

# 敵対的模倣学習

吉田 英樹

2024 年 7 月 2 日

## 概要

本稿は、敵対的模倣学習 (Generative Adversarial Imitation Learning: GAIL)[1] の備忘録である。

## 1 はじめに

敵対的模倣学習 (Generative Adversarial Imitation Learning: GAIL)[1] は、逆強化学習と強化学習を用いた模倣学習のひとつである。逆強化学習 (Inverse Reinforcement Learning: IRL) は、エキスパートの方策  $\pi$  が与えられたとき、報酬関数を推定する。強化学習 (Reinforcement Learning: RL) は、報酬関数が与えられたとき、方策  $\pi$  を推定する。GAIL は、IRL と RL の合成問題で与えられ、式 (1) で定式化ができる。

$$\text{RL} \circ \text{IRL}(\pi_E) = \arg \min_{\pi \in \Pi} D_{\text{JS}}(\rho_\pi, \rho_{\pi_E}) - \lambda H(\pi) \quad (1)$$

ここで、 $D_{\text{JS}}$  は Jensen-Shannon divergence であり、 $H$  はエントロピーによる正則化項である。 $\rho_\pi$  は方策  $\pi$  の occupancy measure[2] である。式 (1) は、方策  $\pi$  の  $(s, a)$  に訪問する確率  $\rho_\pi$  を、エキスパートの方策  $\pi_E$  の  $(s, a)$  に訪問する確率  $\rho_{\pi_E}$  に近づける。式 (1) は、Jensen-Shannon divergence 最小化問題なので、敵対的生成ネットワーク (Generative Adversarial Nets: GAN)[3] のアルゴリズムを適用することができる。

## 2 逆強化学習と強化学習の定式化

逆強化学習は, 式 (2) で定式化できる.

$$\max_{c \in \mathcal{C}} \left( \min_{\pi \in \Pi} -H(\pi) + \mathbb{E}_{\pi}[c(s, a)] \right) - \mathbb{E}_{\pi_E}[c(s, a)] \quad (2)$$

実際, 式 (2) の双対問題を考えれば, 式 (3) を得る. 式 (3) は, 最大エントロピー原理を用いた逆強化学習 [4] に対応することがわかる. 導出は, 次節で説明する.

$$\begin{aligned} \min_{\rho \in \mathcal{D}} \quad & -\bar{H}(\rho) \\ \text{s.t.} \quad & \rho(s, a) = \rho_E(s, a) \end{aligned} \quad (3)$$

強化学習は, 式 (4) で定式化できる.

$$\text{RL}(c) = \arg \min_{\pi \in \Pi} -H(\pi) + \mathbb{E}_{\pi}[c(s, a)] \quad (4)$$

### 3 occupancy measure の特徴づけ

**定義 1** (occupancy measure).  $\rho_\pi : \mathcal{S} \rightarrow \mathbb{R}$  とし, 方策  $\pi$  に従う状態を訪問する標準化されていない確率とする.

$$\rho_\pi(s) = \sum_{t=0}^{\infty} \gamma^t P(s_t = s | \pi) \quad (5)$$

このとき, 方策  $\pi$  の *occupancy measure*  $\rho_\pi(s, a)$  が定まる.

$$\rho_\pi(s, a) = \rho_\pi(s) \pi(a | s) \quad (6)$$

occupancy measure  $\rho_\pi(s, a)$  は, 方策  $\pi$  に従うときの状態と行動  $(s, a)$  の訪問回数の期待値を意味する.

$$\rho_\pi(s, a) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t \mathbf{1}_{(s_t=s \wedge a_t=a)} \middle| p_0, \pi, P \right] \quad (7)$$

occupancy measure は, 総コストの期待値の代わりにの計算式を与える.

$$\mathbb{E}_\pi[c(s, a)] = \sum_{t=0}^{\infty} \gamma^t \sum_s \sum_a P(s_t = s, a_t = a) c(s, a) \quad (8)$$

$$= \sum_{t=0}^{\infty} \gamma^t \sum_s \sum_a P(s_t = s | \pi) \pi(a | s) c(s, a) \quad (9)$$

$$= \sum_{t=0}^{\infty} \gamma^t \sum_s P(s_t = s | \pi) \sum_a \pi(a | s) c(s, a) \quad (10)$$

$$= \sum_s \gamma^t \sum_{t=0}^{\infty} P(s_t = s | \pi) \sum_a \pi(a | s) c(s, a) \quad (11)$$

$$= \sum_s \rho_\pi(s) \sum_a \pi(a | s) c(s, a) \quad (12)$$

$$= \sum_s \sum_a \rho_\pi(s, a) c(s, a) \quad (13)$$

**定義 2** (Bellman flow constraint). *Bellman flow constraint* を下式で定義する.

$$\sum_a \rho(s, a) = p_0(s) + \gamma \sum_{s'} \sum_a P(s | s', a) \rho(s', a) \quad (14)$$

$$\rho \geq 0 \quad (15)$$

定義 2 は,  $\rho(s, a)$  の行動  $a$  の周辺化と,  $\rho(s', a)$  の行動  $a$  の周辺化の期待値が一致しなければならないことを意味している.

**定義 3** ( $\pi$ -specific Bellman flow constraint).  $\pi$ -specific *Bellman flow constraint* を下式で定義する.

$$\rho(s, a) = \pi(a | s) p_0(s) + \pi(a | s) \gamma \sum_{s'} \sum_{a'} \rho(s', a') P(s | s', a') \quad (16)$$

$$\rho \geq 0 \quad (17)$$

**補題 3.1** (Lemma 2 of Syed et al. [2]). 任意の静的方策  $\pi$  において, 方策  $\pi$  の *occupancy measure*  $\rho_\pi$  は,  $\pi$ -specific Bellman flow constraint を満たす.

**証明.**  $\rho_\pi(s, a)$  は明らかに非負である.

$$\rho_\pi(s, a) \tag{18}$$

$$= \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t \mathbf{1}_{(s_t=s \wedge a_t=a)} \right] \tag{19}$$

$$= \sum_{t=0}^{\infty} \gamma^t P(s_t = s, a_t = a) \tag{20}$$

$$= P(s_0 = s, a_0 = a) + \sum_{t=1}^{\infty} \gamma^t P(s_t = s, a_t = a) \tag{21}$$

$$= \pi(a|s)p_0(s) + \sum_{t=0}^{\infty} \gamma^{t+1} P(s_{t+1} = s, a_{t+1} = a) \tag{22}$$

$$= \pi(a|s)p_0(s) + \sum_{t=0}^{\infty} \gamma^{t+1} \sum_{s'} \sum_{a'} P(s_t = s', a_t = a', s_{t+1} = s, a_{t+1} = a) \tag{23}$$

$$= \pi(a|s)p_0(s) + \sum_{t=0}^{\infty} \gamma^{t+1} \sum_{s'} \sum_{a'} P(s_t = s', a_t = a') P(s_{t+1} = s, a_{t+1} = a | s_t = s', a_t = a') \tag{24}$$

$$= \pi(a|s)p_0(s) + \sum_{t=0}^{\infty} \gamma^{t+1} \sum_{s'} \sum_{a'} P(s_t = s', a_t = a') P(s_{t+1} = s | s_t = s', a_t = a') \pi(a|s) \tag{25}$$

$$= \pi(a|s)p_0(s) + \pi(a|s)\gamma \sum_{s'} \sum_{a'} \sum_{t=0}^{\infty} \gamma^t P(s_t = s', a_t = a') P(s_{t+1} = s | s_t = s', a_t = a') \tag{26}$$

$$= \pi(a|s)p_0(s) + \pi(a|s)\gamma \sum_{s'} \sum_{a'} \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t \mathbf{1}_{(s_t=s' \wedge a_t=a')} \right] P(s|s', a') \tag{27}$$

$$= \pi(a|s)p_0(s) + \pi(a|s)\gamma \sum_{s'} \sum_{a'} \rho_\pi(s', a') P(s|s', a') \tag{28}$$

□

**補題 3.2.** *strictly diagonally dominant matrix* は非特異行列である.

**証明.** *strictly diagonally dominant matrix*  $A$  が, 特異行列であると仮定する.

$A$  は特異行列なので,  $Ax = 0$  のとき  $x \neq 0$  が存在する.  $x_i$  は,  $|x_i| \neq 0$  で, 要素の中で絶対値が最大となると仮定する.

$$\sum_j a_{ij}x_j = 0 \quad (29)$$

$$a_{ii}x_i = -\sum_{j \neq i} a_{ij}x_j \quad (30)$$

$$a_{ii} = -\sum_{j \neq i} \frac{x_j}{x_i} a_{ij} \quad (31)$$

$$|a_{ii}| = \left| \sum_{j \neq i} \frac{x_j}{x_i} a_{ij} \right| \quad (32)$$

$$\leq \sum_{j \neq i} \left| \frac{x_j}{x_i} a_{ij} \right| \quad (33)$$

$$\leq \sum_{j \neq i} \left| \frac{x_j}{x_i} \right| |a_{ij}| \quad (34)$$

$$\leq \sum_{j \neq i} |a_{ij}| \quad (35)$$

行列  $A$  が, *strictly diagonally dominant matrix* であることと矛盾する. □

**補題 3.3** (Lemma 3 of Syed et al. [2]). 任意の静的方策  $\pi$  において,  $\pi$ -specific Bellman flow constraint は, 高々1つの解しかない。

**証明.** 行列  $A$  と列ベクトル  $b$  を下式で定義すると,  $A\rho = b$  および  $\rho \geq 0$  となる。

$$A_{((s,a),(s',a'))} = \begin{cases} 1 - \gamma P(s|s', a') \pi(a|s) & \text{if } (s, a) = (s', a') \\ -\gamma P(s|s', a') \pi(a|s) & \text{otherwise} \end{cases} \quad (36)$$

$$b_{((s,a))} = \pi(a|s) p_0(s) \quad (37)$$

$$\rho_{((s,a))} = \rho(s, a) \quad (38)$$

行列  $A$  は, strictly diagonally dominant matrix である。実際,  $\sum_{s'} P(s'|s, a) = 1$ ,  $\sum_a \pi(a|s)$  および  $\gamma < 1$  であるので,

$$\sum_s \sum_a \gamma P(s|s', a') \pi(a|s) = \gamma < 1 \quad (39)$$

$$\Rightarrow 1 - \gamma P(s|s', a') \pi(a|s) = \sum_{s \neq s'} \sum_{a \neq a'} \gamma P(s|s', a') \pi(a|s) \quad (40)$$

$$\Rightarrow |A_{((s,a),(s',a'))}| > \sum_{s \neq s'} \sum_{a \neq a'} |A_{((s,a),(s',a'))}| \quad (41)$$

行列  $A$  は, 補題 3.2 より非特異行列である。よって,  $A\rho = b$  および  $\rho \geq 0$  は, 高々1つの解しかない。□

Bellman flow constraint と occupancy measure の重要な性質として, 定理 3.4 が成り立つ。

**定理 3.4** (Theorem 2 of Syed et al. [2]).  $\mathcal{D} = \{\rho : \rho \geq 0, \sum_a \rho(s, a) = p_0(s) + \gamma \sum_{s'} \sum_a P(s|s', a) \rho(s', a)\}$  とする。  $\rho \in \mathcal{D}$  ならば,  $\rho$  は,  $\pi_\rho(a|s) := \frac{\rho(s, a)}{\sum_{a'} \rho(s, a')}$  の occupancy measure である。さらに,  $\pi_\rho(a|s)$  は一意である。

**証明.** 仮定より,  $\pi_\rho(a|s) := \frac{\rho(s, a)}{\sum_{a'} \rho(s, a')}$  である。  $\rho \in \mathcal{D}$  であるので,

$$\pi_\rho(a|s) = \frac{\rho(s, a)}{\sum_{a'} \rho(s, a')} \quad (42)$$

$$= \frac{\rho(s, a)}{p_0(s) + \gamma \sum_{s'} \sum_a P(s|s', a) \rho(s', a)} \quad (43)$$

よって,

$$\rho(s, a) = \pi_\rho(a|s) p_0(s) + \pi_\rho(a|s) \gamma \sum_{s'} \sum_a P(s|s', a) \rho(s', a) \quad (44)$$

$$\rho(s, a) \geq 0 \quad (45)$$

$\rho(s, a)$  は,  $\pi$ -specific Bellman flow constraint を満たす。

補題 3.1 より,  $\rho$  は,  $\pi_\rho(a|s) := \frac{\rho(s, a)}{\sum_{a'} \rho(s, a')}$  の occupancy measure である。

補題 3.3 より,  $\rho(s, a)$  は高々1つの解しかない。よって,  $\pi_\rho(s, a)$  は高々1つの解しかない。□

## 4 GAIL の最適方策の特徴づけ

逆強化学習にコスト関数  $c$  の正則化項を追加する.

$$\text{IRL}_\psi(\pi_E) = \arg \max_{c \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}} -\psi(c) + \left( \min_{\pi \in \Pi} -H(\pi) + \mathbb{E}_\pi[c(s, a)] \right) - \mathbb{E}_{\pi_E}[c(s, a)] \quad (46)$$

**定理 4.1.**  $RL \circ \text{IRL}_\psi = \arg \min_{\pi \in \Pi} -H(\pi) + \psi^*(\rho_\pi - \rho_{\pi_E})$

**証明.** 論文 [1] の付録を参照せよ.  $\square$

**補題 4.2.**  $\bar{H}$  は *strictly concave* であり,  $H(\pi) = \bar{H}(\rho_\pi)$  であり,  $\bar{H}(\rho) = H(\pi_\rho)$  である.  
ただし,

$$\bar{H}(\rho) = - \sum_s \sum_a \frac{\rho(s, a) \log \rho(s, a)}{\sum_{a'} \rho(s, a')} \quad (47)$$

**証明.** 論文 [1] の付録を参照せよ.  $\square$

**補題 4.3.**  $L(\pi, c) = -H(\pi) + \mathbb{E}_\pi[c(s, a)]$  および  $\bar{L}(\rho, c) = -\bar{H}(\rho) + \sum_s \sum_a \rho(s, a) c(s, a)$  とする.

任意のコスト関数  $c$ , 任意の方策  $\pi \in \Pi$  において,  $L(\pi, c) = \bar{L}(\rho_\pi, c)$ .

任意のコスト関数  $c$ , 任意の *occupancy measure*  $\rho \in \mathcal{D}$  において,  $\bar{L}(\rho, c) = L(\pi_\rho, c)$ .

**証明.** 論文 [1] の付録を参照せよ.  $\square$

**定理 4.4.**  $\psi$  が定数関数で,  $\tilde{c} \in \text{IRL}_\psi(\pi_E)$ ,  $\tilde{\pi} \in RL(\tilde{c})$  のとき,  $\rho_{\tilde{\pi}} = \rho_{\pi_E}$ .

**証明.**  $\bar{L}(\rho, c)$  を定義する.

$$\bar{L}(\rho, c) = -\bar{H}(\rho) + \sum_s \sum_a c(s, a) (\rho(s, a) - \rho_E(s, a)) \quad (48)$$

$$\tilde{c} \in \text{IRL}_\psi(\pi_E) \quad (49)$$

$$= \arg \max_{c \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}} -\psi(c) + \left( \min_{\pi \in \Pi} -H(\pi) + \mathbb{E}_\pi[c(s, a)] \right) - \mathbb{E}_{\pi_E}[c(s, a)] \quad (50)$$

$$= \arg \max_{c \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}} \min_{\pi \in \Pi} \text{constant} - H(\pi) + \mathbb{E}_\pi[c(s, a)] - \mathbb{E}_{\pi_E}[c(s, a)] \quad (51)$$

$$= \arg \max_{c \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}} \min_{\rho \in \mathcal{D}} -\bar{H}(\rho) + \sum_s \sum_a \rho(s, a) c(s, a) - \sum_s \sum_a \rho_E(s, a) c(s, a) \quad (52)$$

$$= \arg \max_{c \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}} \min_{\rho \in \mathcal{D}} \bar{L}(\rho, c) \quad (53)$$

$$= \arg \min_{\rho \in \mathcal{D}} \max_{c \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}} \bar{L}(\rho, c) \quad (54)$$

双対問題を得る.

$$\begin{aligned} \min_{\rho \in \mathcal{D}} \quad & -\bar{H}(\rho) \\ \text{s.t.} \quad & \rho(s, a) = \rho_E(s, a) \end{aligned} \quad (55)$$

$$\tilde{\rho} = \arg \min_{\rho \in \mathcal{D}} \bar{L}(\rho, \tilde{c}) \quad (56)$$

$$= \arg \min_{\rho \in \mathcal{D}} -\bar{H}(\rho) + \sum_s \sum_a \tilde{c}(s, a) (\rho(s, a) - \rho_E(s, a)) \quad (57)$$

$$= \arg \min_{\rho \in \mathcal{D}} -\bar{H}(\rho) + \sum_s \sum_a \tilde{c}(s, a) \rho(s, a) \quad (58)$$

$$= \rho_{\tilde{\pi}} \quad (\because \text{補題 4.3}) \quad (59)$$

$$= \rho_E \quad (\because \text{式 (55)}) \quad (60)$$

$\square$

## 5 実践的な occupancy measure のマッチング

式 (61) の緩和問題から、既存のアルゴリズムが導出できることを示す.

$$\begin{aligned} \min_{\rho \in \mathcal{D}} \quad & -\bar{H}(\rho) \\ \text{s.t.} \quad & \rho(s, a) = \rho_E(s, a) \end{aligned} \quad (61)$$

式 (61) の緩和問題を考える.

$$\min_{\pi \in \Pi} d_\psi(\rho_\pi, \rho_E) - H(\pi) \quad (62)$$

ただし,  $d_\psi(\rho_\pi, \rho_E) := \psi^*(\rho_\pi - \rho_E)$  である. 式 (62) の特殊なケースとして式 (63) がある.

$$\min_{\pi \in \Pi} \max_{c \in \mathcal{C}} \mathbb{E}_\pi[c(s, a)] - \mathbb{E}_{\pi_E}[c(s, a)] \quad (63)$$

実際, 標示関数  $\delta_{\mathcal{C}}$  を定めると,

$$\delta_{\mathcal{C}}(c) = \begin{cases} 0 & \text{if } c \in \mathcal{C} \\ \infty & \text{otherwise} \end{cases} \quad (64)$$

式 (63) より,

$$\max_{c \in \mathcal{C}} \mathbb{E}_\pi[c(s, a)] - \mathbb{E}_{\pi_E}[c(s, a)] \quad (65)$$

$$= \max_{c \in \mathbb{R}^{\mathcal{C} \times \mathcal{A}}} -\delta_{\mathcal{C}}(c) + \sum_s \sum_a (\rho_\pi(s, a) - \rho_{\pi_E}(s, a)) c(s, a) \quad (66)$$

$$= \delta_{\mathcal{C}}^*(\rho_\pi(s, a) - \rho_{\pi_E}(s, a)) \quad (67)$$

$\psi = \delta_{\mathcal{C}}$  とおけば, 式 (62) をえる.

$\mathcal{C}_{\text{linear}} = \{\sum_i w_i f_i : \|w\|_2 \leq 1\}$  とおけば文献 [5] に相当する.

$\mathcal{C}_{\text{convex}} = \{\sum_i w_i f_i : \sum_i w_i = 1, w_i \geq 1\}$  とおけば文献 [2] に相当する.



## 6 敵対的模倣学習:Generative adversarial imitation learning

敵対的模倣学習は, 式 (68) の緩和問題で定式化できる.

$$\min_{\pi \in \Pi} d_{\psi_{\text{GA}}}(\rho_{\pi}, \rho_E) - \lambda H(\pi) \quad (68)$$

ただし,  $d_{\psi_{\text{GA}}}(\rho_{\pi}, \rho_E) := \psi_{\text{GA}}^*(\rho_{\pi} - \rho_E)$  である.

ここで,

$$\psi_{\text{GA}}(c) = \begin{cases} \mathbb{E}_{\pi_E}[g(c(s, a))] & \text{if } c < 0 \\ +\infty & \text{otherwise} \end{cases} \quad (69)$$

$$g(x) = \begin{cases} -x - \log(1 - e^x) & \text{if } x < 0 \\ +\infty & \text{otherwise} \end{cases} \quad (70)$$

式 (68) は, 式変形すれば, 式 (71) をえる. 式変形の詳細は論文 [1] の付録を参照せよ.

$$\min_{\pi \in \Pi} \max_{D \in (0,1)^{S \times A}} \mathbb{E}_{\pi}[\log D(s, a)] + \mathbb{E}_{\pi_E}[\log(1 - D(s, a))] - \lambda H(\pi) \quad (71)$$

式 (71) は, 敵対的生成ネットワーク (Generative Adversarial Nets: GAN)[3] と一致するため, アルゴリズムを適用することができる. 式 (71) は, Jensen-Shannon divergence 最小化問題となる [3].

$$\min_{\pi \in \Pi} D_{\text{JS}}(\rho_{\pi}, \rho_{\pi_E}) - \lambda H(\pi) \quad (72)$$

ここで,  $D_{\text{JS}}$  は Jensen-Shannon divergence であり,  $H$  はエントロピーによる正則化項である.  $\rho_{\pi}$  は方策  $\pi$  の occupancy measure[2] である. 式 (72) は, 方策  $\pi$  の  $(s, a)$  に訪問する確率  $\rho_{\pi}$  を, エキスパートの方策  $\pi_E$  の  $(s, a)$  に訪問する確率  $\rho_{\pi_E}$  に近づける.

## 参考文献

- [1] J. Ho and S. Ermon, “Generative adversarial imitation learning,” *Advances in Neural Information Processing Systems*, 2016.
- [2] M. B. Umar Syed and R. E. Schapire, “Apprenticeship learning using linear programming,” *International Conference on Machine Learning*, 2008.
- [3] I. J. G. et al., “Generative adversarial nets,” *Advances in Neural Information Processing Systems*, 2014.
- [4] B. D. Ziebart, J. A. Bagnell, and A. K. Dey, “Modeling interaction via the principle of maximum causal entropy,” *International Conference on Machine Learning*, 2010.
- [5] P. Abbeel and A. Y. Ng, “Apprenticeship learning via inverse reinforcement learning,” *International Conference on Machine Learning*, 2004.