]．強化学習は，タスクを報酬関数によって記述し，ある制御方策にしたがうことで得られる報酬関数の出力値（報酬信号）の期待値を最大化する最適化の枠組みである．多くの場合，報酬関数はシステムの設計者によって手動で設計されるが，特にタスク，制御対象，環境が複雑な場合に適切な報酬関数を決定することは容易ではない．

そのため，タスクに関するエキスパートの演示を観測し，エキスパートのもつ制御則の背後にある報酬関数を推定する逆強化学習に関する研究がなされてきた [2, 3, 4]．近年では，モデルフリーの手法 [5, 6] 等も考案されており，その適用範囲は広がりつつある．一般的な逆強化学習では，演示を行うエキスパートと推定を行う主体であるエージェントの間で，報酬関数推定を行うための特徴空間は共有される．しかし，人の動作の観測に基づいてロボットの制御方策を獲得するようなタスクでは，身体や環境の違いから両者に共通の特徴空間を設定することが困難な場合がある．

そこで，本稿では特徴空間の明示的な対応関係を用いることなく，身体や外部環境の差異によりヘテロジニアスなエキスパートとエージェント間で報酬関数を転移するための逆強化学習の枠組みを提案する．報酬関数の推定には Relative Entropy Inverse Reinforcement Learning (REIRL) [6] を用いる．REIRL では，“feature matching” [3] に基づいて報酬関数が推定される．エキスパートとエージェントの特徴空間の対応関係が明示的に与えられない場合，feature matching を行うためには，観測される演示を何らかの方法でエージェントの特徴空間に写像する必要がある．本稿では，両者の対応を Least Squares Mutual Information（LSMI）[7] によって近似する．エキスパートとエージェントの状態間の対応点に LSMI を適用し，観測される演示から得られるエージェントの特徴量の事後分布から特徴期待値を算出する．シミュレーションにより，異なる特徴空間における演示から適切な方策を復元可能であることを確認する．

2. エージェントの特徴空間における報酬関数推定

エキスパート及びエージェントの状態をそれぞれ $s_E \in S_E$，$s \in S$ とする．S_E 及び S はエキスパート及びエージェントの状態空間である．本稿では，上記同様に下付き添字 E でエージェントに関する量を表記する．制御方策 $\pi(s,a) : S \times A \to [0,1]$ は状態 s において行動 $a \in A$ が選択される確率である．特徴量 $\boldsymbol{f}(s) : S \to \mathbb{R}^k$，$\pi$ についての特徴期待値を $\mu_i^\pi = E[\sum_{t=0}^{\infty} \gamma^t f_i(s_t,a_t) \mid s_0 = s, \pi]$，パラメータベクトルを $\boldsymbol{\theta} \in \mathbb{R}^k$ とする．ここで，報酬関数を $\boldsymbol{\theta}$ と $\boldsymbol{f}$ の線形結合 $r(s,a) := \sum_{i=1}^{k} \theta_i f_i(s,a)$ と定義すると，状態 s において方策 π にしたがったときの価値関数は $V^\pi(s) = \sum_{i=1}^{k} \theta_i \mu_i^\pi(s,a)$ となる．

REIRL では，任意の軌道 $\tau \in \mathcal{T}$，$\epsilon_i \in \mathbb{R}_+$ として，

$$|\sum_{\tau \in \mathcal{T}} p_\tau(\tau)\mu_i^\tau - \hat{\mu}_i| \leq \epsilon_i \tag{1}$$

なる制約条件の基で $\boldsymbol{\theta}$ を最適化することで報酬関数を推定する．ここで，$\boldsymbol{\mu}_i^\tau$ は τ に沿って観測される特徴期待値，$\hat{\boldsymbol{\mu}}$ はエキスパートの演示から算出される特徴期待値である．したがって，(1) は推定される報酬関数から得られる特徴期待値とエキスパートの方策から得られる特徴期待値を合致させるよう作用する．しかし，エキスパートとエージェントが身体や環境の違いによって特徴空間を共有していない場合，特徴期待値間の相違度を直接的に得ることはできない．

そこで，エキスパートとエージェントの状態空間の間の対応関係を，予め与えられた数点の特徴点の対応から推定する．エージェントの特徴を $\boldsymbol{x} : S \to \mathbb{R}^d$，エキスパートの特徴を $\boldsymbol{y} : S_E \to \mathbb{R}^{d_E}$ とする [1]．c 組の対応点 $C = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{c}$ を入力として LSMI を適用することで，

$$g(\boldsymbol{x},\boldsymbol{y}) = \frac{p_{xy}(\boldsymbol{x},\boldsymbol{y})}{p_x(\boldsymbol{x})p_y(\boldsymbol{y})} \tag{2}$$

が推定される．(2) より，

$$p_x(\boldsymbol{x}|\boldsymbol{y}) = g(\boldsymbol{x},\boldsymbol{y})p_x(\boldsymbol{x}) \tag{3}$$

である．ここで，$p_x(\boldsymbol{x})$ をエージェントに関する事前情報から決定すると，エキスパートの演示の基での $\boldsymbol{x}$ の事後確率が得られる．

演示から得られるエキスパートの特徴期待値の empirical estimate を $\mu_E(\boldsymbol{y}(s_E))$ とし，エージェントの特徴空間における演示の特徴期待値を以下で算出する．

$$\hat{\mu}(\boldsymbol{x}) := \sum_{\boldsymbol{y}} \mu_E(\boldsymbol{y}) g(\boldsymbol{x},\boldsymbol{y}) p_x(\boldsymbol{x}) \tag{4}$$

$\boldsymbol{y} \in Y_E$ は演示に対応する特徴集合である．エージェントの各状態 s に対応する $\boldsymbol{f}(s)$ について，$\hat{\boldsymbol{\mu}}$ を算出する．これにより，パラメータベクトル $\boldsymbol{\theta}$ を REIRL によって推定する．推定された報酬関数を用いて任意の強化学習手法を適用することで制御方策を復元する．

3. シミュレーション

異なる特徴空間をもつエキスパートとエージェントに対して提案手法を適用し，転移される報酬関数から学

[1] LSMI に入力する特徴量は REIRL による報酬関数の推定に用いる特徴量と同一である必要はないため，本稿では表記を区別する．

習される方策について検証を行うシミュレーションを行った．制御方策の算出には方策反復法を用いた．また，エキスパート及びエージェントには，それぞれ 2 リンク及び 3 リンクの平面マニピュレータを模したモデルを用いた．状態は各リンクの角度であり，8 分割した離散状態をとることとした．行動はリンク角度を $-\pi/8, 0, \pi/8$ のいずれか変化させるものとした．

3.1 実験設定

リンク長は全て 1 とした．REIRL に用いる基準方策は，すべての状態において各行動を等確率で選択するものとした．何らかの軌道をサンプルする際の軌道数とステップ数は，全てのシミュレーションを通してそれぞれ 5 本及び 5 ステップとした．また，サンプル軌道を得る際の初期状態は一様分布から生成した．割引率は 0.95，LSMI の基底関数は $\sigma = 1$ の RBF，$p_x(\boldsymbol{x})$ は一様分布とした．

エキスパートとエージェントの対応点は，それぞれの全リンク角度が $m\pi/8$ $(m = 0, \cdots, 7)$ となる 8 点に関する xy 平面座標のペアとした．

3.2 実験結果

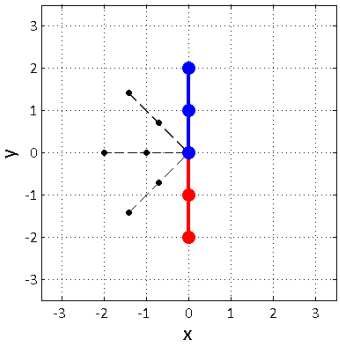
以下の 2 つの設定に関して，それぞれ 20 試行のシミュレーションを行った．エキスパートに与える報酬関数は目標姿勢 $\boldsymbol{\xi}_{E_d}$ において 1，それ以外では 0 とした．エキスパート及びエージェントの初期姿勢を $\boldsymbol{\xi}_{E_0}$，$\boldsymbol{\xi}_0$ とする．

1. $\boldsymbol{\xi}_{E_0} = (3\pi/2, 3\pi/2)$，$\boldsymbol{\xi}_0 = (3\pi/2, 3\pi/2, 3\pi/2)$，$\boldsymbol{\xi}_{E_d} = (\pi/2, \pi/2)$
2. $\boldsymbol{\xi}_{E_0} = (3\pi/2, 3\pi/2)$，$\boldsymbol{\xi}_0 = (3\pi/2, 3\pi/2, 3\pi/2)$，$\boldsymbol{\xi}_{E_d} = (\pi/2, 3\pi/4)$

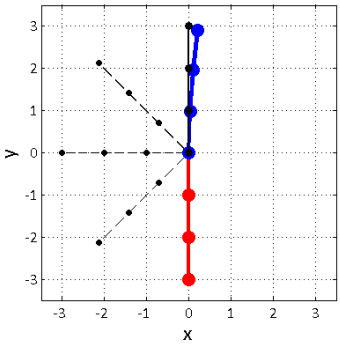
図 1，2 に結果を示す．図中，赤色は初期姿勢，青色は平均終端姿勢を表している．また，平均終端姿勢と xy 座標上で最も近い終端姿勢をもつ試行における状態遷移の例を黒色でプロットしている．

設定 1 の目標姿勢は対応点の 1 つと同一である．エキスパートの目標状態 $\boldsymbol{\xi}_{E_d}$ に対応する姿勢の近傍へ遷移する方策が獲得されていることが確認できる．設定 2 では，対応点ではない状態がエキスパートの目標姿勢となっている．図 2 より，自由度，身体のスケールが異なるにも関わらず，黒色の軌道例のように類似の姿勢を達成する方策が学習されていることが示唆される．しかし，非対応点を終端姿勢とする演示が与えられた場合，エージェントの終端姿勢はエキスパートの演示に依存してばらつく結果となった．これはエキスパートとエージェントの状態間の関係が汎化された結果，より大きなエントロピーをもつ報酬関数が推定されたためであると考えられる．軌道例で終端姿勢に到達するまでの時間ステップが最短でないことも同様の理由によるものである．

次に設定 1 について，適切なエキスパートの演示から得られる方策（REIRL），一様分布から生成した方策（ベースライン），及び提案手法から得られる方策のそれぞれについて，方策損失（policy loss [4]）を評価した．方策損失は，真の報酬関数を用いた場合に得られる方策と推定された方策の一致度を表す指標である．本稿では，価値関数ベクトルのノルムが 1 となるよう

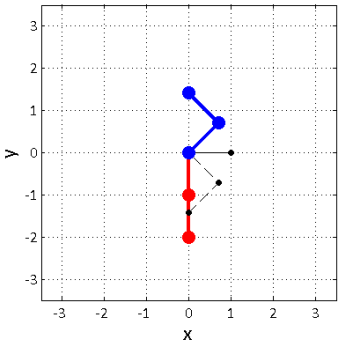


(a) エキスパート

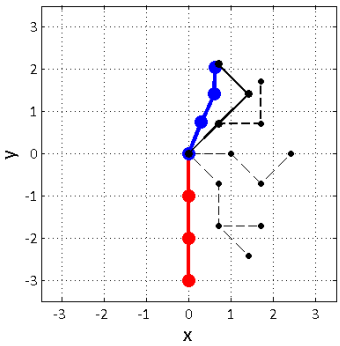


(b) エージェント

図 1: 設定 1



(a) エキスパート



(b) エージェント

図 2: 設定 2

表 1: 方策損失

	REIRL	ベースライン	提案手法
平均	0.00	1.1	0.17
標準偏差	0.00	0.17	0.41

正規化したため，方策損失は 0 から 2 の間の値をとる．真の報酬関数は，$\boldsymbol{\xi}_d = (\pi/2, \pi/2, \pi/2)$ において 1，それ以外の状態で 0 をとるものとした．REIRL では真の報酬関数から最適方策を学習，最適方策から得られたサンプル軌道から特徴期待値を算出した．表 1 に 20 試行の結果を示す．エージェントの環境におけるエキスパートの演示を用いた REIRL では，適切な方策が安定して復元可能であった．提案手法では，REIRL と比較すると若干の劣化はあるものの，おおよそ正しい方策が得られている．提案手法の大半の試行では適切な方策が得られていたものの，推定された方策によって収束する終端状態が真の報酬 0 となるものである場合があったため，大きな方策損失が生じる場合があった．その結果，提案手法の標準偏差はベースラインよりも大きな値をとっている．

4. まとめ

ヘテロジニアスなシステム間での逆強化学習による報酬推定を行う手法を提案した．提案手法では，エキスパートとエージェントの特徴の対応点を用いて確率密度比推定手法を適用し，エージェントの特徴空間における特徴期待値を算出する．算出された特徴期待値を用いて報酬関数を推定することで，自由度・スケールの異なる身体間での報酬関数の転移が可能であることを示した．

今後の課題として，実機による検証が挙げられる．本稿では対応点を用いて獲得された方策を評価したが，提案手法は一般的な模倣学習と異なり，アナロジカルなプロセスをもつものと捉えることができる．そのため，動作の効率性だけでなく，報酬関数を介したロボットによる他エージェントの行動推定にも取り組むことで，インタラクティブなシステムとしての評価を行いたい．

参 考 文 献

[1] M. Wiering, M. van Otterlo: “Reinforcement Learning: State-of-the-Art,” Springer-Verlag, 2012.
[2] A. Ng, S. Russell: “Algorithms for inverse reinforcement learning,” Proc. of the 17th Int. Conf. on Machine Learning, pp.663-670, 2000.
[3] P. Abbeel, A. Ng: “Apprenticeship learning via inverse reinforcement learning,” Proc. of the 21st ACM Int. Conf. on Machine Learning, 2004.
[4] D. Ramachandran, E. Amir: “Bayesian inverse reinforcement learning,” Proc. of the 20th Int. Joint Conf. on Artificial Intelligence, pp.2586-2591, 2007.
[5] E. Klein, B. Piot, M. Geist, O. Pietquin: “A cascaded supervised learning approach to inverse reinforcement learning,” Proc. of the European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2013.
[6] A. Boularias, J. Kober, J. Peters: “Relative entropy inverse reinforcement learning,” Proc of the 14th Int. Conf. on Artificial Intelligence and Statistics, pp.182-189, 2011.
[7] T. Suzuki, M. Sugiyama, T. Kanamori J. Sese: “Mutual information estimation reveals global associations between stimuli and biological processes,” BMC Bioinformatics, vol.10, no.1, pp.S52, 2009.