逆強化学習による学習者の選好を考慮した報酬関数の推定

○増山 岳人 (中央大学)　梅田和昇 (中央大学)

1. はじめに

自律適応的なロボットの制御システムを構築するための方法論の 1 つとして，これまで強化学習 [1] に関する研究が盛んに行われてきた．強化学習では将来に観測される報酬を最大化する方策を，試行錯誤的な環境との相互作用から獲得する．報酬とは，ロボットに与えられるタスクを記述するものである．例えば，ロボットに特定の位置姿勢をとらせたい場合は，所望の位置姿勢に対応する状態に対して高い報酬を設定する．各々の状態に対する報酬を規定する報酬関数は手動で設計されることが多いが，複雑なタスクや制御対象に対して適切な報酬関数を設計することは容易ではない．

この問題に対し，エキスパートの演示からエキスパートのもつ未知の報酬関数を推定する逆強化学習 [2] が提案されている．これまでにヘリコプタの自動運転 [3] 等の複雑なタスクへの応用がなされており，その有用性が示されている．この研究では，ヘリコプタを人間のエキスパートがコントローラで制御することにより教師データを得ているため，エキスパートと学習者で制御対象は同一である．他方，制御対象はロボットだが，教師データとなる特徴量の時系列はロボット以外の観測対象（例えば人間の動作）から得るという問題も考えられる．

このとき，学習者であるロボットは，観測対象とは異なる制約条件のもとで方策を獲得しなければならない場合がある．例えば，学習者が人の生活空間で運用される移動ロボットであれば，タスクとは無関係な人との間には安全性を担保するための距離や速度に関する制約条件が与えられる．このような制約条件は，我々人間が日常生活の中で感覚的に課している同様の制約と比較して，厳密に守られなければならないものとなるはずである．したがって，学習者固有の制約を考慮せずに，エキスパートの方策に忠実にしたがうよう推定された報酬関数をそのままロボットの強化学習に利用することは困難な場合がある．ロボットが方策を学習する際，ロボットのもつ制約に対応した報酬関数を推定された報酬関数に手動で加えればこの問題を解決することは可能である．しかし，制約条件を侵害しない限りは，元々の報酬関数にしたがって探索的な行動選択が行われる．そのため，エキスパートの模倣と制約条件を侵害するような探索行動の選択確率の低下を同時に実現する報酬関数を得ることは困難であり効率的ではない．

そこで，本稿では学習者のもつ固有の制約条件を考慮した報酬関数の推定を行う逆強化学習手法を提案する．以後，上述の学習者固有の制約条件を選好と呼ぶ．エキスパートの演示から得られる報酬関数と推定報酬関数の差に一定のマージンを設け，エキスパートの演示をどの程度忠実に模倣するかを決定する．与えられたマージンに応じて学習者の選好が報酬関数に反映されるよう制約条件を設定し，逆強化学習問題を解く．提案手法により，タスク実行に対する整合性を保ちつつ，学習者の選好を反映した報酬関数を獲得可能であることを 2 次元グリッドワールドにおけるシミュレーション実験により示す．

2. feature matching への学習者の選好の導入

本稿では Abbeel らの feature matching [4] に基いて報酬関数の推定を行う．まず，エキスパートのもつ真の報酬関数を $R_E(s) = w_E^T \phi(s)$，推定する報酬関数を $R(s) = w^T \phi(s)$ と定義する．ここで S は状態空間，$s \in S$ は状態，$\phi : S \mapsto [0,1]^k$ は特徴量，$w, w_E \in \mathbb{R}^k$ は重みベクトルである．エキスパートの演示 $\xi_i (i = 1, \cdots, m)$ を出力する方策 π_E が未知の R_E に対する最適方策であることを仮定し，π_E に近い性能をもつ方策 π を与える，特徴量から報酬関数への線形写像を求めることがここでの問題である．

ある状態 $s_0 \in S$ における方策 π の価値は割引率 $\gamma \in [0,1)$ とすると

$$V^\pi(s_0) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) | \pi\right] = w^T E\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t) | \pi\right] \quad (1)$$

式 (1) における割り引きされた特徴量の累積期待値を $\mu(\pi) := E[\sum_{t=0}^{\infty} \gamma^t \phi(s_t) | \pi]$ とする．なお，エキスパートの特徴期待値は $\hat{\mu}_E = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{\infty} \gamma^t \phi(s_t^{(i)})$ によって得るものとする．

Abbeel らの手法では，十分に小さな $\varepsilon \in \mathbb{R}_+$ に対して

$$\|w^T \mu_E - w^T \mu\| \leq \varepsilon \quad (2)$$

を満足する，エキスパートの演示を再現可能な方策の学習を目的としている．本稿では ε を学習者が自身の選好を報酬関数に反映させるためのマージンとみなし，以下では M で置き換える．M の値に応じてエキスパートの演示の再現と選好に対する優先度を変化させる．

学習者はエキスパートの演示から報酬関数を推定するために用いる特徴量 $\phi(s)$ に加えて，自身の方策を強化学習により学習する際に用いる特徴量 $\phi_r(s)$ 及び対応する特徴期待値 μ_r を観測することとする．学習者の選好を表す関数を $L\phi_r$ とする [1]．$L\phi_r$ は学習者固有の事前情報に応じて決定する．エキスパートは ϕ_r を観測，あるいは考慮せず ϕ のみによって演示を行っていると仮定すると，エキスパートの方策の再現と学習者の選好の両者を完全に満足することは多くの場合不

[1] 選好を表す関数には様々な形式が考えられるが，3. 節では $L \in \mathbb{R}^k$ としている．

可能である．そこで，学習者が得る報酬関数を新たに $w^T\phi(s) + L\phi_r(s)$ とし，以下のように制約条件を設定する．

$$w^T\mu_E - (w^T\mu + L\mu_r) \leq M \tag{3}$$

式 (3) の制約条件のもと逆強化学習問題を解くことで，学習者はマージン M によって許容される範囲で選好によってエキスパートの方策を修正した新たな方策を与える報酬関数を獲得する．

本稿では，以下の 2 次計画問題を解くことで報酬関数 $R(s) = w^T\phi$ を推定する．

$$\begin{aligned} &\underset{w}{\text{minimize}} \quad \|w\|_2 \\ &\text{subject to} \quad w^T\mu_E - M \geq \max_{\mu}(w^T\mu + L\mu_r) \end{aligned} \tag{4}$$

最適化には，Maximum Margin Planning（MMP）[5] を用いる．本稿では，MMP におけるスラック変数はマージン M によって置き換えている．以上の定式化により，学習者の選好を反映した報酬関数の推定を行う．

3. シミュレーション実験

提案手法による学習者の選好が，推定されるエキスパートの報酬関数に反映されることを確認するため，2 次元グリッドワールドにおけるシミュレーション実験を行った．

3.1 実験条件

状態空間は 32×32 のグリッドワールドとし，エージェントの行動は上下左右の 4 方向への 1 セルの移動とした．ただし，エージェントの行動は 30% の確率で失敗し，ランダムな行動が選択される．エキスパートの真の報酬関数は $[29, 32] \times [29, 32]$ の領域で 1，割引率 γ は 0.9 とした．エキスパートの演示は，真の報酬関数により得られる最適方策にしたがって，ランダムな初期状態から 10 ステップの行動選択を行い得られる状態系列とした．演示の数 m は 10 とした．

以下の 3 条件について実験を行った．

条件 1 図 2(a) の灰色領域を選好

条件 2 図 3(a) の灰色領域を選好

条件 3 図 4(a) の灰色領域を選好

$\phi \in \mathbb{R}^{|S|}$ は現在の状態において 1，それ以外の状態で 0 をとるベクトルとし，$\phi_r(s)$ はグリッドワールドの各状態に対して灰色領域では 1，白色領域では 0 の値をとるものとした．各条件で図中の白色領域内に対して 1 のコストを設定した．また，$M = 3$ である．

上記 3 条件を与えた提案手法及び MMP によって 10 回の試行を行った結果の平均を以下に示す．なお，比較のため報酬関数は全状態における報酬の 2 乗平均が状態数に一致するよう正規化した．

3.2 実験結果

各条件で推定された報酬関数を図 2(b)，3(b)，4(b) に示す．図 1 は MMP によって推定された報酬関数である．

図 2(b) では，報酬関数の最小点が白色領域内に移動し，外周を回りこむような勾配がみられる．図 3(b) では，白色領域の全体が低い報酬をもっていることが確認できる．図 4(b) では，全体的に状態空間の中心周りの時計回り方向に報酬が移動したような推定がなされた．図 1 の MMP によるフラットな推定報酬関数と比較して，それぞれの条件で灰色領域における報酬が高くなり，コストの設定された白色領域の報酬関数が低くなる傾向がみられた．同時に MMP による推定，すなわちエキスパートの方策に忠実な報酬関数と概形は

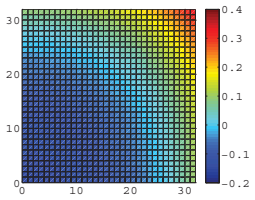
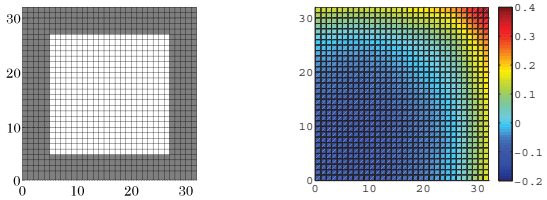
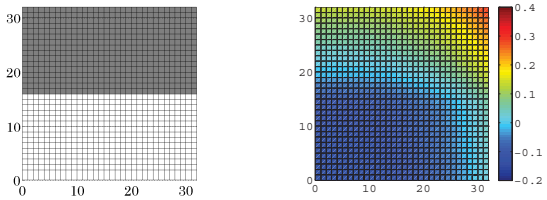


図 1 報酬関数（MMP）



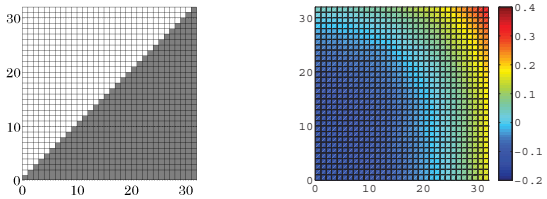
(a) コストマップ（条件 1）　(b) 報酬関数（条件 1）

図 2 条件 1



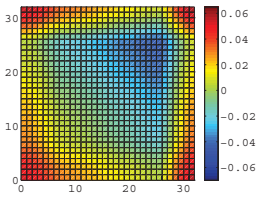
(a) コストマップ（条件 2）　(b) 報酬関数（条件 2）

図 3 条件 2

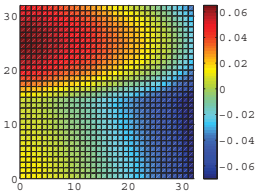


(a) コストマップ（条件 3）　(b) 報酬関数（条件 3）

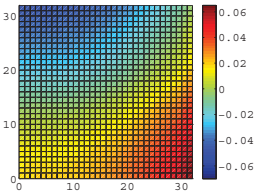
図 4 条件 3



(a) 条件 1



(b) 条件 2



(c) 条件 3

図 5MMP 及び提案手法の推定報酬関数の差

	条件 1	条件 2	条件 3
灰色領域	0.2107	0.1458	0.1885
白色領域	-0.2107	-0.1626	-0.2007

表 1 各条件及び領域毎の報酬関数の差

一致している．

図 5 は提案手法及び MMP で得られた報酬関数の差を各条件について図示したものである．図 2(a)，3(a)，4(a) の灰色領域で相対的に高い値がとられていることがみてとれる．真の報酬関数は図中右上の領域（目標領域）において高い報酬を与えるものである．エキスパートは真の報酬関数に対する最適方策をもつため，目標領域へ向かうような行動を選択する．図 3(a) 中，左下の領域の報酬が増加しているのは，左下の領域から目標領域へ向かう経路の途中で自身の選好をより長く満足できる可能性が高いためである．逆に目標領域における報酬が減少傾向にあるのは，エキスパートの演示から推定される報酬が十分に高く，マージン内での選好とのトレードオフにより，目標領域外での報酬が増加した結果である．

表 1 は図 5 の報酬関数の差に関して灰色領域，白色領域のそれぞれで平均をとったものである．各条件で灰色領域における報酬が白色領域における報酬より大きく，設定した学習者の選好をエキスパートの演示から推定される報酬に反映できていることが確認できる．

4. まとめ

強化学習問題における報酬関数をエキスパートの演示から推定する逆強化学習を利用して，学習者の選好を反映した報酬関数の推定を行う手法を提案した．提案手法では feature matching の枠組みに，エキスパートの方策の再現を図る特徴量とは独立した特徴量により表現される，学習者固有の選好を取り入れ報酬関数を推定する．2 次元グリッドワールドにおけるシミュレーションにより，エキスパートの方策を模倣すると同時に一定のマージンにしたがって学習者の選好を報酬関数に反映可能であることを示した．

これまでに，エキスパートの演示をクラスタリングし，各クラスタについて逆強化学習問題を解く手法が提案されている [6, 7]．今後の課題としては，学習者の選好と整合する演示を抽出することで，エキスパートの方策と矛盾の少ない報酬関数の推定手法を構築する．また，特徴量に関する線形写像で報酬関数を表現しているため，非線形な報酬関数への対応も今後の課題である．

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”, Proceedings of the 17th International Conference on Machine Learning, pp.663–670, 2000.

[3] P. Abbeel, A. Coates M. Quigley, A. Ng: “An Application of Reinforcement Learning to Aerobatic Helicopter Flight”, Advances in Neural Information Processing Systems, 2007.

[4] P. Abbeel, A. Ng: “Apprenticeship Learning via Inverse Reinforcement Learning”, Proceedings of the 21st International Conference on Machine Learning, pp.1–8, 2004.

[5] N. Ratliff, J. Bagnell, M. Zinkevich: “Maximum Margin Planning”, Proceedings of the 23rd International Conference on Machine Learning, pp.729–736, 2006.

[6] M. Babes, V. Marivate, K. Subramanian, M. Littman: “Apprenticeship Learning About Multiple Intentions”, Proceedings of the 28th International Conference on Machine Learning, pp.897–904, 2011.

[7] J. Choi, K. Kim: “Nonparametric Bayesian Inverse Reinforcement Learning for Multiple Reward Functions”, Advances in Neural Information Processing Systems, pp.305–313, 2012.