スコアに基づく逆強化学習のための動的計画法による軌道の自己生成

Self-generation of trajectories via dynamic programming for score-based inverse reinforcement learning

○ 渡邉 夏美（中央大） 増山 岳人（中央大） 正 梅田 和昇（中央大）

Natsumi WATANABE, Chuo University
Gakuto MASUYAMA, Chuo University
Kazunori UMEDA, Chuo University

This paper presents a generation method of trajectories for score-based inverse reinforcement learning. While most inverse reinforcement learning methods require demonstrations of the expert, score-based inverse reinforcement learning can estimate the expert's reward function from scores of arbitrary trajectories. This study brings active learning into score-based inverse reinforcement learning to estimate a reward function from few trajectories. An agent generates informative trajectories for reward estimation and pose them as queries to the expert. The proposed method generates the trajectories using expectation of discounted accrued features which is calculated by dynamic programming. The informativeness of the trajectories is evaluated by criteria for queries. Simulation results in a cart-pole domain demonstrate that the proposed method efficiently estimates the reward function from few trajectories.

Key Words: Reinforcement learning, Inverse reinforcement learning, Active learning

1 緒言

未知環境下でのロボットの自律的な行動学習の実現を目的とし，強化学習に関する研究が盛んに行われている [1]．強化学習では，ロボットなどの学習者をエージェントと呼び，エージェントは環境との相互作用から行動の選択基準となる方策を学習する．選択された行動に対する評価は報酬と呼ばれる数値信号によって与えられる．したがって，タスクを表現する報酬によって，エージェントの獲得する方策の性能は変化する．しかし，多様で複雑な環境における報酬の設定は人手では困難であり，報酬の設定が適切でなかった場合ロボットによるタスクの達成は不可能である．

逆強化学習 [2, 3, 4, 5] では他者の演示からその振る舞いを説明する報酬関数を推定するため，人手による報酬の設定が不要となる．逆強化学習では，目的のタスクにおける熟練者であるエキスパートの演示から報酬関数を推定する．推定した報酬関数を用いて強化学習を行うことで，人手により報酬を設定することなく方策を学習することができる．しかし，適切な報酬関数を推定するためにしばしば多数の演示が要求される．

Daniel らが提案する逆強化学習 [6] では，ロボットが自己生成した軌道に対するスコアをエキスパートに要求しそのスコアから報酬関数を推定する．そのため，一般的な逆強化学習と異なりエキスパートの演示が不要である．この逆強化学習では方策と報酬関数の学習を並行して行っており，方策に基づき軌道を生成する．しかし，方策更新のための強化学習の反復と報酬関数の GP モデルの学習が必要であり，ロボットの制御問題への適用は容易ではない．

Burchfiel らが提案する逆強化学習 [7] は，Daniel らの逆強化学習とは異なり，任意の軌道のスコアから報酬関数を推定する．したがって，用いる軌道の選択が適切であれば，少数の軌道からの効率的な報酬関数の推定が可能であると考えられる．本稿では，Burchfiel らの逆強化学習に能動学習を導入し，効率的な報酬関数推定を可能とする軌道の自己生成手法を提案する．報酬関数の推定に対する寄与度が大きい軌道をエージェントが自己生成することで，推定に要する軌道の数を削減する．

2 自己生成した軌道のスコアに基づく逆強化学習

2.1 軌道のスコアに基づく逆強化学習

Burchfiel らの軌道のスコアに基づく逆強化学習（Distance Minimization Inverse Reinforcement Learning; DM-IRL）[7] では，任意の軌道に付与されるスコアからエキスパートの報酬関数を推定する．軌道に対する評価であるスコアのみをエキスパートに要求し，最適な演示を不要とする．

まず，エージェントがとり得る状態を $s \in S$，とり得る行動を $a \in A$ とする．本稿では，方策は決定論的であるとし，方策 $\pi : S \mapsto A$ はエージェントが状態 s において選択する行動を示す．$\boldsymbol{\phi} : S \mapsto \mathbb{R}^k$ を状態 s においてエージェントが観測する特徴量とする．エキスパートの報酬関数 $R : S \mapsto \mathbb{R}$ は $R(s) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(s)$ で表される．したがって，ここでの報酬関数の推定は重み $\boldsymbol{w} \in \mathbb{R}^k$ の推定と等価である．

軌道は状態列であり，軌道 $\tau_i = \{s_0^i, \cdots, s_{|\tau_i|-1}^i\}$ の累積特徴量 $\boldsymbol{\psi}(\tau_i)$ を式 (1) で定義する．

$$\boldsymbol{\psi}(\tau_i) = \sum_{t=0}^{|\tau_i|-1} \gamma^t \boldsymbol{\phi}(s_t^i) \tag{1}$$

ただし，$\gamma \in [0, 1)$ は割引率である．

累積特徴量を用いて，軌道 τ_i のスコアを式 (2) で表す．

$$\begin{aligned} v_i &= \sum_{t=0}^{|\tau_i|-1} \gamma^t \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(s_t^i) \\ &= \boldsymbol{w}^{\mathrm{T}} \sum_{t=0}^{|\tau_i|-1} \gamma^t \boldsymbol{\phi}(s_t^i) \\ &= \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\psi}(\tau_i) \end{aligned} \tag{2}$$

r 本の軌道とそのスコアが与えられたとき，式 (3) の線形回帰により重みを推定する．

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \|M\boldsymbol{w} - \boldsymbol{v}\| \tag{3}$$

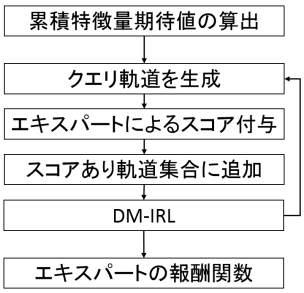


Fig.1 Flow of estimation

ただし，$\boldsymbol{v} \in \mathbb{R}^r$ は軌道のスコア $v_1, \cdots, v_r$ を要素とするベクトル，$M \in \mathbb{R}^{r \times k}$ は軌道 $\tau_1, \cdots, \tau_r$ の累積特徴量からなる式 (4) の行列である．

$$M = \begin{pmatrix} \boldsymbol{\psi}^{\mathrm{T}}(\tau_1) \\ \vdots \\ \boldsymbol{\psi}^{\mathrm{T}}(\tau_r) \end{pmatrix} \tag{4}$$

最小二乗法による線形回帰を行う場合，M のランク数が特徴量の次元数以上である必要がある．そこで，M のランク数が特徴量の次元数に満たない場合には LASSO 回帰を用いることで，軌道の本数が少数である場合にも線形回帰による推定を可能とする．

推定した重み $\hat{\boldsymbol{w}}$ を用いることで，任意の軌道のスコアを推定することができる．

2.2 エージェントによる軌道の自己生成

DM-IRL では，複数の軌道とそれらに対してエキスパートが付与するスコアから推定を行う．用いる軌道はエキスパートによる演示軌道である必要がなく，任意のものでよい．しかし，とり得る軌道の報酬関数推定に対する寄与度は一様でないため，寄与度が小さい軌道ばかりを用いた場合，軌道の本数が増えても推定精度は向上しない．

少数の軌道からの効率的な推定を可能にする方法として，能動学習の導入が考えられる．能動学習は，教師あり学習において必要なラベルありデータの数の削減を目的とした手法である [8]．エージェントによって推定に対する寄与度が大きい軌道を自己生成することができれば，少数の軌道からの報酬関数の推定が可能となる．

本稿で提案する能動学習では，推定に対する寄与度が大きい軌道をエージェントが自己生成し，エキスパートに対するクエリとする．ここでのクエリとは，エキスパートに対するスコアの問い合わせである．以降，エージェントにより生成される軌道をクエリ軌道と呼ぶ．軌道の生成には，動的計画法により算出した累積特徴量期待値を用い，推定に対する寄与度が大きい軌道を生成するためにクエリ軌道生成基準を設ける．

本手法を導入した DM-IRL による報酬関数推定の流れを図 1 に示す．

3 軌道の自己生成手法

3.1 累積特徴量期待値を用いた軌道生成

DM-IRL は軌道の累積特徴量を用いて推定を行うため，推定に対する軌道の寄与度はその軌道の累積特徴量に従う．よって，効率的な推定を行うには，推定に対する寄与度が大きい累積特徴量をもつ軌道が生成できればよい．

本稿では，所望の累積特徴量をもつ軌道を生成するために，推定開始前の処理として累積特徴量の期待値を算出する．各状態における，その状態が初期状態である軌道がとる累積特徴量の期待値は，軌道を通して選択される行動は一様分布に従うとして式 (5) で表される．

$$\hat{\boldsymbol{\psi}}(s) = \sum_{t=0}^{\infty} \gamma^t \underset{a}{E}[\boldsymbol{\phi}(s_t) | s_0 = s] \tag{5}$$

累積特徴量期待値の算出は，後述する動的計画法を用いて行う．

初期状態から式 (6) で表される方策に従い行動を選択することで，累積特徴量 $\bar{\boldsymbol{\psi}}$ をもつクエリ軌道を生成する．

$$\pi(s) = \underset{a}{\operatorname{argmin}} \underset{s' \sim s,a}{E} \left[C_{\bar{\psi}} \left(\hat{\boldsymbol{\psi}}(s') \right) \right] \tag{6}$$

ここで，$C : \mathbb{R}^k \mapsto \mathbb{R}$ は後述するクエリ軌道生成基準である．

3.2 動的計画法による累積特徴量期待値の算出

累積特徴量の期待値の算出には動的計画法を用いる．動的計画法とは，計算量の膨大な問題を部分問題の計算結果の利用により解く手法である．強化学習の解法としても動的計画法が用いられており，方策反復や価値反復などがこれにあたる [9]．

本稿では，価値反復を応用して累積特徴量期待値の算出を行う．まず，累積特徴量期待値 $\hat{\boldsymbol{\psi}}(s)$ を全ての状態について任意の値で初期化する．軌道を通して選択される行動は一様分布に従うとして，各状態について式 (7) で更新する．

$$\hat{\boldsymbol{\psi}}(s) \leftarrow \underset{a}{E}[\boldsymbol{\phi}(s') + \gamma \hat{\boldsymbol{\psi}}(s')] \tag{7}$$

更新前後の変化量が十分小さくなるまで更新を繰り返し，最終的に得られる $\hat{\boldsymbol{\psi}}(s)$ を用いて軌道の自己生成を行う．

4 クエリ軌道生成基準

DM-IRL の目的は任意の軌道に対するエキスパートのスコアの推定であり，推定スコアとエキスパートによる真のスコアとの誤差の最小化といえる．よって，報酬関数の推定に対する寄与度が最大である軌道とは，スコア推定誤差を最小化する軌道である．

このような軌道を生成するために，本稿では 3 つのクエリ軌道生成基準 C を設ける．推定開始時のクエリ軌道生成には全特徴量の観測を条件とするクエリ軌道生成基準を用い，その後は推定スコア最大のクエリ軌道とスコア推定誤差に基づくクエリ軌道の 2 つを用いる．

4.1 全特徴量の観測

線形回帰により重みを推定する際，全ての軌道で観測されなかった特徴量は無視される．そのため，その特徴量が重要なものであった場合，適切な報酬関数の推定は困難である．

そこで，学習開始時のクエリ軌道には全特徴量がいずれかの軌道で必ず観測される複数の軌道を用いる．学習に用いる軌道のうち少なくとも 1 本の軌道で観測されれば，その特徴量は以降の報酬関数推定で必ず考慮される．学習開始時に全特徴量の観測を条件とするクエリ軌道生成基準を設けることで，特徴量の無視を回避する．

ある特徴量 ϕ_j を観測する軌道を生成する場合，初期状態から式 (8) で表される方策に従い行動を選択し，それにより得られる状態列を軌道とする．

$$\pi(s) = \underset{a}{\operatorname{argmin}} \underset{s' \sim s,a}{E} [|\hat{\psi}_j(s') - \bar{\psi}_j|] \tag{8}$$

$\bar{\psi}_j$ の値は，$\bar{\boldsymbol{\psi}}$ が ϕ_j を観測する軌道の累積特徴量に相当するよう設定する．

全特徴量が観測されたら，次節以降の推定スコア最大のクエリ軌道の生成とスコア推定誤差に基づいたクエリ軌道生成を行う．

4.2 推定スコア最大のクエリ軌道

DM-IRL においては，タスクに有用である軌道ほど，エキスパートから高いスコアが与えられる．高いスコアをもつ軌道に関するスコア推定誤差の最小化は，目的のタスクの達成に有効であり，任意の軌道に対するスコア推定精度の向上にも効率的であると考えられる．

そこで，全特徴量観測後のクエリ軌道生成基準の 1 つを，推定した報酬関数から得られる推定スコアが最大である軌道の生成とする．軌道生成の際には，推定した重み $\hat{\boldsymbol{w}}$ を用いて，初期状態から式 (9) で表される方策に従い行動を選択し，それにより得られる状態列を軌道とする．

$$\pi(s) = \underset{a}{\operatorname{argmax}} \underset{s' \sim s,a}{E} [\hat{\boldsymbol{w}}^{\mathrm{T}} \hat{\boldsymbol{\psi}}(s')] \tag{9}$$

Table 1 Terminal states

	最小値	最大値
カートの位置 x[m]	-2.4	2.4
振子の角度 θ[°]	-12	12

Table 2 Features and true weights

$\boldsymbol{\phi}$	$\boldsymbol{w}^*$
$-\epsilon_x \leq x \wedge x \leq \epsilon_x$	1
$-\epsilon_{\dot{x}} \leq \dot{x} \wedge \dot{x} \leq \epsilon_{\dot{x}}$	0
$-\epsilon_\theta \leq \theta \wedge \theta \leq \epsilon_\theta$	1
$-\epsilon_{\dot{\theta}} \leq \dot{\theta} \wedge \dot{\theta} \leq \epsilon_{\dot{\theta}}$	0
$x\dot{x} > 0$	-1
$x\dot{x} < 0$	1
$\theta\dot{\theta} > 0$	-1
$\theta\dot{\theta} < 0$	1

4.3 スコア推定誤差に基づく軌道生成

スコアあり軌道集合のうち，推定された報酬関数を用いて表される推定スコアとエキスパートにより付与された真のスコアとの誤差が大きい軌道に着目する．このような軌道に対するスコア推定に必要な情報は相対的に不足しているといえる．したがって，類似の累積特徴量をもつ軌道をクエリ軌道とすることで，効率的なスコア推定精度の向上が期待できる．そこで，スコア推定誤差に基づくクエリ軌道生成基準を設けることで不足している情報を獲得する．

スコア推定誤差に基づくクエリ軌道生成基準では，回帰により得た重みを用いた推定スコアと実際に付与されたスコアとの誤差が最も大きい軌道に類似した軌道を生成する．軌道生成の際は，初期状態から式 (10) で表される方策に従い行動を選択し，それにより得られる状態列をクエリ軌道とする．ただし，$\bar{\boldsymbol{\psi}}$ はスコア推定誤差が最も大きい軌道に類似した累積特徴量である．

$$\pi(s) = \underset{a}{\operatorname{argmin}} \underset{s' \sim s,a}{E} [\|\hat{\boldsymbol{\psi}}(s') - \bar{\boldsymbol{\psi}}\|_2] \tag{10}$$

5 シミュレーション

提案手法の有用性を検証するため，倒立振子制御問題を用いたシミュレーションを行った．シミュレーションで用いる環境設定は，文献 [10] に述べられているものを使用している．

5.1 シミュレーションの設定

状態は，カートの位置 x とその速度 $\dot{x}$，振子の角度 θ とその角速度 $\dot{\theta}$ から与えられ，それぞれを 7 つの領域に分割することで離散化する．終端状態は，x または θ の表 1 に示す最大値または最小値への到達とする．選択され得る行動は，右方向，左方向へのカートの加速である．

特徴量は 8 つの命題の真理値を用い，それぞれに重みを設定する．表 2 に特徴量 ϕ と設定した真の重み w^* の値を示す．ただし，ϵ はそれぞれの観測値に関して設定した微小な値である．軌道に付与するスコアは，設定した重みを用いて式 (2) により算出する．用いる割引率は $\gamma = 0.99$ とする．

5.2 シミュレーション条件

クエリ軌道本数を 30 本とし，以下の条件でシミュレーションを行った．ただし，全ての条件において学習開始時は全特徴量の観測を条件とするクエリ軌道生成基準に従い軌道を生成しクエリとした．全特徴量観測後のクエリ軌道として，

(a) 推定スコア最大の軌道を生成．

(b) スコア推定誤差に基づく軌道を生成．

(c) スコアあり軌道集合のスコア推定誤差の最大値が 1 以下であった場合は推定スコア最大の軌道を生成し，その他の場合にはスコア推定誤差に基づく軌道を生成．

Table 3 Scores of test data

最小値	最大値	平均
1.0	112.0	39.5

また，クエリ軌道を全てランダムな軌道とした場合を比較手法としてシミュレーションを行った．

評価用データとしてランダムな軌道とそのスコアの組を 100 個用意し，これらに対するスコア推定誤差を用いて評価を行った．評価用データのスコアの最小値，最大値，平均を表 3 に示す．

それぞれ 30 回の試行を行い，その平均を用いて評価した．

5.3 シミュレーション結果

(a), (b), (c) における学習過程全体に渡るスコア推定誤差の平均と標準偏差の変化を図 2 に示す．また，軌道本数 15 本から 30 本での (a), (b), (c) のスコア推定誤差の平均と標準偏差の変化を図 3 に示し，軌道本数 30 本でのスコア推定誤差の平均と標準偏差を表 4 に示す．ただし，図 2，図 3 において，実線と点線はスコア推定誤差の平均を，色塗りされた領域は標準偏差を表す．ただし，図 3 では平滑化を行った結果を示している．

全ての条件において，比較手法に比べ誤差の平均，標準偏差ともに大きく減少した．軌道本数約 20 本から各条件における結果に違いがみられた．推定スコア最大の軌道をクエリとした場合では 20 本以降での標準偏差の減少が小さいが，スコア推定誤差に基づくクエリ軌道を用いた場合は 30 本になるまで標準偏差の減少がみられる．また，推定スコア最大のクエリ軌道とスコア推定誤差に基づくクエリ軌道を組み合わせた場合では，軌道本数約 10 本から標準偏差の減少がみられ，30 本では誤差の平均と標準偏差ともに他の条件に比べ最小となる結果が得られた．

5.4 考察

シミュレーション結果から，クエリ軌道生成基準を用いることで少数のスコアあり軌道から精度よく任意の軌道のスコアが推定可能となることが確認できた．

推定スコア最大のクエリ軌道を用いた場合では標準偏差の減少が小さくなったが，これは特徴空間におけるクエリ軌道の偏りによるものだと考えられる．推定スコアが最大となる軌道をクエリとしているため，クエリ軌道から観測される特徴量に類似性が生じる．よって，クエリ軌道に対する類似性をもつ軌道に対しては精度のよいスコア推定が可能だが，類似しない軌道に対しては推定誤差が大きくなると考えられる．また，スコア推定誤差に基づくクエリ軌道のみを用いた場合，スコアあり軌道集合のスコア推定誤差が小さいときには，情報が不足した軌道をクエリとすることができず，効率的なクエリ軌道の生成が困難となる．よって，推定スコア最大のクエリ軌道とスコア推定誤差に基づくクエリ軌道を取り入れた手法ではより精度のよい報酬関数の推定が可能となると考えられる．

また，このシミュレーションでは特徴量は 8 次元であったため，全特徴量の観測を条件とするクエリ軌道生成基準の適用が容易であったが，特徴量次元が高次である場合の適用は困難である．よって，クエリ軌道生成基準の改良やより汎用的な基準の採用が必要である．

また，軌道生成に必要な累積特徴量の期待値は行動が一様分布であるとして算出したが，行動の分布が適切であるかについての評価は行っていない．行動の選択方法に検討の余地がある．

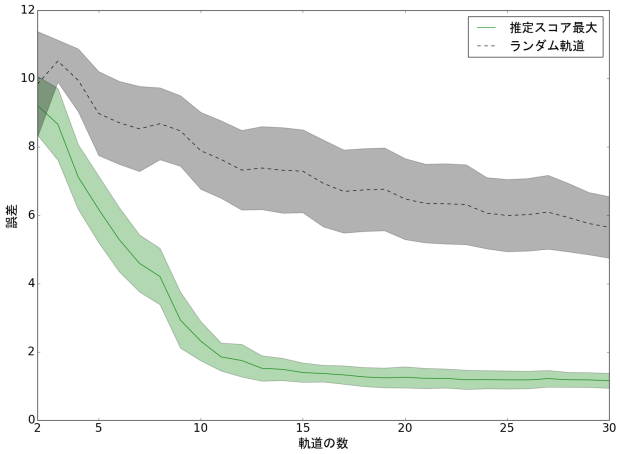
6 結言

本稿では，軌道のスコアに基づく逆強化学習に能動学習を導入し，効率的な報酬関数推定を可能とする軌道の自己生成手法を提案した．また，倒立振子制御問題を用いたシミュレーションを行い，提案手法により少数の軌道からの報酬関数推定が可能となることを確認した．

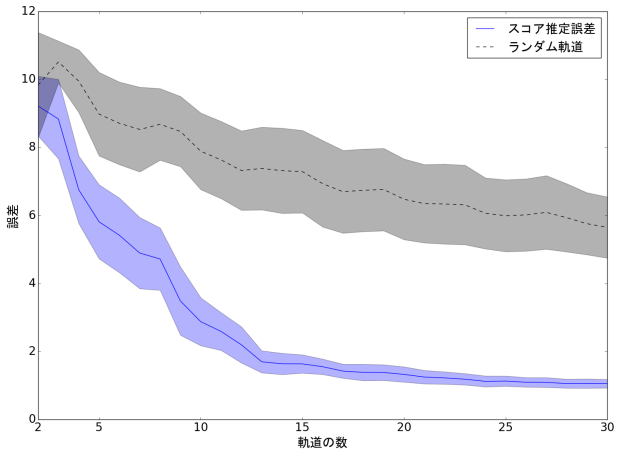
今後の展望として，特徴量が高次である問題への適用が挙げられる．これには，累積特徴量期待値の計算量削減や新たなクエリ軌道生成基準の検討が必要であると考えられる．

謝辞

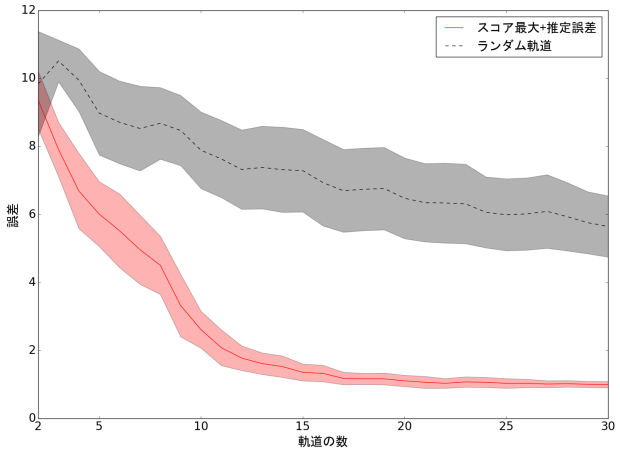
本研究は JSPS 科研費 16K16132 の助成を受けたものです．



(a) Maximum estimate



(b) Maximum error



(c) Maximum estimate/error

Fig.2 Mean and standard deviation of error of estimated score

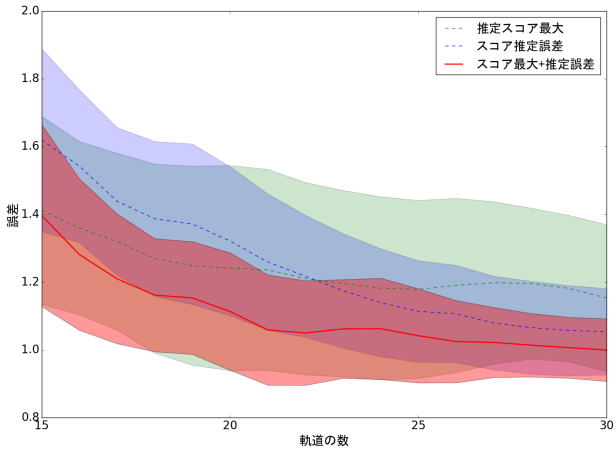


Fig.3 Error of estimated score（15-30 trajectories）

Table 4 Error of estimated score after 30 queries

クエリ軌道	平均	標準偏差
推定スコア最大	1.156	0.435
スコア推定誤差	1.054	0.252
スコア最大＋推定誤差	0.998	0.183

参考文献

[1] J. Kober, J. A. Bagnell and J. Peters, “Reinforcement Learning in Robotics: A Survey,” the 17th International Journal of Robotic Reserch, vol.32, no.11, pp.1238-1274, 2013.

[2] A. Y. Ng and S. Russell, “Algorithms for Inverse Reinforcement Learning,” the 17th International Confernce on Machine Learning, pp.663-670, 2000.

[3] P. Abbeel and A. Y. Ng, “Apprenticeship Learning via Inverse Reinforcement Learning,” the 21st International Confernce on Machine Learning, pp.1-8, 2004.

[4] B. D. Ziebart, A. Maas, J. A. Bagnell and A. K. Dey, “Maximum Entropy Inverse Reinforcement Learning,” the 23rd AAAI Conference on Artificial Intelligence, pp.1433-1438, 2008.

[5] A. Boularias, J. Kober and J. Peters, “Relative Entropy Inverse Reinforcement Learning,” the 14th International Conference on Artificial Intelligence and Statistics, vol.15, pp.20-27, 2011.

[6] C. Daniel, M. Viering, J. Metz, O. Kroemer and J. Peters, “Active Reward Learning,” Robotics: Science and Systems, 2014.

[7] B. Burchfiel, C. Tomasi and R. Parr, “Distance Minimization for Reward Learning from Scored Trajectries,” the 30th AAAI Conference on Artificial Intelligence, pp. 1-7, AAAI Press, 2016.

[8] B. Settles, “Active Learning Literature Survey,” Computer Sciences Technical Report 1648, University of WisconsinMadison, 2010.

[9] R. S. Sutton and A. G. Barto, “Reinforcement Learning: An Introduction,” MIT Press, Cambridge, Massachusetts, 1998.

[10] A. G. Barto, R. S. Sutton and C. W. Anderson, “Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Ploblems,” IEEE Transactions on System, Man, and Cybernetics, vol.13, no.5, pp.834-846, 1983.