The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

複数の報酬関数を推定可能なタスク条件付き敵対的模倣学習

Task-Conditional Generative Adversarial Imitation Learning That Infers Multiple Reward Functions

小林 京一郎 *1 Kyoichiro Kobayashi　堀井 隆斗 *2*3 Takato Horii　岩城 諒 *1 Ryo Iwaki　長井 志江 *3 Yukie Nagai　浅田 稔 *1 Minoru Asada

*1大阪大学 Osaka University　*2電気通信大学 The University of Electro-Communications　*3情報通信研究機構 National Institute of Information and Communications Technology

In this work, we propose a new framework of imitation learning that is designed to infer the multiple reward functions. We introduce latent variables to discriminator and generator in Generative Adversarial Imitation Learning (GAIL) to learn different reward functions and policies for different tasks. In order to control the balance between imitate expert directly (early convergence) and to enhance variance of policy (sample various data and learning robust reward), we introduce entropy regularized correction term in generator's objective function. We guarantee that the objective function has the unique optimal solution by the same discussion as GAIL. In the experiment at the grid world problem, we show that our framework can infer multiple reward functions and policies that represent different tasks efficiently.

1. はじめに

模倣学習は，教示者の行動データを所与として，学習主体が自らの行動を獲得する学習手法である． この手法は， 教示者の行動から環境の報酬関数を推定する逆強化学習問題（Inverse Reinforcement Learning, IRL）と，報酬から得られる収益の期待値を最大化する方策を学習する強化学習問題（Reinforcement Learning, RL）として扱う事ができる [Russell 98]．近年，逆強化学習と強化学習の合成問題を一つの目的関数からなる最適化問題として定式化した Generative Adversarial Imitation Leranig (GAIL) [Ho 16] が提案された．GAIL は報酬関数の分布の形状を仮定しないため，IRL と RL を用いた従来の模倣学習手法の問題点である報酬関数のパラメータ更新における尤度の計算量を解決できる学習則である．一方， 模倣学習において，環境のダイナミクスを共有している教示者の様々な行動に対する複数の報酬関数を推定することで， 異なる目的を達成するタスクを単一の学習器で表現可能であることが期待できる．GAIL は単一の報酬関数を想定しているため， そのまま複数タスクの模倣への適用はできないが，GAIL に対して，多様な動作を学習させるためのアプローチとして方策に対して潜在変数を導入した学習則が提案されている [Li 17, Hausman 17]．この手法はある一つのタスクを達成するための多様な手段を学習可能であるが，報酬関数に対してタスク条件付けがなされていないため，目的が異なるタスクの学習に適用することは困難である．

そこで， 本研究では GAIL を拡張し， 複数の報酬関数の推定と模倣を同時に行うアルゴリズムを提案する． 提案モデルは， GAIL における生成器と識別器の両方に， 潜在変数による条件付け可能な構造を持たせ，複数の報酬関数と方策を同時に推定する．さらに，GAIL と同様の議論によって提案する目的関数が唯一の最適解を持つことを示す．また， 生成器の学習則に対してエントロピー正則化の係数補正項を導入することで， 学習速度と獲得する方策の性能の向上を図る．本研究では， グリッドワールド内の異なる座標へ到達するような， 複数の報酬関数を仮定した教示者の行動を模倣する実験において，提案モデルが異なる報酬関数を同時に推定し，各報酬関数に対して方策の学習が可能であることを示す．

連絡先: 小林京一郎, kyoichiro.kobayashi@ams.eng.osaka-u.ac.jp

2. 関連研究

2.1 単一の報酬関数および方策に対する模倣学習則

本研究では， 強化学習と逆強化学習を用いた模倣学習の代表的なアプローチの一つである GAIL [Ho 16] に着目する．逆強化学習及び強化学習の合成問題は以下の最適化問題として定式化可能である [Ho 16].

$$\mathrm{RL} \circ \mathrm{IRL} = \underset{\pi \in \Pi}{\operatorname{argmin}} \; -H(\pi(a|s)) + \psi^*(\rho(s,a) - \rho_E(s,a)) \qquad (1)$$

ここで， $H(\pi)$ はエントロピー，$\rho(s,a)$ は状態 s と行動 a の同時分布である．ベルマン制約を満たす $\rho(s,a)$ の集合 $\mathcal{M} = \{\rho_\pi : \pi \in \Pi\} = \{\rho : \rho \geq 0 \text{ and } \sum_a \rho(s,a) = \rho_0(s) + \gamma \sum_{s'} P(s|s',a)\rho(s,a)\}$ を考えた時，$\rho \in \mathcal{M}$ を満たす ρ は方策 π と一対一で対応し，$\pi(a|s) = \rho_\pi(s,a) / \sum_{a'} \rho_\pi(s,a')$ の関係式が成立することが示されている [Syed 08]． また， $\psi^* : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \bar{\mathbb{R}}$ はコスト関数 $l(s,a)$ (報酬関数の異符号) を変数とした凸正則化関数 $\psi(l) : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \bar{\mathbb{R}}$ の共役関数であり，

$$\psi^*(\rho - \rho_E) = \sup_{l \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}} \sum_{s,a} (\rho(s,a) - \rho_E(s,a)) l(s,a) - \psi(l)$$

である．ただし，$\bar{\mathbb{R}}$ は拡張実数である．式 (1) の目的関数は，$\rho(s,a)$ と $l(s,a)$ を変数とする関数として見たときに鞍点が存在し， 唯一の最適解を持つことが保証される．さらに， この最適化問題を，方策を表現する生成器とコスト関数を推定する識別器の学習として捉えることで，Generative Adversarial Nets (GAN) [Goodfellow 14] の学習則を適用することが可能である．

$$\underset{\boldsymbol{w}}{\text{maximize}} \;\; \mathbb{E}_{\pi_E}[\log(1 - D_{\boldsymbol{w}}(s,a))] + \mathbb{E}_{\pi_\theta}[\log D_{\boldsymbol{w}}(s,a)]$$

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \;\; \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[\log D_{\boldsymbol{w}}(s,a)] - \lambda H(\pi_{\boldsymbol{\theta}})$$

ただし， $\boldsymbol{w}$ および $\boldsymbol{\theta}$ はそれぞれ， 識別器と生成器のパラメータである．$D(\cdot)$ は識別器の出力であり，入力された状態 s, 行動 a が教示者のものである確率を示す．識別器は入力された状態行動対を生成した分布が生成器と教示者のどちらであるか正しく識別するように学習を行う．一方， 生成器は入力された状態に対して， 識別器が識別を誤るような行動の選択確率を出力するように学習する．

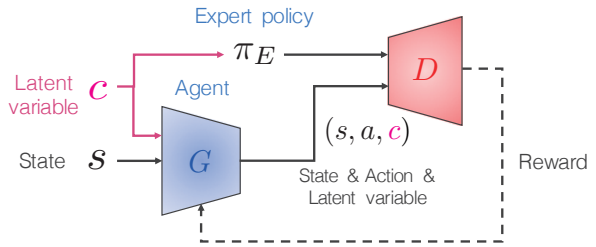


図 1: 提案手法の概念図.

2.2 潜在変数を導入した方策の学習

多様な行動を表現可能な方策を学習することを目的として，単一の方策に対して潜在変数 c を導入する手法が提案されている [Li 17, Hausman 17]．InfoGAIL [Li 17] は生成器の目的関数に潜在変数 c と軌道 τ の相互情報量を最大化する補正項を導入することで， 潜在変数によって条件付け可能な方策 $\pi_\theta(a|s,c)$ を学習する．

$$\underset{\theta}{\text{minimize}}\ \mathbb{E}_{\pi_\theta}[\log D_w(s,a)] - \lambda_1 H(\pi_\theta) - \lambda_2 I(c;\tau\sim\pi_\theta)$$

ただし， $I(c;\tau)$ は潜在変数 c と軌道 τ の相互情報量である．

InfoGAIL は潜在変数を切り替えることで，ある一つのタスクを達成する多様な手段を表現することが可能である．しかし， 報酬関数を表現する識別器に対して潜在変数を導入していないため，同じ状態において，最適な行動が異なるようなタスクそれぞれに対応する報酬を表現することができない．近年提案された conditional GAIL [Merel 17] は，モーションキャプチャから人の複雑な動きを模倣することを目的として，GAIL の枠組みを拡張した研究である．conditional GAIL は生成器と識別器の両方に対して潜在変数を導入しているため，複数のタスクに対応する報酬関数を表現可能であり，結果として単一の学習器で複数のタスクを学習できる．しかし，実験的には，一つのタスクを達成する複数の手段を獲得したこと示しているのみである．また，conditional GAIL の学習則は最適解の存在が保証されていない．特に生成器の目的関数にエントロピー正則化項が含まれていないため，GAIL [Ho 16] の問題設定と異なり，解の一意性に関する議論が適用できない可能性がある．

3. 提案手法

提案手法は，生成器および識別器の両方に対して潜在変数 c を導入することで，複数のタスクに対応する報酬関数と方策 $\pi(a|s,c)$ を学習する．図 1 に，提案手法の概念図を示す．本研究では， 教示者の行動データに対して事前にタスクを表現する潜在変数が与えられていると仮定する．

生成器（Generator，図中の G）は状態 s と潜在変数 c を観測して， 行動 a の選択確率を出力する．一方， 識別器（Discriminator，図中の D）は状態 s， 行動 a および潜在変数 c を入力として受け取り， それらを生成した分布が生成器と教示者のどちらであるかを識別する．識別器の出力は $[0,1]$ のスカラー値であり， 教示者の行動データに対して高い値を出力するように学習を行う．対して生成器は， 識別器が誤識別するような行動選択確率を出力するように学習を行う．

提案手法の目的関数を以下に示す．GAIL と同様に，二つの学習モデルを用いて報酬関数と最適方策 π_θ を同時に求める．加えて，識別器と生成器の両方に潜在変数を導入することで，異なるタスクに対応する報酬関数および方策 $\pi(a|s,c)$ を獲得することが可能である．

$$\underset{w}{\text{minimize}}\ \mathbb{E}_{\pi_\theta}[\log(D_w(s,a,c))] + \mathbb{E}_{\pi_E}[\log(1-D_w(s,a,c))] \quad (2)$$

$$\underset{\theta}{\text{maximize}}\ \mathbb{E}_{\pi_\theta}[\log(D_w(s,a,c))] - \mathbb{E}_{\pi_\theta}[\log(1-D_w(s,a,c))] \quad (3)$$

ただし， $\boldsymbol{w}$ および $\boldsymbol{\theta}$ はそれぞれ，識別器と生成器のパラメータである．ここで， 提案手法は Adversarial Inverse Reinforcement Learning (AIRL) [Fu 17] で提案された識別器の構造を適用して，

$$D_w(s,a,c) = \frac{\exp(f_w(s,a,c))}{\exp(f_w(s,a,c)) + \pi_\theta(a|s,c)} \quad (4)$$

と定義する．式 (3) に対して (4) を代入することで，GAN の学習則より，最適解に収束した時に以下が成立する．

$$f^*(s,a,c) = \log\pi^*(a|s,c) = Q^*(s,a,c) - V^*(s,c) = A^*(a,s,c) \quad (5)$$

ただし， $\pi^*(a|,s,c)$ は最適方策， $V^*(s,c)$ は最適状態価値関数， $Q^*(s,a,c)$ は最適行動価値関数， $A^*(s,a,c)$ は $\pi^*(a|,s,c)$ のもとでのアドバンテージ関数である．式 (5) の二つ目の等号は，エントロピー正則化付マルコフ決定過程 [Nachum 17] において，正則化の係数が 1 の場合のみに成立する．

さらに， 提案した目的関数が GAIL と同様の議論によって最適解を唯一持つことを保証する．つまり， 以下の命題が成立する．

命題 1. 提案手法の目的関数は以下の最適化問題と等価である．

$$\underset{\pi\in\Pi}{\text{argmin}}\ -H(\pi(a|s,c)) + \psi^*(\rho(s,a,c) - \rho_E(s,a,c)) \quad (6)$$

命題 1 は潜在変数を導入したエントロピー正則化付マルコフ決定過程を考え，GAIL と同様に二つの目的関数 (2)，(3) が式 (6) に帰着可能であることを示すことで証明可能である．GAIL との差異は潜在変数を導入した同時分布 $\rho(s,a,c)$， エントロピー $H(\pi) = \mathbb{E}_{\pi,c}[-\log\pi(a|s,c)]$ を考慮している点である．

本研究ではさらに， 学習の効率化を目的として，生成器の目的関数に対してエントロピー正則化項に対する係数補正項 $\beta\log\pi(a|s,c)$ を導入する，

$$\begin{aligned}\underset{\theta}{\text{maximize}}\ &\mathbb{E}_{\pi_\theta}[\log(D_w(s,a,c))] - \mathbb{E}_{\pi_\theta}[\log(1-D_w(s,a,c))] \\ &\quad + \beta\mathbb{E}_{\pi_\theta}[\log(\pi_\theta(a|s,c))] \\ &= \mathbb{E}_{\pi_\theta}[f_w(s,a,c)] - (1-\beta)\mathbb{E}_{\pi_\theta}[\log\pi_\theta(a|s,c)] \end{aligned} \quad (7)$$

エントロピー正則化項に対する係数補正のパラメータ $\beta\in[0,1]$ は方策の学習において識別器から得られる報酬の最大化とエントロピー最大化の重要度を調整する機能を果たす．学習初期は β を大きく設定することで，エントロピー正則化項を小さくし， 教示者の方策を近似する方向に学習を進める．一方，学習後期はエントロピー正則化項を大きくすることで， 生成器の探索を促し，状態，行動空間全体を埋めるような汎化性能の高い方策を学習することが可能になる．

提案手法のアルゴリズムを Algorithm1 に示す．本研究では， 生成器のパラメータを最適化するアルゴリズムとして，方策探索の学習手法である Trust Region Policy Optimization (TRPO) [Schulman 15] を用いた．

Algorithm 1

Input:
- Expert trajectories, latent variables, entropy-regularization correction parameter and scheduling parmeter

  $\tau_E \sim \pi_E,\ c_E \sim p(c),\ \beta,\ \Delta\beta$
- Initial Generator and Discriminator parameters and learning rate for Discriminator

  $\boldsymbol{\theta} = \boldsymbol{\theta}_0,\ \boldsymbol{w} = \boldsymbol{w}_0,\ \alpha_{\boldsymbol{w}}$

Learnig:

for $i = 0, 1, 2, \ldots$ **do**
- Sample latent variables and trajectories

  $c_i \sim p(c),\ \tau_i \sim \pi_{\theta_i}$
- Discriminator update

  $\boldsymbol{w}_{i+1} = \boldsymbol{w}_i + \alpha_{\boldsymbol{w}} \Delta \boldsymbol{w}_i$

  where $\Delta \boldsymbol{w}_i = \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[\nabla_{\boldsymbol{w}} \log(1 - D_{\boldsymbol{w}}(s,a,c))]$

  $+\mathbb{E}_{\pi_E}[\nabla_{\boldsymbol{w}} \log(D_{\boldsymbol{w}}(s,a,c))]$
- Generator update (using policy gradient algorithm) with following reward function

  $\mathcal{R}_{\pi_{\boldsymbol{\theta}}} = \log(D_{\boldsymbol{w}_{i+1}}(s,a,c))$

  $- \log(1 - D_{\boldsymbol{w}_{i+1}}(s,a,c)) + \beta \log(\pi_\theta(a|s,c))$
- Schedule Entropy correntionparameter

  $\beta \leftarrow \beta + \Delta\beta$

end for

4. 実験

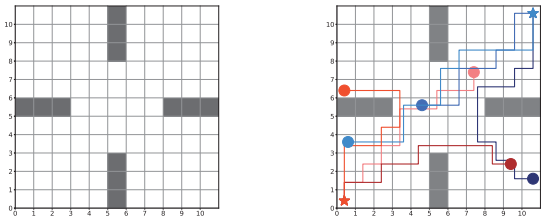
学習するタスクを，グリッドワールド内で目標の位置へ向かう迷路タスクとする．学習エージェントの目的は異なる座標へ向かう二つのタスクを模倣することである．図 2(a) に実験で使用したグリッドワールドを示す．本研究では，11×11 の環境を用いた．状態 s は $[0, 1]$ に正規化した x 座標，y 座標の離散 2 次元ベクトルとした．各状態での行動 a は（右, 上, 左, 下）のうち選択した行動のみが値 1 を持つ one-hot の 4 次元ベクトルで表現した．本研究は決定的な状態遷移則を用いた．ただし，網掛け部分は壁であり，学習エージェントはこの座標へ遷移することはできない．また，本研究ではタスクを表現する潜在変数 c を 3 次元 one-hot ベクトルで表現した．

教示者の行動データは，グリッドワールドからランダムにサンプリングした初期状態から，目的の座標に最短ステップ数で到達する軌道を各タスク 30 個ずつ用意した（図 2(b)）．ただし，二つのタスクのうち「タスク 1」は右上 (1, 1),「タスク 2」は左下 (0.0) へ向かうタスクと定義する．

図 3 に方策のみに潜在変数を導入した既存手法と性能を比較した結果を示す．40 個初期状態を選択してエピソードを開始した時のゴールに到達した試行数を評価する．本実験では，$\beta = 0.9,\ \Delta\beta = 0.0$ とした．結果から，同じ状態において矛盾するような行動をとる場合，既存手法である InfoGAIL，および通常の InfoGAIL+AIRL では学習に失敗しているが，提案手法は二つのタスクに対する方策を学習できていることが分かる．

提案手法が獲得した報酬関数に従う状態価値を図 4 に示す．提案手法は図 4 に示すように，タスク情報を潜在変数という形で導入することでそれぞれのタスクに応じた報酬を正しく表現可能であるため学習に成功した．一方，既存手法 InfoGAIL では報酬関数を全タスク同一のものとして学習してしまうため，正しく価値が学習できず失敗したと考えられる．

しかし，既存手法 InfoGAIL+AIRL with ERC 条件は最適方策（教示者）の約 7 割の性能を獲得している．図 5 に InfoGAIL+AIRL（通常，with ERC）が獲得した「タスク 1」の価



(a) グリッドワールド　(b) 教示者の軌道例

図 2: グリッドワールド環境および，教示者の軌道の一例．軌道中の ● は初期状態，★ は終端状態を示す．

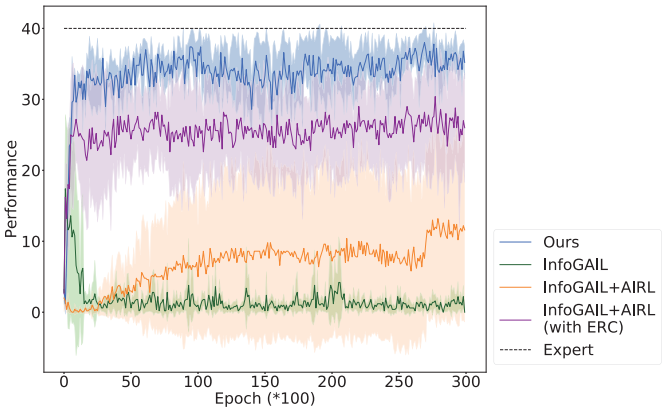
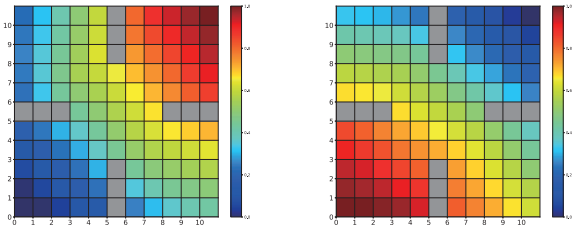


図 3: 学習性能の比較．横軸は学習エポック，縦軸はランダムに 40 個初期状態を選択してエピソードを開始した時のゴールに到達した試行数を示す．40 個のうち，タスク 1 とタスク 2 をぞれぞれ 20 個ずつ生成した．また，InfoGAIL+AIRL の ERC はエントロピー正則化補正項（Entropy Regularized Correction term）を示す．また，教示者の性能は理論値である

値関数 $V^\pi(s,c)$ を示す．AIRL の学習則は InfoGAIL と組み合わさることで，生成器が受け取る擬似的な報酬 $\hat{R}$ を以下の式で表現する．

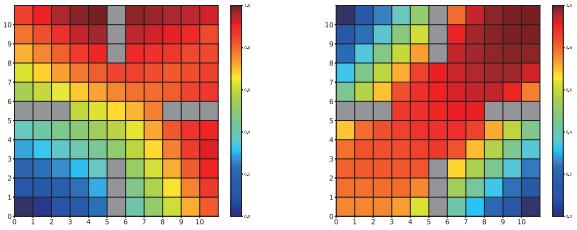
$$\hat{R} = f(s,a) - (1-\beta)\log\pi(a|s,c)$$

ただし，通常の条件では $\beta = 0$ である．これは，模倣の精度を評価する項 $f(s,a)$ とエントロピー正則化項に分解でき，さらに，$\pi(a|s,c)$ の値が潜在変数を条件として大きく変わることで，生成器の受け取る報酬 $\hat{R}$ の値が変化しうることを意味する．また，本実験では，二つのタスクの目標座標は対角に位置するため，各タスクの報酬関数は鏡像関係にあり，二つの報酬関数を重ね合わせることで，目標座標付近を部分的に表現可能である．つまり，今回は，タスクの違いを潜在変数 c による方策 $\pi(a|s,c)$ の値の差のみで表現することが可能である．しかし，通常の InfoGAIL+AIRL 条件ではエントロピー正則化項の重要度が，模倣の精度を評価する $f(s,a)$ に比べて非常に大きくなる．その結果，$f(s,a)$ が無視されるため，図 5(a) に示すように，目標座標付近以外の状態価値が高くなり，学習に失敗したと考えられる．一方，エントロピー正則化の係数を補正することで，$f(s,a)$ とエントロピー正則化項の重要度が適切に設定される．そのため，InfoGAIL+AIRL with ERC 条件は，模倣するという目的を失わず，タスク間，つまり潜在変数によって方策 $\pi(a|s,c)$ の値を切り替えて，タスクごとに異なる擬似報酬



(a) タスク 1　(b) タスク 2

図 4: 提案手法の獲得した価値関数 $V^{\pi}(s,c)$.



(a) 通常：タスク 1　(b) ERC あり：タスク 1

図 5: InfoGAIL+AIRL 条件の獲得した価値関数 $V^{\pi}(s,c)$.

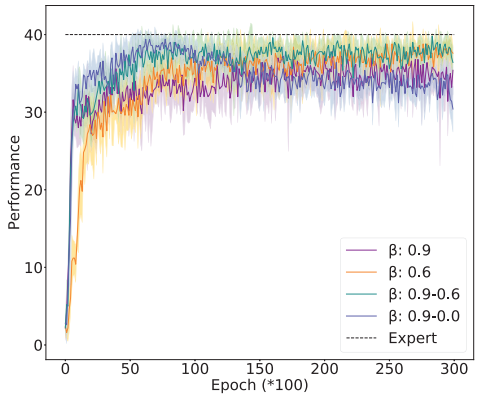


図 6: エントロピー正則化の係数補正項を変化させた時の提案手法の学習結果.

$\hat{R}$ を受け取ることができ，7 割程度の学習に成功したと考える（図 5(b)）．ただし，提案手法は潜在変数の入力が，$f(s,a,c)$ に直接作用するため，提案手法の方がタスクの違いをより顕著に表現でき，高い性能を示している．

次に，エントロピー正則化の係数補正項による学習速度と獲得した方策の性能の影響について検証した．エントロピー正則化は方策の分散を大きくすることから，環境中の探索を促す効果が期待できる．そこで今回は教示者の行動データを各タスクについて 5 個に減らすことで，教示者の行動データを模倣するだけでなく，汎化が必要な問題設定にした．

図 6 にエントロピー正則化の係数補正項を変化させた時の提案手法の学習結果を示す．β の値を固定した条件間で学習曲線を比較した場合，β の値が小さい条件（$\beta = 0.6$）は探索の結果，汎化が行われるため，比較的獲得した方策の性能は高いが，収束までに時間を要している．一方，β の値を大きく設定（$\beta = 0.9$）することで，エントロピー正則化項が抑制され，学習初期に教示者の行動を模倣する方向に学習が進むため，収束は速い．しかし，探索を行いにくくなるため，汎化された方策を獲得できず，高い性能を示さない．β の値を学習中に適切に変動させる条件（$\beta = 0.9 \rightarrow 0.6$）が，収束の速さ，獲得した方策の性能ともに最も良いことが分かる．ただし，β の値が学習中に極端に変動する条件（$\beta = 0.9 \rightarrow 0.0$）では，一度模倣した方策から大きく離れていくため，性能の悪化が確認された．

5. おわりに

本稿では，生成器と識別器の両方に対して潜在変数を導入した敵対的模倣学習の学習則を提案した．これにより，同じ状態であっても，最適な行動が異なるようなタスクそれぞれ対応する報酬関数を推定することができた．また，理論解析により，提案した目的関数について最適解の存在を保証した．さらに，方策の学習に対してエントロピー正則化の係数補正項を導入する学習則を提案した．これにより，方策の学習において識別器から得られる報酬の最大化とエントロピー最大化の重要度を適切に調整することで，学習速度と獲得した方策の性能の向上を可能にした．今後の課題として，教示者の行動データから自動的にタスク数を推定する学習則への拡張が挙げられる．

謝辞

本研究は，JST CREST「認知ミラーリング」（課題番号：JPMJCR16E2）の支援を受けたものである．

参考文献

[Fu 17] Fu, J., Luo, K., and Levine, S.: Learning Robust Rewards with Adversarial Inverse Reinforcement Learning, in *Proceedings of the 6th International Conference on Learning Representations* (2017)

[Goodfellow 14] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y.: Generative Adversarial Nets, in *Advances in Neural Information Processing Systems*, Vol. 27 (2014)

[Hausman 17] Hausman, K., Chebotar, Y., Schaal, S., Sukhatme, G., and Lim, J. J.: Multi-Modal Imitation Learning from Unstructured Demonstrations using Generative Adversarial Nets, in *Advances in Neural Information Processing Systems*, Vol. 30 (2017)

[Ho 16] Ho, J. and Ermon, S.: Generative Adversarial Imitation Learning, in *Advances in Neural Information Processing Systems*, Vol. 29 (2016)

[Li 17] Li, Y., Song, J., and Ermon, S.: InfoGAIL: Interpretable Imitation Learning from Visual Demonstrations, in *Advances in Neural Information Processing Systems*, Vol. 30 (2017)

[Merel 17] Merel, J., Tassa, Y., Srinivasan, S., Lemmon, J., Wang, Z., Wayne, G., and Heess, N.: Learning Human Behaviors From Motion Capture by Adversarial Imitation, *arXiv preprint arXiv:1707.02201* (2017)

[Nachum 17] Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D.: Bridging the Gap Between Value and Policy Based Reinforcement Learning, in *Advances in Neural Information Processing Systems*, Vol. 30 (2017)

[Russell 98] Russell, S.: Learning Agents For Uncertain Environments (Extended Abstract), in *Proceedings of the 11th Annual Conference on Computational Learning Theory* (1998)

[Schulman 15] Schulman, J., Levine, S., Moritz, P., Jordan, M., and Abbeel, P.: Trust Region Policy Optimization, in *Proceedings of the 32nd International Conference on Machine Learning* (2015)

[Syed 08] Syed, U., Bowling, M., and Schapire, R. E.: Apprenticeship Learning Using Linear Programming, in *Proceedings of the 25th International Conference on Machine Learning* (2008)