# 複数のマルコフ決定過程における逆強化学習に関する研究

千葉大学大学院融合理工学府

地球環境科学専攻 都市環境システムコース

中田 勇介

2021年2月

(千葉大学審査学位論文)

# 複数のマルコフ決定過程における逆強化学習に関する研究

千葉大学大学院融合理工学府

地球環境科学専攻 都市環境システムコース

中田　勇介

2021年2月

# 論文要旨

逆強化学習はタスクの手本であるエキスパートの方策や軌跡データからエキスパートの報酬と方策を推定する機械学習手法である．既存の逆強化学習手法の多くは，ある単一の環境で得られた軌跡からエキスパートの報酬を推定し，この推定報酬を用いた強化学習によってエキスパートの方策を学習する．しかし，エキスパートの方策と推定報酬に対する最適方策が一致した場合においても，推定報酬がエキスパートの報酬と一致する保証はない．そこで本論文では，報酬と方策の推定精度向上を目的として，状態遷移確率が異なる複数の環境におけるエキスパートのデータを用いる逆強化学習問題を提案する．本論文では複数の環境における逆強化学習問題に三つの定式化を与える．一つ目は，線形計画問題としての定式化である．線形計画法を用いる提案法では複数の環境におけるエキスパートの方策に基づく報酬推定によって，推定報酬の精度が改善することを示す．二つ目はベイズ推定としての定式化で，離散状態行動空間の複数環境のエキスパートの軌跡を用いて報酬の事後分布を推定する方法を提案する．三つ目の定式化は敵対的最大エントロピー逆強化学習で，連続状態行動空間の複数の環境におけるエキスパートの軌跡から報酬と方策を推定する．実験では，各提案手法を既存手法と比較し，状態遷移確率が異なる複数の環境におけるエキスパートのデータを用いた逆強化学習が報酬，方策の推定精度向上に寄与することを示す．

# 目 次

# 図 目 次

# 表 目 次

# 第1章　序論

## 1.1　研究の背景

逆強化学習は，あるタスクの解法を知るエージェント（以下，エキスパート）が示す手本を用いてタスクを自動化する方法である．計算機科学分野におけるタスクを自動化する方法は，自動化に用いる情報によって大きく四つに分類される．一つ目は，タスクの手続きの記述に基づくルールベース，二つ目は知識に基づく論理推論 [Huth 04] やエキスパートシステム [Huth 04]，三つ目はタスクの実行結果の評価に基づく強化学習 [Sutton 18] や遺伝的アルゴリズム [Whitley 94]，そしてタスクの手本に基づいて学習する逆強化学習 [Ng 00] である．タスクの手本に基づいて学習する方法の利点は，タスクに必要な情報を具体的な表現（手続きの記述，知識，評価）を作成する必要がない点にある．したがって，人が持つ暗黙知に依存するタスク（例: 自動車の運転，熟練の技術者による機械の制御）においては，タスクの手本に基づく逆強化学習が有用である．

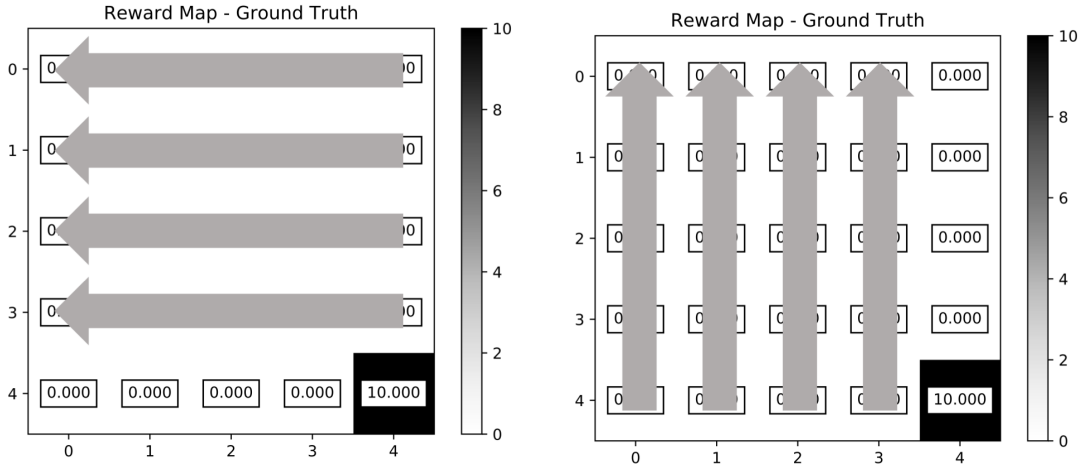
逆強化学習は，タスクの解法を知るエキスパートの多段階の意思決定をマルコフ決定過程（i. e. 環境）[Sigaud 13] でモデル化し，エキスパートの報酬（タスクの評価関数）と方策（意思決定のルール）を推定する手法である [Russell 98, Ng 00]．具体的には，エキスパートの方策が，エキスパートが取り組むタスクの報酬に対して最適であると仮定し，エキスパートの方策が最適となる報酬を推定する．逆強化学習の推定報酬の用途は大きく二つに分けられる．一つは，エキスパートの意思決定の解析である．人や動物，虫などの意思決定が最適となる報酬を推定できれば，その意思決定の目的の解釈が可能になる [Kitani 12, Hirakawa 18]．もう一つの用途は，エキスパートの方策の学習である．エキスパートの報酬を推定できれば，推定報酬に対する最適方策を強化学習 [Barto 89, Sutton 18] で学習することによって，エキスパートの方策を学習できる [Abbeel 04]．ここで重要な点は，エキスパートの方策を学習できる環境はエキスパートの軌跡が得られた環境に留まらないことである．推定報酬は，状態遷移確率が異なる環境に対しても転移可能であるため，転移先の環境におけるエキスパートの方策を学習することができる [Fu 18]．

逆強化学習問題の難しさは，推定報酬が一意に定まらないこと（i.e. 問題の不良設定性）にある [Ng 00]．逆強化学習はエキスパートの方策が最適となる報酬を推定するが，エキスパートの方策が最適となる報酬は複数存在する．そのため，逆強化学習の推定報酬がエキスパートの報酬と一致する保証はない．推定報酬がエキスパートの報酬と異なると，意思決定の解析 [Kitani 12, Hirakawa 18] において誤った解釈が生じたり，転移先の環境におけるエ

キスパートの方策を学習できない [Amodei 16] などの問題が生じる．そのため，エキスパートの報酬に可能な限り近い報酬を推定する必要がある．

## 1.2 研究の目的

本論文の目的は，複数のマルコフ決定過程におけるエキスパートのデータを対象とする逆強化学習手法を提案することによって，逆強化学習の推定性能を改善することである．具体的には，状態遷移確率対象が異なる複数の環境におけるエキスパートの方策，または状態と行動の系列（軌跡）から報酬を推定する逆強化学習を提案する．状態遷移確率が異なる複数のマルコフ決定過程の例を図 1.1 に示す．

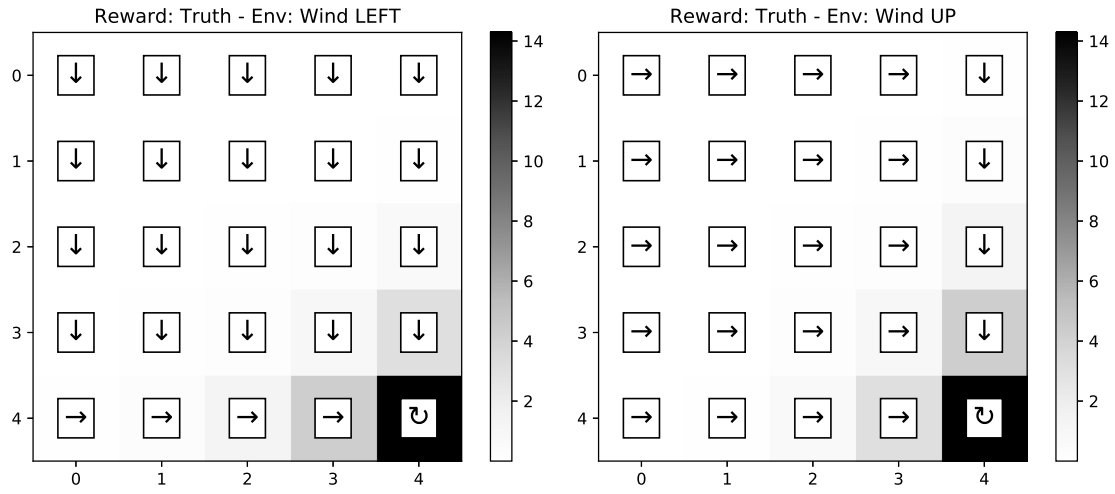



(a) Wind left　　(b) Wind up

図 1.1: 状態遷移確率が異なる複数のマルコフ決定過程（環境）の例

この複数の環境に共通するエキスパートの目的は，右下の報酬が最も高い状態に最短経路で到達することである．これらの環境の違いは，図 1.1 中の灰色の矢印が示す風が吹く方向で，風が吹く状態（マス目）ではエージェントが選択した行動（移動の方向）に関わらず一定確率で風が吹く方向に遷移する．したがって，これら二つの環境では特定の状態で同じ行動を選択しても各環境で異なる状態遷移確率に従って次の状態が決まる．このように図 1.1 に示した複数の環境は「状態遷移確率が異なる複数の環境」である．本論文では簡単のため，「状態遷移確率が異なる複数の環境」を複数の環境と呼ぶ．

図 1.1 の各環境におけるエキスパート方策を図 1.2 に示す．




(a) Wind left (b) Wind up

図 1.2: 複数の環境におけるエキスパート方策の例

図 1.2 の各状態の矢印は，矢印がある状態における最適な行動を示す．各環境におけるエキスパート方策は，風が吹く状態から無風の状態に最短経路で移動した後に，一番右下のゴールに到達する方策である．このように，複数の環境におけるエキスパートの報酬（目的）が共通していても，状態遷移確率が異なる複数の環境においてはエキスパートの方策が異なることが確認できる．

本論文では，状態遷移確率が異なる複数の環境と，図 1.2 が示す各環境におけるエキスパートの方策，またはエキスパートの方策に従って生成されたエキスパートの軌跡から報酬を推定する逆強化学習手法を提案する．複数の環境のデータを用いることによって，報酬の推定に利用できるデータが増加し，単一の環境のデータを用いる場合と比較して報酬の推定結果が改善することが期待できる．

状態遷移確率が異なる複数の環境におけるタスクの例としては，スーパーマリオなどのゲームの異なるステージ（ゴールへの到達や，コインの獲得といった目的はステージ間で共通しているが，各ステージの構成が異なる．）[Choi 11]，異なる交通・路面・車種での運転 [Kuderer 15]，異なる供給条件の下でのプラントの制御などが挙げられる．

「複数の環境間の状態遷移確率の差異」が生じる二つの要因を，車のブレーキの制御を例に説明する．状態遷移確率が異なる一つ目の要因はエージェントの個体差である．車においては，その重量，ブレーキの磨耗量などのがエージェントの個体差に該当する．車の個体差があったとき，複数の車のブレーキを同じ力で制御しても各車の減速度が，その個体差（例：重量，ブレーキの磨耗の仕方）によって異なることが分かる．

もう一つの要因は，エージェントが観測できない環境の差異である．例えば，車が走行する路面の摩擦係数は，塗装の種類や天候によって異なり，エージェントは路面の摩擦係数を観測できない．路面の摩擦係数が異なれば，特定の車のブレーキを同じ力で制御した時に生

じる減速度も異なる．これら，二つの要因，エージェントの個体差，エージェントが観測できない環境の差異，によって生じる車の減速度の違いは，マルコフ決定過程における状態遷移確率の差異として等価に扱かうことができる．これら状態遷移確率が異なる複数の環境においてブレーキを制御するエキスパートが一つの報酬に従うと仮定すると，単一の環境だけでなく，複数の環境におけるエキスパートのデータを矛盾なく説明する報酬を推定することによって，よりエキスパートに近い報酬が推定されることが期待できる．

本論文では，状態遷移確率が異なる複数の環境における逆強化学習問題に対する四つの解法を示す．ここでは，各提案手法の目的を整理する．

### 複数のマルコフ決定過程における線形逆強化学習

複数のマルコフ決定過程における線形逆強化学習の目的は，複数の環境におけるエキスパートの方策からエキスパートの報酬を推定することである．この方法の利点は，複数の環境におけるエキスパート方策が最適となる報酬が推定されることを保証できる点にある．実験では，ベンチマーク問題である風向きが異なる複数の Windy grid world 環境に対して提案法を適用し，複数の環境におけるエキスパートの方策が最適となる報酬を推定できることを確認する．また，推定報酬を学習時と異なる状態遷移確率の環境に転移し，転移先の環境におけるエキスパート方策を学習する実験を行い提案法のエキスパート方策の再現率が，一環境から報酬を推定する既存法と比較して優れていることを確認する．

### 複数のマルコフ決定過程におけるベイジアン逆強化学習

状態遷移確率が異なる複数のマルコフ決定過程におけるベイジアン逆強化学習の目的は，エキスパートが軌跡から報酬を推定することである．ベイズ推定を用いるもう一つの利点としては，設計者が持つ報酬に関する事前知識を事前分布として導入できる点が挙げられる．適切な事前分布を導入することによって，エキスパートの軌跡の数が少ない場合にも，エキスパートに近い報酬が推定できることが期待される．実験では，ベンチマーク問題である風向きが異なる複数の Windy grid world 環境に対して提案法を適用し，複数環境におけるエキスパートの軌跡から報酬を推定できることを確認する．また，提案法が，単一環境から報酬を推定する既存法と比較して，エキスパート方策の再現率が高いことを定量的に示す．

### 複数のマルコフ決定過程におけるミニバッチベイジアン逆強化学習

複数の環境におけるミニバッチベイジアン逆強化学習手法は，複数の環境におけるベイジアン逆強化学習手法の計算量を削減する．具体的には，複数環境のベイジアン逆強化学習の各マルコフ連鎖モンテカルロステップにおける動的計画法の回数を，エキスパートの軌跡が得られた環境の数よりも小さい任意のミニバッチ数に削減する．

実験では，ミニバッチベイジアン逆強化学習手法に関する二つのことを確認する．一つ目は，環境数よりも小さいミニバッチ数を設定しても，適切にエキスパートの報酬が推定できることである．もう一つは，ミニバッチ数を固定した下で，エキスパートの環境数の増加に応じて推定報酬が改善することである．

### 複数のマルコフ決定過程における敵対的最大エントロピー逆強化学習

複数のマルコフ決定過程における敵対的最大エントロピー逆強化学習の目的は，連続状態行動空間の複数の環境におけるエキスパートの軌跡からエキスパートの報酬と方策を推定することである．この連続状態行動空間の問題は線形計画・ベイジアン・ミニバッチベイジアン逆強化学習手法が扱うことができない．実験では，プラントの内部状態を昇温，昇圧する問題を用いて提案法の有用性を示す．具体的には，供給条件が異なる複数のプラントにおけるエキスパートの軌跡に提案法を適用し，複数の供給条件のプラントにおいて適用可能な方策が獲得できることを確認する．そして，提案法を用いて学習した方策を，学習時と異なる供給条件の環境へと転移する実験を通して，提案法の転移の成功確率が既存手法の転移の成功確率よりも高いことを確認する．

## 1.3 論文の構成

本論文の構成を図 1.3 を用いて説明する．

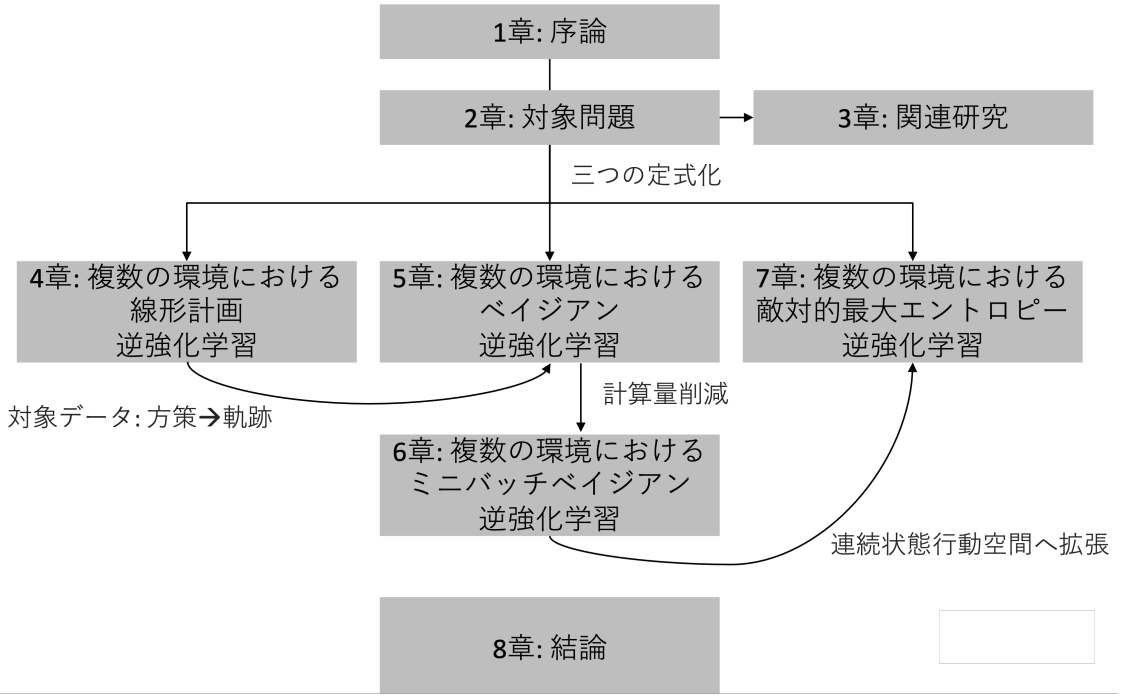



図 1.3: 本論文の構成

本論文は著者がこれまでおこなってきた四つの研究，複数環境における線形計画逆強化学習，ベイジアン逆強化学習，ミニバッチベイジアン逆強化学習，敵対的最大エントロピー逆強化学習について記す．第 2 章では，四つの研究の対象問題を示す．そして第 3 章で四つ研究の関連研究をまとめ，関連研究の課題を示す．第 4 章では，複数環境におけるエキスパート方策から報酬を推定する線形計画逆強化学習を定式化し解法を示す．第 5 章では，複数環境におけるエキスパートの軌跡から報酬を推定するベイジアン逆強化学習を定式化し解法を提案する．. 第 6 章では，5 章で提案した解法の計算量を削減する近似推論の方法を提案する．第 7 章では，逆強化学習の実応用が期待される連続量を入出力とするプラント制御問題に適用可能な敵対的最大エントロピー逆強化学習手法を提案する．最後に，第 8 章で本論文の結論の述べる．

# 第2章　対象問題

本章では，本論文の対象問題を定義する．対象問題の基礎となるマルコフ決定過程，強化学習問題，逆強化学習問題について簡単に整理した後に，状態遷移確率が異なる複数のマルコフ決定過程における逆強化学習問題について記す．

## 2.1　マルコフ決定過程（MDP）

マルコフ決定過程 (Markov Decision Processes) は，エージェントの行動による状態遷移にマルコフ性を仮定した数学モデルである．

マルコフ決定過程 $\mathcal{M} = (E, R)$ は，環境 $E = \langle\, \mathcal{S},\, \mathcal{A},\, T,\, \gamma \,\rangle$ と報酬 $R$ からなる．ここで，$\mathcal{S}$ は状態集合，$\mathcal{A}$ は行動集合，$T(s'|s,a)$ は状態 $s \in \mathcal{S}$ で行動 $a \in \mathcal{A}$ を取ったとき次状態 $s' \in \mathcal{S}$ に遷移する確率である状態遷移確率，$\gamma \in [0,1)$ は割引率，$R : \mathcal{S} \to \mathbb{R}$ は状態 $s \in \mathcal{S}$ の報酬 $r$ を返す報酬関数を表す．

エージェントは時刻 $t$ において状態 $s_t \in \mathcal{S}$ を観測し，自身の方策 $\pi : \mathcal{S} \to \mathcal{A}$ に基づいて行動 $a_t \in \mathcal{A}$ を選択する．その後，時刻 $t+1$ では $s_t$，$a_t$ によって確率的に次状態 $s_{t+1}$ に遷移し，報酬 $R(s_{t+1})$ を得る．

マルコフ決定過程におけるエージェントと環境の関係を図 2.1 に示す．

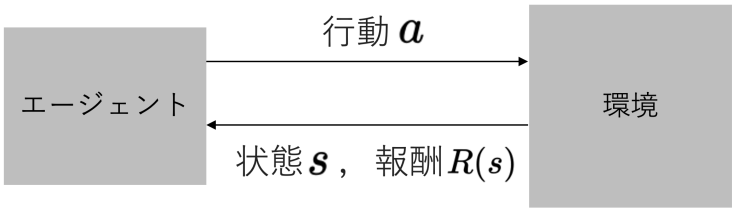



図 2.1: マルコフ決定過程の概略図

エージェントは環境から受け取った状態に基づいて行動を決定する．そして，状態 $s$ で行動 $a$ を選択したときの状態遷移確率 $T(S'|s,a)$ に従って次の状態がサンプルされる．マルコフ決定過程における各エピソードの流れを図 2.2 に示す．

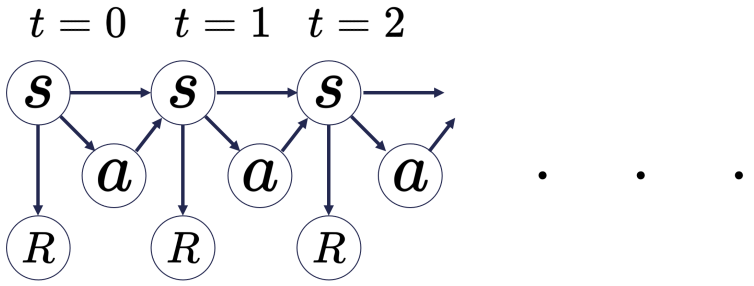



図 2.2: マルコフ決定過程のエピソードの概念図

図 2.2 の矢印は各変数の依存関係を表す．エージェントは状態に基づいて行動を決定するため，状態 $s$ から行動 $a$ に矢印が伸びている．また，報酬 $R$ もエージェントの行動と同様に状態に依存しているため．状態 $s$ から報酬 $R$ に矢印が伸びている．次状態は，直前の状態とエージェントの行動から定まる状態遷移確率に基づいてサンプルされるため，ステップ $t$ の状態 $s$ と行動 $a$ からステップ $t+1$ の状態に矢印が伸びている．

## 2.2 強化学習問題

強化学習問題は，マルコフ決定過程における試行錯誤を通して最適方策を獲得する問題である．最適方策とは，方策 $\pi$ の下で得られる累積報酬の期待値が最大の方策である．

方策 $\pi$ の下で得られる報酬の期待値は，状態価値関数 $V^\pi : \mathcal{S} \to \mathbb{R}$ と行動価値関数 $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ で表される．状態価値 $V^\pi(s)$，行動価値 $Q^\pi(s,a)$ の定義式をそれぞれ式 (2.1)，式 (2.2) に示す．

$$V^\pi(s) = E_\pi \left[ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | s_t = s \right] \tag{2.1}$$

$$Q^\pi(s,a) = E_\pi \left[ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | s_t = s, a_t = a \right] \tag{2.2}$$

ここで，$R_t$ はエージェントが時刻 $t$ に獲得する報酬を指す．状態価値 $V^\pi(s)$ は，方策 $\pi$ の下で状態 $s$ 以降に得られる報酬の期待値，行動価値 $Q^\pi(s,a)$ は，方策 $\pi$ の下で状態 $s$ で行動 $a$ を選択した時に得られる報酬の期待値を示す．

状態価値と行動価値は，それぞれ式 (2.3), 式 (2.4) のベルマン方程式を満たす．

$$V^\pi(s) = R(s) + \gamma \sum_{s'} T(s'|s, \pi(s)) V^\pi(s') \tag{2.3}$$

$$Q^{\pi}(s,a) = R(s) + \gamma \sum_{s'} T(s'|s,a) V^{\pi}(s') \tag{2.4}$$

最適方策 $\pi^*$ の状態価値 $V^{\pi^*}(s)$ は任意の状態 $\forall s \in \mathcal{S}$ において最大値を取る．最適方策 $\pi^*$ が満たす条件，最適ベルマン方程式を式 (2.5) に示す．

$$V^{\pi^*}(s) = R(s) + \gamma \max_{\pi} \sum_{s'} T(s'|s,\pi(s)) V^{\pi}(s') \quad \forall s \in \mathcal{S} \tag{2.5}$$

最適方策 $\pi^*$ の行動価値 $Q^{\pi^*}(s,a)$ を式 (2.6) に示す．

$$Q^{\pi^*}(s,a) = R(s) + \gamma \sum_{s'} T(s'|s,a) V^{\pi}(s') \tag{2.6}$$

マルコフ決定過程における最適方策は，最適行動価値 $Q^{\pi^*}(s,a)$ を用いて式 (2.7) のように表される．

$$\pi^*(s) = \max_{a} Q^{\pi^*}(s,a) \tag{2.7}$$

強化学習問題は，式 (2.7) に示す最適方策を学習する問題である．

## 2.3 逆強化学習問題

逆強化学習問題とは，エキスパートがある環境 $E$ において生成したデータセット $D$，またはエキスパートの方策 $\pi_{\text{exp}}$ を所与として，報酬 $R$ を推定する問題である．推定報酬は，エキスパート方策が最適となる報酬である．エキスパート方策の最適性を式 (2.8) に示す．

$$\pi_{\text{exp}} \in \arg\max_{\pi} \mathbb{E}_{\pi} \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t) \right] \tag{2.8}$$

式 (2.8) は，次の不等式と等価である．

$$\mathbb{E}_{\pi_{\text{exp}}} \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t) \right] \geq \max_{\pi} \mathbb{E}_{\pi} \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t) \right] \tag{2.9}$$

逆強化学習問題は，式 (2.9) を満たす報酬を推定する問題であると言える．エキスパート方策ではなく，エキスパートが生成したデータセット $D$ から報酬を推定する場合には，式 (2.9) ではなく式 (2.10) を満たす報酬を推定する．

$$\mathbb{E}_{D} \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t) \right] \geq \max_{\pi} \mathbb{E}_{\pi} \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t) \right] \tag{2.10}$$

逆強化学習問題は，式 (2.9), (2.10) を満たす報酬を推定する問題である．

式 (2.9) を満たす報酬は複数存在し，式 (2.11) に示す全状態で一定の値を持つ自明な報酬も不等式 (2.9) を満たす．

$$R(s) = c \quad \forall s \in \mathcal{S} \tag{2.11}$$

したがって，不等式 (2.9) を満たす非自明な報酬を推定するためには，マージン最大化 [Ng 00, Abbeel 04]，ベイズ推定 [Ramachandran 07]，最大エントロピー原理 [Ziebart 08] などの方法を導入する必要がある．

## 2.4 複数のマルコフ決定過程における逆強化学習問題

本研究が対象とする問題は状態遷移確率が異なる $M$ 個の環境 $\{E_m\}_{m=1}^{M} = \{\langle\, \mathcal{S}, \mathcal{A}, T_m, \gamma \,\rangle\}_{m=1}^{M}$ と，各環境 $E_m$ におけるエキスパートの方策 $\{\pi_{\mathrm{exp,m}}\}_{m=1}^{M}$，またはデータセット $\{D_m\}_{m=1}^{M}$ からエキスパートの報酬を推定する問題である．推定報酬は，単一環境における逆強化学習の制約条件式 (2.8) を複数環境に拡張した式 (2.12) を満たす必要がある．

$$\pi_{\mathrm{exp,m}} \in \arg\max_{\pi} \mathbb{E}_{\pi,T_m}\left[\sum_{t=0}^{\infty}\gamma^t R\left(s_t\right)\right] \quad \forall m \in \{1,\ldots,M\} \tag{2.12}$$

式 (2.12) を満たす報酬は以下の不等式を満たす．

$$\mathbb{E}_{\pi_{\mathrm{exp,m}},T_m}\left[\sum_{t=0}^{\infty}\gamma^t R\left(s_t\right)\right] \geq \max_{\pi_m} \mathbb{E}_{\pi_m,T_m}\left[\sum_{t=0}^{\infty}\gamma^t R\left(s_t\right)\right] \quad \forall m \in \{1,\ldots,M\}, \tag{2.13}$$

$$\mathbb{E}_{D_m}\left[\sum_{t=0}^{\infty}\gamma^t R\left(s_t\right)\right] \geq \max_{\pi_m} \mathbb{E}_{\pi_m,T_m}\left[\sum_{t=0}^{\infty}\gamma^t R\left(s_t\right)\right] \quad \forall m \in \{1,\ldots,M\} \tag{2.14}$$

状態遷移確率が異なる複数環境における逆強化学習問題は，式 (2.13), (2.14) を満たす報酬を推定する問題である．状態遷移確率が異なる複数環境における逆強化学習問題の制約条件の数は，単一の環境における逆強化学習問題の制約条件式 (2.9), (2.10) と比較して多い．しかし依然として式 (2.11) に示した自明な報酬も複数環境における逆強化学習問題の制約条件式 (2.13), (2.14) を満たすことに注意されたい．したがって，複数環境における逆強化学習問題もマージン最大化 [Ng 00, Abbeel 04]，ベイズ推定 [Ramachandran 07]，最大エントロピー原理 [Ziebart 08] などの方法を導入する必要がある．

# 第3章　関連研究

本章では，本論文の関連研究である逆強化学習と，マルコフ決定過程における転移学習についてまとめる．

## 3.1　逆強化学習

### 単一の環境における逆強化学習

逆強化学習は，マルコフ決定過程におけるエキスパートの方策が最適となる報酬を推定する問題である [Russell 98]．この逆強化学習問題に対するアプローチは大きく三つに分けられる．一つ目は，エキスパートの方策と，それ以外の方策の割引期待獲得報酬の差を最大化するマージン最大化 [Ng 00, Abbeel 04, Ratliff 06] である．二つ目は，ある報酬の下でエキスパートの軌跡が得られる確率をモデル化し，ベイズ推定を行うアプローチである [Ramachandran 07, Choi 11, Choi 12, Choi 13, Surana 14a, Surana 14b]．三つ目は，最大エントロピー原理 [Grünwald 04a] に基づいてエキスパートに軌跡から報酬を推定する最大エントロピー逆強化学習 [Ziebart 08, Wulfmeier 15, Finn 16, Fu 18] である．図 3.2 に三つのアプローチ，マージン最大化，ベイズ推定，エントロピー最大化の代表的な手法を示す．

表 3.1: 代表的な逆強化学習手法のアプローチによる分類

| アプローチ | 代表的な逆強化学習手法 |
|---|---|
| マージン最大化 | 線形計画逆強化学習 (LPIRL)[Ng 00],<br>Apperenticeship Learning[Abbeel 04],<br>Max Margin Planning [Ratliff 06] |
| ベイズ推定 | ベイジアン逆強化学習（BIRL）[Ramachandran 07],<br>Map IRL[Choi 11] |
| 最大エントロピー原理 | 最大エントロピー逆強化学習 (Max Ent IRL)[Ziebart 08],<br>Deep Max Ent IRL[Wulfmeier 15],<br>Guided Cost Learning [Finn 16],<br>Adversarial IRL(AIRL)[Fu 18] |

マージン最大化は，式 (3.1) に示すエキスパートの期待獲得報酬とエキスパート以外の方策の期待獲得報酬の差を最大化するアプローチである．

$$\underset{R}{\text{maximize}} \quad \mathbb{E}_{\pi_{\text{exp}}}\left[\sum_{t=0}^{\infty}\gamma^t R\left(s_t\right)\right] - \max_{\pi\in\Pi\setminus\pi_{\text{exp}}} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t R\left(s_t\right)\right] \tag{3.1}$$

Ng らは，エキスパートの方策が最適となる制約の下で，式 (3.1) を最大化する問題を線形計画問題として定式化した [Ng 00]．Apprenticeship learning[Abbeel 04] は，この問題を報酬関数を状態の特徴量 $\phi(s)$ を用いて線形関数，

$$R(s) = \theta \cdot \phi(s) \tag{3.2}$$

でモデル化し，射影法を応用した解法を示した．また，Max margin planning[Ratliff 06] は，Apprenticeship learning[Abbeel 04] と同様に報酬を線形関数でモデル化し，サポートベクターマシンを用いた解法を示した．

ベイジアン逆強化学習は，ある報酬の下でエキスパートの軌跡が得られる確率を行動価値式 (2.2) を用いてモデル化し，ベイズ推定を行うアプローチである．逆強化学習問題にベイズ推定を導入する利点の一つは推定の頑健性にある．具体的には，エキスパートの軌跡のデータの中に準最適な方策から生成された軌跡を含む場合にも推定結果が大きく変わらないという性質がある．実問題においては，エキスパートが常に最適な方策（例: 目標値から全くずれない機械の制御）を持っていない場合があるため，このような頑健性は逆強化学習問題を解く上で重要である．しかし，ベイズアン逆強化学習はマルコフ連鎖モンテカルロ法を用いる必要があり，その計算量が実応用上の課題となっている．

最大エントロピー逆強化学習は，ある報酬の下でエキスパートの軌跡が得られる確率を最大エントロピー原理を用いてモデル化し，報酬を推定するアプローチである．Ziebart らは最大エントロピー原理を導入することによって，対象問題が目的関数が凸関数で線形制約を持つ制約付き最適化問題になることを示し勾配法を用いた解法を提案した [Ziebart 08, Ziebart 10]．勾配法が逆強化学習問題に適用可能になったことによって，関数の表現能力が高いモデルとして知られるニューラルネットワーク [Mnih 13] を使用できるようになり，深層エントロピー最大化逆強化学習が提案された [Wulfmeier 15]．Finn らは，最大エントロピー逆強化学習と，敵対的生成モデルの学習法として知られる Generative Advesarial Network[Goodfellow 14] との定式化の類似性を指摘 [Finn 17] し，Ho らが最大エントロピー逆強化学習問題がルジャンドル変換 [Rockafellar 70] によって敵対的学習問題となることを示した [Ho 16]．そして，この敵対的学習問題に対して Fu らは敵対的最大エントロピー逆強化学習手法である Adversarial Inverse Reinforcement Learning (AIRL)[Fu 18] を提案した．

図 3.2 に三つのアプローチ，マージン最大化，ベイズ推定，エントロピー最大化の利点を整理する．

表 3.2: 代表的な逆強化学習手法のアプローチによる分類

| アプローチ | 準最適なエキスパートデータ | 勾配法の適用 |
|---|---|---|
| マージン最大化 | | |
| ベイズ推定 | ✓ | |
| 最大エントロピー原理 | ✓ | ✓ |

マージン最大化は逆強化学習問題に対して解法を示した点で逆強化学習の発展に対する貢

献は大きい一方で，準最適なエキスパートのデータを扱うことができないという課題があった．この問題に対してベイジアン逆強化学習は，エキスパートの軌跡が生成される過程を確率モデルを用いてベイズ推定の問題に定式化した．ベイジアン逆強化学習に必要なマルコフ連鎖モンテカルロ法の計算量という課題に対して，最大エントロピー逆強化学習は凸性などの優れた性質を問題を定式化し勾配法を用いた解法を示した．

### 複数の環境または報酬を扱う逆強化学習

本論文で単一の環境における逆強化学習を複数の環境におけるエキスパートのデータを扱うことができるように拡張する．そこで，本論文の関連研究として，単一の環境における逆強化学習を拡張した手法について整理し，本論文との違いを整理する．

表 3.3 に，環境モデルの数と，推定報酬の数に基づいた逆強化学習の分類を示す．

表 3.3: 環境と推定報酬の数に基づく逆強化学習手法の分類

| | | 推定報酬 | |
|---|---|---|---|
| | | 単一 | 複数 |
| 環境のモデル | 単一 | 線形計画逆強化学習 [Ng 00],<br>ベイジアン逆強化学習<br>[Ramachandran 07],<br>最大エントロピー逆強化学習<br>[Ziebart 08] | ALMI [Babes 11],<br>BNIRL sMDP [Surana 14a] |
| | 複数 | Repeated IRL [Amin 17],<br>本論文 | Meta IRL [Li 17] |

ある環境で，異なる報酬に従う複数のエキスパートのデータが混在するデータセットから，各エキスパートの報酬を推定する方法には，EM アルゴリズムを用いた逆強化学習 (ALMI) [Babes 11], ノンパラメトリックベイズを用いた逆強化学習 (BNIRL sMDP)[Surana 14a] がある．これらの手法は，状態遷移確率が単一の環境で,（異なる報酬関数を持つ）複数のエキスパートが生成したデータを扱う．これに対して，提案法は，状態遷移確率が異なる複数の環境を扱う点が，ALMI [Babes 11], BNIRL sMDP [Surana 14a] と異なる．

複数の環境のモデルを扱い，複数の報酬を推定する逆強化学習手法に Meta IRL [Li 17] がある．Li らは，各エキスパートの推定報酬の差を小さくするペナルティ項を最大エントロピー逆強化学習 [Ziebart 08] の目的関数に追加し，各エキスパート間で類似した複数の報酬を推定する方法を提案している．各エキスパート間で類似した報酬を推定することによって，各環境におけるエキスパートの軌跡のデータ量が少ない時に，推定報酬のエキスパートの軌跡への過学習を抑制している．

Meta IRL は，複数の環境のモデルを扱う点で提案法と類似しているが，各環境におけるエキスパートの報酬に対する仮定が異なる．Meta IRL は，各環境におけるエキスパートが異なる報酬を持つという仮定の下で，複数の報酬を推定する．そのため，Meta IRL によっ

て複数の環境のデータと無矛盾な報酬が推定されることはない．一方，提案法は，各エキスパートが同じ報酬を持つと仮定し，推定報酬が複数の環境のデータと無矛盾であることが理論的に保証される．

複数の環境から単一の報酬を推定する手法に Repeated IRL [Amin 17] がある．Repeated IRL は，複数の環境に対する情報が逐次的に与えられる場合の報酬更新方法として提案されている．したがって，その推定報酬は環境が与えられる順番と報酬更新の閾値（ハイパーパラメータ）に依存する．一方，本論文は逐次的な更新ではなく，複数の環境におけるエキスパートのデータがバッチで与えられる場合を対象とし，複数の環境間の無矛盾性を保証することを目的とする点で立場が異なる．また，Repeated IRL は，複数の環境における準最適な方策や確率的方策から生成されたエキスパートの軌跡を扱うことができないが，本論文が提案する複数の環境におけるベイジアン・ミニバッチベイジアン・敵対的最大エントロピー逆強化学習は，準最適な方策や確率的方策から生成されたエキスパートの軌跡から報酬を推定することができる．

## 3.2 マルコフ決定過程における転移学習

提案法を含む上記の逆強化学習手法は，エキスパートの方策や軌跡から報酬を推定し，推定報酬を転移することによって，転移先の環境におけるエキスパート方策を学習できる．学習環境と異なる環境におけるエキスパート方策を学習する方法には，報酬を推定し推定する方法ではなく，方策を転移するアプローチもある．以下で，学習環境と異なる環境に方策を転移するアプローチについて述べる．

Behavioral Cloning[Pomerleau 89] や Dagger[Ross 11] は，エキスパートの軌跡からエキスパートの方策を学習する模倣学習法である．エキスパートの報酬を推定することなく，直接方策を学習する Behavioral Cloning の目的関数を式 (3.3) に示す．

$$L_{\mathrm{bc}}(\theta) = \sum_{(s,a)\in\mathcal{D}} \log \pi_\theta(s,a) \tag{3.3}$$

式 (3.3) は，エキスパートの状態行動対 $(s,a)$ に対する尤度最大化問題である．尤度最大化問題は，ニューラルネットワーク [Goodfellow 16] を用いた一般的な教師あり学習 [Crisci 12] で解くことができる．このアプローチにはマルコフ決定過程における試行錯誤が不要であるという利点があるものの，エキスパートのデータが存在しない状態における方策を適切に学習できないという課題がある [Shimodaira 00]．一方，逆強化学習は環境における試行錯誤を通してエキスパートのデータが存在しない状態においても推定報酬に対する最適方策を学習できる．このように，方策を直接学習する Behavioral Cloning[Pomerleau 89] や Dagger[Ross 11] などの手法には課題が存在するが，Behavioral Cloning や Dagger で学習した方策を Meta-Learning[Thrun 98, Finn 17, Duan 17, Clavera 19] や Domain Adaptation[Daume III 06, Jiang 08, Pan 17]，Fine-tuning[Lamblin 10, Bengio 12, Yosinski 14] を適用することによっ

て転移可能な方策を学習するアプローチも考えられる．しかし，これらのアプローチは転移先の環境におけるエキスパートの軌跡を要するという欠点がある．一方，逆強化学習は推定報酬を転移する際，転移先の環境におけるエキスパートの軌跡は不要である．

文献 [Yan 17] では，Dagger と Domain Randomization[Tobin 17, Peng 18, Andrychowicz 18] を組み合わせることによって，転移先のエキスパートの軌跡を用いることなく転移可能な方策の学習法を提案している．ただし，Domain Randamization は学習に用いる環境と転移先の環境の状態遷移確率が類似している必要がある．これに対して,「報酬の転移」では環境の状態遷移確率の類似性を仮定する必要はない．エキスパートの報酬が推定できれば，任意の状態遷移確率の環境におけるエキスパートの方策を学習できるからである．

# 第4章　複数のマルコフ決定過程における線形計画逆強化学習

本章では,「状態遷移確率が異なる複数の環境下で得られるエキスパート方策」から報酬を推定する逆強化学習手法を提案する．具体的には，複数の環境から報酬を推定する問題を線形計画問題として定式化し,「状態遷移確率だけが異なる複数の環境のモデル」と「各環境でのエキスパートの方策」から，各環境におけるエキスパート方策が推定された報酬に対して最適方策であることを保証する．

実験では，ある一定の確率で風が吹く方向に遷移する Windy grid world 環境を用いる．風向きが異なる複数の Windy grid world 環境と，各環境におけるエキスパート方策を用意し，既存手法の推定報酬が，各環境におけるエキスパート方策と矛盾することを確認する．また，提案法の推定報酬が，各環境におけるエキスパート方策に無矛盾であることを示す．

## 4.1　準備

本章では，強化学習，線形計画法の基礎と線形計画逆強化学習（Linear Programing Inverse Reinforcement Learning; LPIRL）[Ng 00] について述べる．

### 4.1.1　強化学習

強化学習はマルコフ決定過程における最適方策を求める方法である．ここでは，離散状態行動空間の状態遷移確率が既知である場合に用いられる Value Iteration と，状態遷移確率を明示的に利用せず環境との試行錯誤によって最適方策を学習する Q 学習について説明する．

Value Iteration のアルゴリズムを Algorithm 1 に示す．

**Algorithm 1** Value Iteration

**INPUT:** Transition matrix $T$, Reward $R$,threshold $\theta$

**OUTPUT:** Deterministic policy, $\pi \simeq \pi^*$, such that $\pi(s) = \arg\max_a R(s) + \gamma \sum_{s'} T(s'|s,a)V(s')$

Initialize $V(s) \quad \forall s \in \mathcal{S}$, arbitrarily excpet that $V(terminal) = 0$
**loop**
    $\Delta \Leftarrow 0$
    **for** e **do**ach state $s \in \mathcal{S}$
        $v \leftarrow V(s)$
        $V(s) \leftarrow \max_a R(s) + \gamma \sum_{s'} T(s'|s,a)V(s')$
        $\Delta \Leftarrow \max_a(\Delta, |v - V(s)|)$
    **end for**
**end loop** if $\Delta \leq \theta$

Value Iteration は，状態価値を報酬と次状態の状態価値に基づいて更新することによって，式 (2.7) を満たす状態価値を求めている．最適ベルマン法定式を満たす状態価値は，最適方策の状態価値になっている．

状態遷移確率を明示的に利用せず，環境との試行錯誤によって最適方策を学習する Q 学習のアルゴリズムを 1 に示す．

**Algorithm 2** Q learning

**INPUT:** step size $\alpha \in (0, 1]$, small $\epsilon > 0$

**OUTPUT:** State action values $Q(s,a)\forall(s,a) \in \mathcal{S} \times \mathcal{A}$

Initialize $Q(s,a)\forall(s,a) \in \mathcal{S} \times \mathcal{A}$, arbitrarily excpet that $Q(terminal, \cdot) = 0$
**for** each episode **do**
    Initalize $S$
    **loop** each step of episode:
        Choose action $A$ from $S$ using policy derived from Q(e.g., $\epsilon$-greedy)
        Take action $A$, observe $R, S'$
        $Q(S,A) \leftarrow Q(S,A) + \alpha \left[R + \gamma \max_a Q\left(S', a\right) - Q(S,A)\right]$
        $S \leftarrow S'$
    **end loop** if $S$ is terminal
**end for**

### 4.1.2 線形計画法

線形計画問題は，線形の目的関数と制約を持つ最適化問題であり，最大化問題，最小化問題，等式制約付き問題，不等式制約付き問題などの様々な形の問題が存在する．これら様々な線形計画問題において，線形等式制約と非負制約の下で線形の目的関数を最小化する問題を線形計画問題の標準形と言う．線形計画問題の標準形を以下に示す．

$$\text{minimize}: \quad w = c_1x_1 + c_2x_2 \qquad + \cdots c_nx_n \tag{4.1}$$

$$\text{s.t.}: \quad a_{11}x_1 + a_{12}x_2 \quad + \cdots a_{1n}x_n = b_1 \tag{4.2}$$

$$a_{21}x_1 + a_{22}x_2 \quad + \cdots a_{2n}x_n = b_2 \tag{4.3}$$

$$\vdots$$

$$a_{m1}x_1 + a_{m2}x_2 \quad + \cdots a_{mn}x_n = b_m \tag{4.4}$$

$$x_i \quad \geq 0 (i = 1, \cdots, n) \tag{4.5}$$

ここで，係数 $a_{ij}$

ここで，$a_{ij}, b_i, c_i$ は定数で，$m \leq n$ である．標準形の制約条件は等式であるが，対象問題が不等式制約条件，

$$\sum_{j=1}^{n} a_{ij}x_j \leq b_i \tag{4.6}$$

を含む場合には，非負のスラック変数 $x_{n+1}$ を導入することによって等式制約と非負制約に変換できる．不等式制約式 (4.6) の等式制約と非負制約への変換を以下に示す．

$$\sum_{j=1}^{n} a_{ij}x_j + x_{n+i} = b_i \tag{4.7}$$

$$x_{n+i} \geq 0 \tag{4.8}$$

非負制約を持たない変数 $x_i$ は，2 つの非負の変数の差として表すことができる．自由変数 $x_i$ の変換を以下に示す．

$$x_i = x_i^+ - x_i^- \tag{4.9}$$

$$x_i^+ \geq 0 \tag{4.10}$$

$$x_i^- \geq 0 \tag{4.11}$$

線形計画問題の標準形が与えられたとき，以下のことが成り立つことが知られている．

1. 実行可能解が存在するとき，実行可能基底解が存在する．
2. 最適解が存在するならば，実行可能基底解の中に最適解が存在する．

ここで，実行可能解は線形計画問題の制約を満たす解で，実行可能基底解は基底解のうち全ての変数が非負の解である．線形計画問題の基底解は，等式制約の係数行列 $A$ から $m$ 本の独立な列ベクトルを並べた基底行列 $B \in \mathbb{R}^{m \times m}$ を用いて定義される．基底行列に対応する変数のベクトルを基底変数ベクトル $\boldsymbol{x}_B$，残りの変数を非基底変 $\boldsymbol{x}_N$，非基底変数に対応する列ベクトルの行列を $N \in \mathbb{R}^{m \times (n-m)}$ とおけば，等式制約 $A\boldsymbol{x} = \boldsymbol{b}$ は $B\boldsymbol{x}_B + N\boldsymbol{x}_N = \boldsymbol{b}$ と記述できる．そして，非基底変数を零ベクトルとおけば，基底変数は $\boldsymbol{x}_B = B^{-1}\boldsymbol{b}$ と一意に定まる．このように定まる解 $\boldsymbol{x} = [\boldsymbol{x}_B, \boldsymbol{x}_N]^T$ を基底解と呼ぶ．

次に，線形計画法の代表的な解法である単体法について説明する．単体法は 1 組の実行可能基底解が与えられたとき，目的関数の値が減少する新しい実行可能解を効率的に求め，これを繰り返すことによって最適な実行可能基底解を求める方法である．

単体法のアルゴリズムを Algorithm 3 に示す．

**Algorithm 3** 単体法

**INPUT:** 任意の実行可能基底形式

**OUTPUT:** 線形計画問題の最適解

1: 相対費用係数 $\bar{\boldsymbol{c}}_i$ が非負ならば，最適解であるため終了．非負でなければ添字番号 $q = \underset{i}{\operatorname{argmin}}\ \bar{c}_i$ を求める．

2: $\bar{\boldsymbol{a}}_q \leq 0$ ならば，目的関数が下に有界でないので終了．$\bar{a}_{iq} > 0$ となる成分が存在するならば，$p = \underset{i}{\operatorname{argmin}}\ \frac{\bar{b}_i}{\bar{a}_{iq}}$ を求める．

3: $\bar{a}_{pq}$ をピボットとする掃き出しを実行し，$x_q$ を基底変数とする新たな実行可能基底形式を作成．

4: ステップ 1 へ戻る．

Algorithm 3 内の変数 $\bar{\boldsymbol{a}}, \bar{\boldsymbol{b}}, \bar{\boldsymbol{c}}$ の定義を以下に示す．

$$\bar{\boldsymbol{a}}_i = B^{-1}\boldsymbol{a}_i \quad (i = m+1, \cdots, n), \tag{4.12}$$

$$\bar{\boldsymbol{b}} = B^{-1}\boldsymbol{b}, \tag{4.13}$$

$$\bar{\boldsymbol{c}_N} = \boldsymbol{c}_N - (B^{-1}A_N)^T\boldsymbol{c}_B, \tag{4.14}$$

ここで $A_N$ は非基底変数に対応する制約条件の係数行列，$\boldsymbol{c}_N$ は非基底変数に対応する目的関数の係数ベクトルである．

### 4.1.3 線形計画逆強化学習

逆強化学習は，環境 $E$ から報酬 $R$ を除いた環境のモデル $E$ と，エキスパートの方策 $\pi_{\text{exp}}$ や軌跡を所与として，エキスパート方策 $\pi_{\text{exp}}$ が最適となる報酬 $R$ を推定する．Ng らは，エキスパートの報酬を推定する線形計画問題を三つ定式化した [Ng 00]．これら三つの定式化は，状態行動空間とエキスパートのデータに対する前提が異なる．一つ目は，離散状態行動空間の環境のモデル $E$ と，エキスパートの方策 $\pi_{\text{exp}}$ を所与とする定式化，二つ目は，連続状態行動空間の環境のモデル $E$ と，エキスパート方策 $\pi_{\text{exp}}$ を所与とする定式化，三つ目は，エキスパートの軌跡が所与の場合の定式化である．

以下で，提案法の基礎となる Ng らが提案した一つ目の定式化（離散状態行動空間の環境のモデル $E$ と，エキスパートの方策 $\pi_{\text{exp}}$ を所与とする定式化）について述べる．一つ目の定式化では，各状態の報酬を独立に推定するものの，二つ目と三つ目の定式化は，報酬関数を線形関数で近似する．以下で，離散状態行動空間の環境のモデル $E$ と，エキスパートの

方策 $\pi_{\text{exp}}$ を所与として，エキスパート方策 $\pi_{\text{exp}}$ が最適となる報酬 $R$ を推定する線形計画問題の定式化（LPIRL）を示す．

Ng ら [Ng 00] は，環境のモデル $E$ とエキスパート方策 $\pi_{\text{exp}}$ を所与とし，エキスパートの報酬 $R$ を推定する問題を，線形計画問題として定式化した．

Ng らの定式化を式 (4.15), (4.16) に示す．

$$\text{maximize}: \quad \sum_{i=1}^{|\mathcal{S}|} \min_{a \in \mathcal{A} \backslash a_1} \{(\boldsymbol{T}_{a_1}(i) - \boldsymbol{T}_a(i))(\boldsymbol{I} - \gamma \boldsymbol{T}_{a_1})^{-1} \boldsymbol{R}\} - \lambda ||\boldsymbol{R}||_1 \quad (4.15)$$

$$\text{subject to}: \quad (\boldsymbol{T}_{a_1} - \boldsymbol{T}_a)(\boldsymbol{I} - \gamma \boldsymbol{T}_{a_1})^{-1} \boldsymbol{R} \succeq 0 \quad \forall a \in \mathcal{A} \backslash a_1 \quad (4.16)$$

$$|\boldsymbol{R}_i| \leq R_{\max}, \quad i = 1, \ldots, |\mathcal{S}| \quad (4.17)$$

ここで，$a_1$ は状態 $s$ におけるエキスパートの行動 $\pi_{\text{exp}}(s)$ を，$|\mathcal{S}|$ は状態数を指す．状態遷移行列 $\boldsymbol{T}_a$ は状態遷移確率 $T(s_m|a, s_i)$ を $(i, j)$ 成分とする $|\mathcal{S}| \times |\mathcal{S}|$ 行列，状態遷移ベクトル $T_a(i)$ は $\boldsymbol{T}_a$ の第 $i$ 行ベクトルを表す．$\lambda$ は報酬の総量を調節するペナルティ係数である．

式 (4.15) に示す目的関数の第一項はエキスパートの行動と 2 番目によい行動の行動価値の差の総和である．式 (4.15) の第二項は，報酬の総量に比例するペナルティで，報酬の総量を調節する役割を果たす．

式 (4.16) は，エキスパート方策 $\pi_{\text{exp}}$ が報酬 $R$ に対して最適方策となることを保証する制約条件であり，式 (4.18) と表すことができる．

$$Q^{\pi_{\text{exp}}}(s, a_1) \geq Q^{\pi_{\text{exp}}}(s, a) \quad \forall a \in \mathcal{A} \backslash a_1, \ \forall s \in \mathcal{S} \quad (4.18)$$

式 (4.18) を満たす報酬の下で，エキスパート方策 $\pi_{\text{exp}}$ は最適方策である．なぜなら，式 (4.18) を満たす報酬の下では，任意の状態におけるエキスパートの行動 $a_1$ の行動価値 $Q(s, a_1)$ は，それ以外の行動 $a \in A \backslash a_1$ の行動価値 $Q(s, a)$ 以上の値をとるからである．式 (4.18) と式 (4.16) は等しいため，式 (4.16) を満たす報酬の下でエキスパート方策 $\pi_{\text{exp}}$ は最適方策である．

一方，式 (4.16) を満たさない報酬の下では，エキスパート方策 $\pi_{\text{exp}}$ は最適方策でない．そのため，式 (4.16) を満たすことは，推定報酬 $\hat{R}$ がエキスパートの報酬と一致する必要条件である．

### 状態と行動に依存する報酬 $R(s, a)$ の推定

式 (4.15), (4.16), (4.17) からなる線形計画問題は，状態に依存し，行動に依存しない報酬 $R : \mathcal{S} \to \mathbb{R}$ を推定する定式化である．

状態と行動の両方に依存する報酬 $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ を推定する線形計画問題も，エキスパート方策 $\pi_{\exp}$ が報酬 $R$ に対して最適方策となることを保証する制約条件を導出することによって定式化できる．以下に，制約条件の導出を示す．はじめに，状態と行動の両方に依存する報酬の下での行動価値は式 (4.19) と定義される．

$$Q^{\pi}(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a) V^{\pi}(s') \tag{4.19}$$

各状態で行動 $a$ をとった時に得られる報酬 $R(s,a)$ を要素とするベクトルを $\boldsymbol{R}^a$ とおけば，式 (4.18), (4.19) から，エキスパート方策 $\pi_{\exp}$ が報酬 $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ に対して最適方策となることを保証する制約条件，式 (4.20) が導かれる．

$$\begin{aligned}
&R(s,a_1) + \sum_{s'} T(s'|s,a_1) V^{\pi_{\exp}}(s') \geq R(s,a) + \sum_{s'} T(s'|s,a) V^{\pi_{\exp}}(s') \\
&\qquad\qquad \forall a \in \mathcal{A}\backslash a_1,\ \forall s \in \mathcal{S} \\
\Leftrightarrow &\boldsymbol{R}^{a_1} + \gamma \boldsymbol{T}_{a_1} \boldsymbol{V}^{\pi_{\exp}} \succeq \boldsymbol{R}^{a} + \gamma \boldsymbol{T}_{a} \boldsymbol{V}^{\pi_{\exp}} \quad \forall a \in \mathcal{A}\backslash a_1 \\
\Leftrightarrow &\boldsymbol{R}^{a_1} + \gamma \boldsymbol{T}_{a_1} \left(\boldsymbol{I} - \gamma \boldsymbol{T}_{a_1}\right)^{-1} \boldsymbol{R}^{a_1} \\
&\qquad \succeq \boldsymbol{R}^{a} + \gamma \boldsymbol{T}_{a} \left(\boldsymbol{I} - \gamma \boldsymbol{T}_{a_1}\right)^{-1} \boldsymbol{R}^{a_1} \quad \forall a \in \mathcal{A}\backslash a_1 \\
\Leftrightarrow &\boldsymbol{R}^{a_1} - \boldsymbol{R}^{a} + \gamma (\boldsymbol{T}_{a_1} - \boldsymbol{T}_{a}) (\boldsymbol{I} - \gamma \boldsymbol{T}_{a_1})^{-1} \boldsymbol{R}^{a} \succeq 0 \\
&\qquad\qquad \forall a \in \mathcal{A}\backslash a_1
\end{aligned} \tag{4.20}$$

式 (4.20) はエキスパート方策 $\pi_{\exp}$ が報酬 $R$ に対して最適方策となることを保証する制約条件，式 (4.16) と対応する．式 (4.20) に基づいて，目的関数，式 (4.16) と対応する制約条件，各状態行動対の報酬の最大値を定める制約条件，からなる線形計画問題を解くことによって，状態と行動の両方に依存する報酬を推定できる．

## 4.2 対象問題

はじめに，本章が対象とする「複数の環境間の状態遷移確率の差異」が生じる要因を具体例を用いて説明する．次に，状態遷移確率が異なる複数の環境のモデルと，各環境におけるエキスパートの方策から報酬を推定する対象問題を，マルコフ決定過程の枠組みで定式化する．

本章では，状態遷移確率が異なる複数の環境におけるエキスパートが，ある一つの報酬 $R$ に従う仮定の下で，エキスパートの報酬 $R$ を推定する．具体的には，状態遷移確率のみ異なる環境のモデルの集合 $\{E_m\}_{m=1}^{M}$ と，各環境 $E_m$ におけるエキスパート方策 $\pi_{\exp}^m$ の集合 $\Pi_{\exp} = \{\pi_{\exp}^m\}_{j=1}^{M}$ を所与として，各環境 $E_m$ におけるエキスパート方策 $\pi_{\exp}^m \in \Pi_{\exp}$ が最適となる報酬 $R$ を推定する．

式 (2.7) より，状態遷移確率 $T_m$ の環境のモデル $E_m$ においてエキスパート方策 $\pi^m_{\exp}$ が最適となる報酬 $R$ は，式 (4.21) を満たす．

$$V^{\pi^m_{\exp}}(s) = R(s) + \gamma \max_{\pi} \sum_{s'} T_m(s'|s, \pi(s)) V^{\pi}(s') \ \forall s \in \mathcal{S} \tag{4.21}$$

以上のことから，本章の対象問題は，所与の $E_m$ と $\pi^m_{\exp}$ のすべての組 $(E_m, \pi^m_{\exp})$ が，式 (4.21) を満たす報酬 $R$ の推定である．

## 4.3 アプローチ

提案法は，LPIRL[Ng 00] を拡張し，状態遷移確率が異なる複数の環境のモデルと，各環境におけるエキスパート方策から報酬を推定する．

提案法の定式化を以下に示す．

$$\underset{R}{\text{maximize}}: \quad \sum_{m=1}^{M} \sum_{i=1}^{|\mathcal{S}|} \min_{a^m \in \mathcal{A} \backslash a^m_1} \{(\boldsymbol{T}^m_{a^m_1}(i) - \boldsymbol{T}^m_a(i))(\boldsymbol{I} - \gamma \boldsymbol{T}^m_{a^m_1})^{-1} \boldsymbol{R}\} - \lambda ||\boldsymbol{R}||_1 \tag{4.22}$$

$$\begin{aligned} \text{subject to}: \quad (\boldsymbol{T}^1_{a^1_1} - \boldsymbol{T}^1_a)(\boldsymbol{I} - \gamma \boldsymbol{T}^1_{a^1_1})^{-1} \boldsymbol{R} &\succeq 0 \\ \forall a \in \mathcal{A} &\backslash a^1_1 \\ &\vdots \\ (\boldsymbol{T}^M_{a^M_1} - \boldsymbol{T}^M_a)(\boldsymbol{I} - \gamma \boldsymbol{T}^M_{a^M_1})^{-1} \boldsymbol{R} &\succeq 0 \\ \forall a \in \mathcal{A} &\backslash a^M_1 \end{aligned} \tag{4.23}$$

$$|\boldsymbol{R}_i| \leq R_{\max}, \quad i = 1, \ldots, |\mathcal{S}| \tag{4.24}$$

式 (4.22) の第一項は各環境における LPIRL の目的関数，式 (4.15) の総和である．式 (4.22) の第二項は，報酬の総量に対するペナルティを指す．式 (4.23) は各環境のモデル $E_m$ の下で，エキスパート方策 $\pi^m_{\exp} \in \Pi_{\exp}$ が最適となる制約条件をすべて列挙したものであり，各制約条件式 (4.23) は，式 (4.16) と等しい．状態遷移行列 $\boldsymbol{T}_a$ は状態遷移確率 $T(s_m|a, s_i)$ を $(i, j)$ 成分とする $|\mathcal{S}| \times |\mathcal{S}|$ 行列，$a^m_1$ は環境 $E_m$ の各状態 $s$ におけるエキスパートの行動 $\pi^m_{\exp}(s)$ を示す．

提案法の定式化は，推定報酬 $\hat{R}$ がエキスパートの報酬と一致するための必要条件から導かれる．推定報酬 $\hat{R}$ がエキスパートの報酬と一致する必要条件とは，各環境のモデル $E_m$ におけるエキスパート方策 $\pi^m_{\exp} \in \Pi_{\exp}$ が，推定報酬 $\hat{R}$ に対して最適となることである．

提案法の制約条件と既存法の制約条件の比較を図 4.1 に示す．

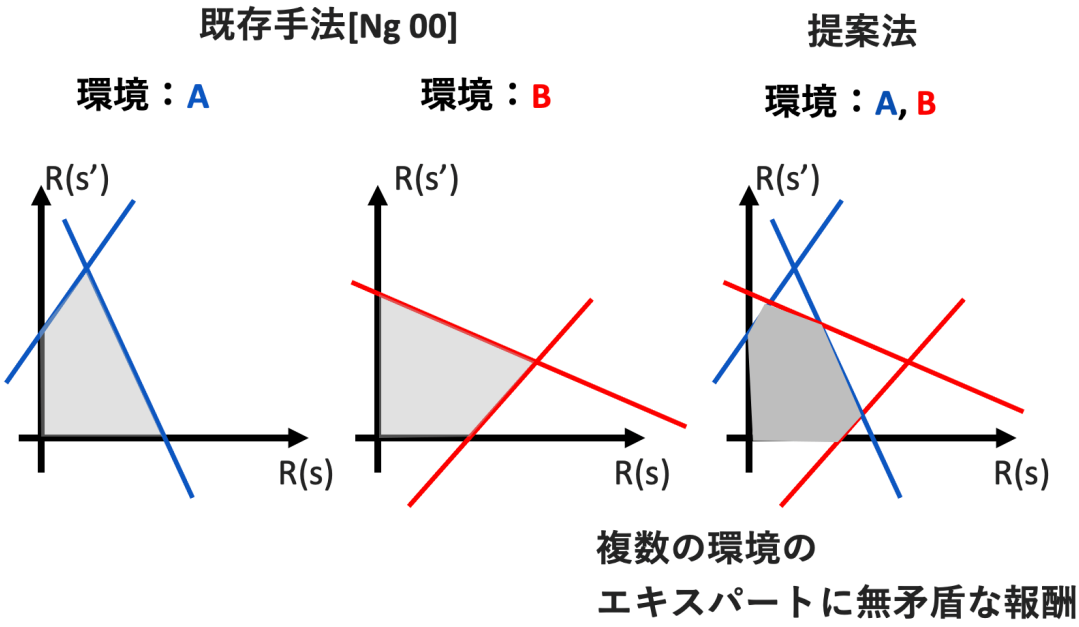



図 4.1: 既存研究 [Ng 00] と提案法の制約条件の比較

図 4.1 の各図は，状態数 $|S|$ が二つのマルコフ決定過程における報酬の空間を示している．各図の青と赤色の線は式 (4.20) や式 (4.24) の制約条件の式を表している．図 4.1 の左二つの図は，既存の線形計画逆強化学習の制約条件を満たす報酬の領域の模式図であり，一番右の図は提案法の制約条件式 (4.24) の模式図である．提案法の制約条件を示す一番右の図は，環境 A, B の両方の制約を満たす領域を示しており，単一の環境における実行可能領域よりも小さい領域となっている．エキスパートの報酬は，複数の環境における制約条件を満たす領域に存在することから，提案法を用いることによって，エキスパートに近い報酬を推定されることが期待できる．

提案法の推定報酬 $\hat{R}$ が，エキスパートの報酬と一致する必要条件を満たす理由を以下で述べる．ある環境 $E$ において，報酬 $R$ に対してエキスパート方策 $\pi_{\text{exp}}$ が最適となる条件は，式 (4.16) で表される．よって，各環境 $E_m$ におけるエキスパート方策 $\pi^m_{\text{exp}} \in \Pi_{\text{exp}}$ が最適となるすべての制約条件式 (4.23) を，報酬 $R$ が満たすとき，各環境のモデル $E_m$ における各エキスパート方策 $\pi^m_{\text{exp}} \in \Pi_{\text{exp}}$ は報酬 $R$ に対して最適方策である．よって，提案法の推定報酬は，エキスパートの報酬と一致する必要条件「各環境のモデル $E_m$ におけるエキスパート方策 $\pi^m_{\text{exp}} \in \Pi_{\text{exp}}$ が，推定報酬 $\hat{R}$ に対して最適となる」を満たす．

提案法の定式化の実行可能領域を，LPIRL の実行可能領域と比較し説明する．各環境 $E_m$ において，エキスパート方策 $\pi^m_{\text{exp}} \in \Pi_{\text{exp}}$ が最適となる報酬 $R$ の集合は，各環境の制約条件式 (4.16) を満たす実行可能領域の共通集合と等価である．提案法の定式化の制約条件式 (4.23) は，各環境の制約条件式 (4.16) から構成されるため，提案法の定式化の実行可能領域もまたこれと等しい．

提案法の実行可能領域は，各環境のモデル $E_m$ と，それに対応するエキスパート方策 $\pi_{\exp}^m \in \Pi_{\exp}$ から構成される LPIRL[Ng 00] の実行可能領域の共通集合と一致する．そのため，提案法の定式化の下で，報酬を推定できれば，推定報酬 $\hat{R}$ と各環境のモデル $E_m$ におけるエキスパート方策 $\pi_{\exp}^m \in \Pi_{\exp}$ との無矛盾性を保証できる．

本章で定式化した問題は，線形計画問題であるため，シンプレックス法などの線形計画問題を解く標準的なアルゴリズムを用いて解が求まる．ここでは，提案法の定式化におけるシンプレックス法の計算量を LPIRL の定式化と比較する．シンプレックス法の計算量は，実行可能基底解の数に依存する．線形計画問題の実行可能基底解の数の上界は．その標準形における変数の数 $j$，制約条件の数 $k$ を用いて，

$$j \left[ k\frac{\beta}{\alpha} \log k\frac{\beta}{\alpha} \right] \tag{4.25}$$

と表される [Kitahara 13]．ここで $\alpha$，$\beta$ はそれぞれ，全ての実行可能基底解の正の要素の最小値，最大値である．線形計画問題が非退化の時，実行可能基底解の数の上界はシンプレックス法の反復回数の上界となる．

提案法の定式化の標準形における変数の数 $j$ と，制約条件の数 $k$ はそれぞれ，

$$j = |\mathcal{S}|(M+2) \tag{4.26}$$

$$k = 2M|\mathcal{S}|(|\mathcal{A}|-1) + 4|\mathcal{S}| \tag{4.27}$$

である．式 (4.25), (4.26), (4.27) から，提案法の定式化における実行可能基底解の数の上界は，環境の数 $M$，状態数 $|\mathcal{S}|$，行動数 $|\mathcal{A}|$，に対して $\mathcal{O}\left(|\mathcal{A}|(|\mathcal{S}|M)^2 \log |\mathcal{A}||\mathcal{S}|M\right)$ のオーダーで増加することがわかる．

本章で提案したアプローチは，式 (4.20) を用いることによって，状態と行動に依存する報酬 $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ の推定にも適用できる．

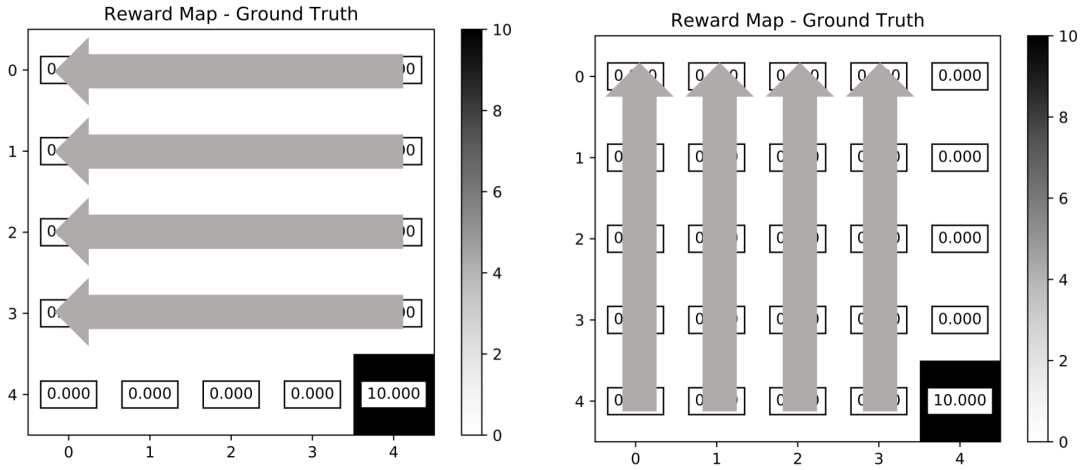
## 4.4 実験

提案法の推定報酬 $\hat{R}$ が，各環境のモデル $E_m$ の下で，エキスパート方策 $\pi_{\exp}^m \in \Pi_{\exp}$ に無矛盾であることを示す．実験には一定の確率で風が吹く方向に遷移する Windy grid world 環境を用いる．はじめに，実験設定を説明し，実験結果について述べる．

### 実験設定

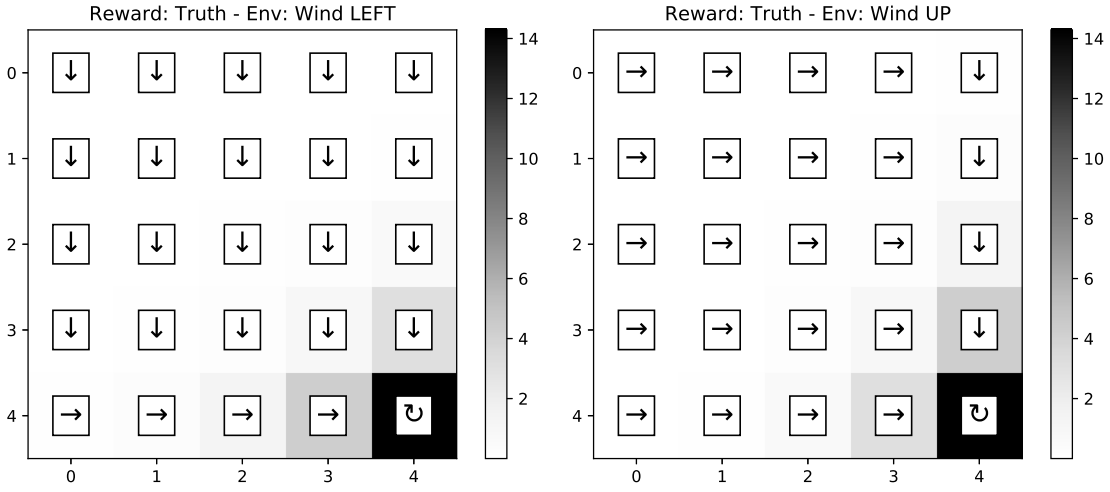
風向きが異なる二つの Windy grid world 環境を用いて，LPIRL [Ng 00] と提案法を比較する．Windy grid world 環境とは，風が吹く迷路問題で，風が吹く状態ではエージェントの行動に関わらず，ある一定の確率で風が吹く方向に遷移する．エージェントの行動は五つで，上下左右へ移動する四つの行動とその場に留まる行動からなる．

図 4.2 に風向きが異なる二つの実験環境を示す．




(a) Wind left　　(b) Wind up

図 4.2: 風向きが異なる Windy grid world 環境




(a) Wind left　　(b) Wind up

図 4.3: 各 Windy grid world 環境におけるエキスパート方策

図 4.2 の灰色の矢印は，矢印のある状態で吹く風の方向を示す．本実験では風が吹く方向に遷移する確率を 30%とした．各マス目の色の濃淡は報酬の大きさに比例し，右側のバーは色の濃淡に対応する報酬の値を示す．各マス目の数値は各状態 $s \in \mathcal{S}$ の報酬 $R(s)$ であり，バーが示す報酬の値に対応する．最適方策 $\pi^*$ は報酬の期待値を最大化する方策であり，各環境 $E_m$ における最適方策 $\pi^*_m$ はその風向きによって異なる．

図 4.3 に，図 4.2 の各環境における最適方策を示す．図 4.3 の各状態の矢印は，各状態に

おける最適行動を指す．最適行動の計算には割引率 $\gamma = 0.3$ の価値反復 [Sutton 18] を用いた．これら二つの環境における，風が吹く状態の最適行動は，風が吹いていない状態へと最短経路で移動する行動である．具体的には，風が吹く状態の最適行動は，風向きが左の状態で下，風向きが上の状態で右である．

本章では，これら二つの環境それぞれの最適方策を，各環境におけるエキスパート方策とし，エキスパートの報酬を推定する．報酬を推定する際，既存手法，提案法の報酬の最大値 $R_{\max}$ を定める必要がある．ここでは，制約条件，式 (4.17), (4.24) に示す報酬の最大値 $R_{\max}$ をエキスパートの報酬の最大値と同じ値 10 として実験を行った．

報酬のスパース性を調整するパラメータ $\lambda$ は，最もスパースな解（全状態の報酬が 0 の解を除く）が得られるパラメータ $\lambda$ の最小値を，0.1 刻みのグリッドサーチで探索し選択した．ここでは，既存手法の $\lambda$ を 1.1，提案法の $\lambda$ を 0.1 とした．

本実験で用いる評価指標について述べる．本実験では，提案法の推定報酬と既存法の推定報酬を，エキスパートの報酬との近さの指標である Expected Value Difference（EVD）[Levine 11] を用いて評価する．EVD の定義式を (4.28) に示す．

$$\mathrm{EVD} = E_{\pi_{\exp}}\left[\sum_{t=0}^{\infty}\gamma^t R_{\exp}(s_t)\right] - E_{\hat{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t R_{\exp}(s_t)\right] \tag{4.28}$$

ここで，$R_{\exp}$ はエキスパートの報酬，$\pi_{\exp}$ はエキスパートの方策，$\hat{\pi}$ は推定報酬 $\hat{R}$ に対する最適方策を指す．EVD の値が小さいほど推定報酬がエキスパートの報酬と類似している．エキスパートの方策 $\pi_{\exp}$ はエキスパートの報酬 $R_{\exp}$ に対する最適方策であるため EVD は非負で，その最小値は 0 である．

### 実験結果

既存手法と提案法で報酬を推定し，各環境におけるエキスパート方策と，推定報酬に対する最適方策を比較する．

表 4.1 に各手法による推定報酬と，推定報酬に対する図 4.2 の各環境の最適方策を示す．

表 4.1: 既存手法と提案法の比較．1 列目の各図は報酬，2 列目の各図は風が左へ吹く環境における最適方策，3 列目の各図は風が上へ吹く環境における最適方策である．

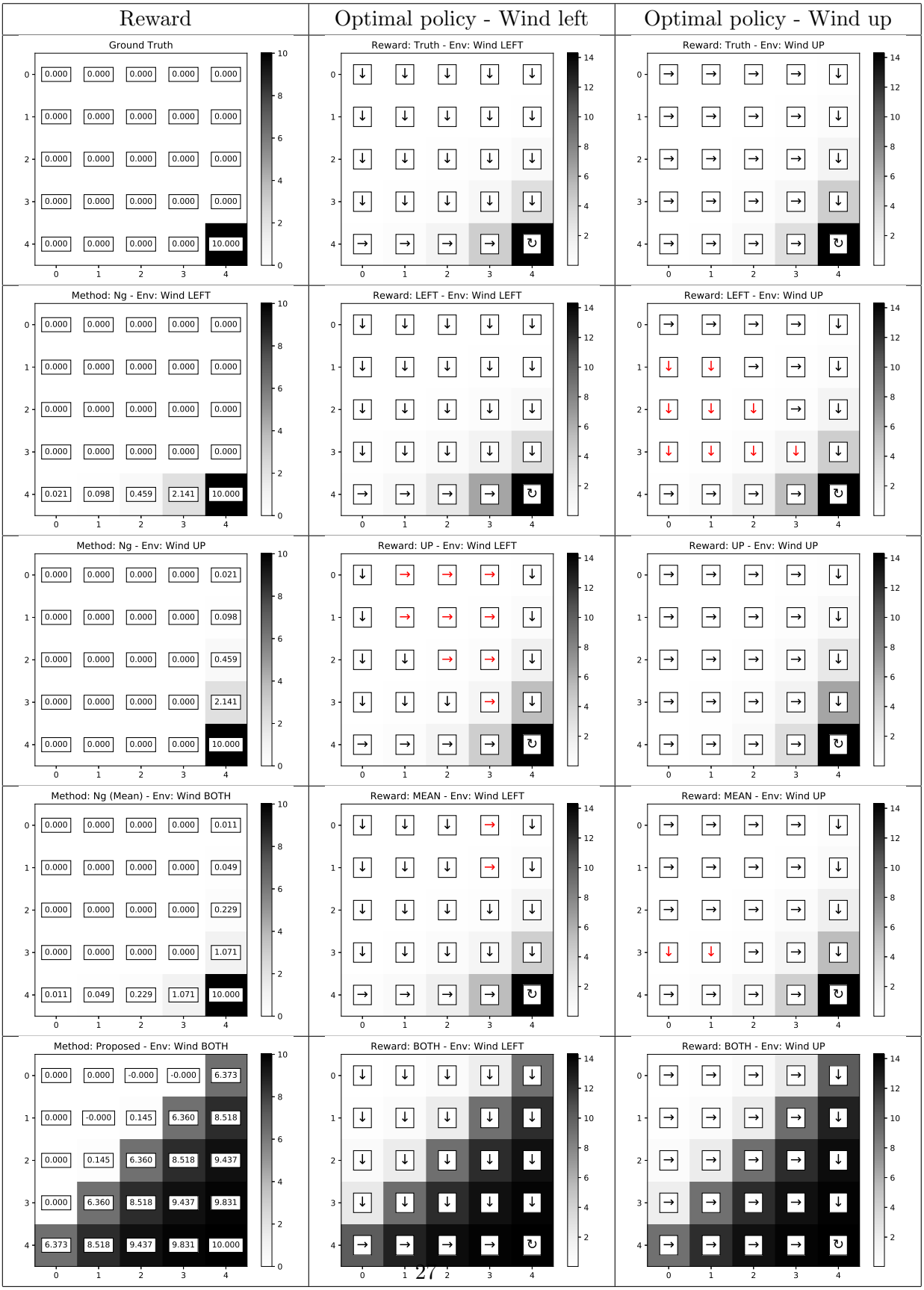


表 4.1 の 1 行目は，各列の図の内容を示す．具体的には，1 列目の各図は報酬，2 列目の各図は風が左へ吹く環境における最適方策，3 列目の各図は風が上へ吹く環境における最適方策である．

1 列目の各図は上から，エキスパートの報酬，風が左へ吹くの環境のモデルとエキスパート方策に基づく推定報酬（既存手法），風が上へ吹くの環境のモデルとエキスパート方策に基づく推定報酬（既存手法），二つの環境のモデルと各環境におけるエキスパート方策に基づく推定報酬（提案法）である．各状態の値は，その状態 $s$ の報酬 $R(s)$ を表す．2 列目の各図は，風が左に吹く環境における，各行の 1 列目の報酬に対する最適方策で，3 列目の各図は，風が上に吹く環境における各行の 1 列目の報酬に対する最適方策である．

ここで，右の二列の各図中の各状態における矢印は，その状態における最適行動を指す．各状態の最適行動を表す矢印の色は，エキスパート方策との一致，不一致を示す．黒色の矢印は，最適行動とエキスパートの行動との一致を示し，赤色の矢印は，最適行動とエキスパートの行動との不一致を示す．

右の二列の各図中の各状態の色の濃淡は状態価値を表し，色の濃淡は状態価値に比例する．各図の右側のバーは色の濃淡に対応する状態価値の値を示す．

同じ報酬に対する最適方策から報酬を推定しても，その最適方策と，状態遷移確率が異なれば，異なる報酬が推定される．例えば，表 4.1 の 3 行 1 列目の図は，風向きが左の環境のモデルと最適方策から推定した報酬，4 行 1 列目の図は，風向きが左の環境のモデルと最適方策から推定した報酬を示す．これら二つの推定報酬の値が異なることから，同じ報酬に対する最適方策から報酬を推定しても，その最適方策と，状態遷移確率が異なれば，推定報酬も異なる場合があることが確認できる．

既存手法では，風向きが異なる環境において，推定報酬に対する最適方策はエキスパート方策と一致しない．例えば，表 4.1 の 3 行 3 列目の図は，風向きが左の環境で推定した報酬に対する，風向きが上の環境における最適方策であり，図の座標 (2,2) における最適行動は「下」である．しかし，表 4.1 の 2 行 3 列目の図から分かるように，エキスパートの報酬に対する最適行動は「右」である．このように，同じ報酬に対するエキスパート方策を用いて報酬を推定しても，エキスパート方策とは異なる方策が最適方策となる．

5 行目は，二つの環境における既存手法の推定報酬を平均した結果である．複数の環境における推定報酬の平均値に対する最適方策はエキスパート方策と一致しないことが確認できる．

提案法は，既存手法と異なり，風向きが異なる各環境におけるエキスパート方策が最適となる報酬を推定した．これは，環境の風向きに関わらず，各環境におけるエキスパート方策（表 4.1，2 行目）と推定報酬に対する最適方策（表 4.1，最下行）が一致したことから確認できる．

次に，推定報酬とエキスパートの報酬の近さを Expected Value Difference（EVD）[Levine 11] を用いて評価する．EVD は，推定報酬に対する最適方策とエキスパートの方策の近さを表す

指標で，EVD の値が小さいほど推定報酬もエキスパートの報酬と類似している．推定報酬の EVD は，エキスパート方策で得られるエキスパートの報酬の期待値から，推定報酬に対する最適方策で得られるエキスパートの報酬の期待値を引くことで求まる．ここでは，EVD の評価には各状態の風向きをランダムに定めた 500 個の環境を用いる．

表 4.2 に，各手法による推定報酬の EVD を示す．

表 4.2: Expected Value Difference の比較

| Method | Environment | EVD (Mean ± Std) |
|---|---|---|
| Ng | Left | $6.97 \times 10^{-3} \pm 8.55 \times 10^{-3}$ |
| Ng | Up | $6.51 \times 10^{-3} \pm 8.23 \times 10^{-3}$ |
| Ng (Mean) | Both | $5.89 \times 10^{-4} \pm 8.09 \times 10^{-4}$ |
| Proposed | Both | $\mathbf{3.91 \times 10^{-4} \pm 7.97 \times 10^{-4}}$ |

提案法の EVD が最も小さいことから，提案法によって，よりエキスパートに近い報酬が推定されたことが確認できる．

## 4.5 考察

本章では，4.4 の実験結果に基づき以下の四点を考察する．

### 状態遷移確率が異なれば，同じ報酬のエキスパート方策から報酬を推定しても，異なる報酬が推定される要因

表 4.1 の 3 行 1 列目，4 行 1 列目は，異なる状態遷移確率の環境の下での，既存手法の推定報酬である．各図を比較すると，各環境間で，推定報酬が異なることが確認できる．ここでは，同じ報酬に対するエキスパート方策を用いても，状態遷移確率の異なる環境間で，推定報酬が異なる原因を考察する．

状態遷移確率の異なる環境間で，推定報酬が異なる原因は二つある．一つ目は，状態遷移確率が異なる環境間での，線形計画問題の実行可能領域の違いである．線形計画問題の制約条件の式は，状態遷移確率を含むため，状態遷移確率が異なれば，線形計画問題は異なる制約条件を持つ．そのため，状態遷移確率が異なる環境間では，実行可能領域が異なり，異なる報酬が推定される場合がある．

二つ目の原因は，状態遷移確率が異なる環境間の線形計画問題の目的関数の違いである．一つ目の要因である，実行可能領域の違いに関する議論が，目的関数に対しても同様に成り立つ．異なる目的関数を持つ問題の解は異なる場合があるため，状態遷移確率が異なる環境間で，異なる報酬が推定される場合がある．これら二つが，同じ報酬に対するエキスパート方策を用いても，状態遷移確率の異なる環境間で推定報酬が異なる原因である．

### 報酬を推定した環境と異なる状態遷移確率の環境において，既存手法の推定報酬に対する最適方策が，エキスパート方策と一致しない要因

表 4.1 の 3 行目と，4 行目の図中に赤い矢印がある．赤い矢印の状態 $s$ は，推定報酬に対する最適方策 $\pi^*(s)$ とエキスパート方策 $\pi_{\exp}(s)$ が異なる．ここでは，状態遷移確率が異なる環境において，既存手法の推定報酬に対する最適方策と，エキスパート方策が異なる要因を考察する．

状態遷移確率が異なる環境において，既存手法の推定報酬に対する最適方策が，エキスパート方策と一致しない要因は，推定報酬が複数の環境の線形計画問題の制約条件を，同時に満たしていないことにある．各環境の状態遷移確率，エキスパート方策から構成される線形計画問題の制約条件は,「各環境においてエキスパート方策が最適となる」という制約である．表 4.1 の 3 行目と，4 行目の図に赤い矢印が存在することから，既存手法による推定報酬は，所与の環境の制約を満たすが，他の環境の制約条件を満たさない．つまり，既存手法による推定報酬は，二つの環境の線形計画問題の実行可能領域の差集合の要素であると言える．

### 複数の環境における推定報酬の平均値に対する最適方策が，エキスパートの方策と一致しない要因

表 4.1 の 5 行目は，複数の環境における推定報酬の平均値を用いた結果である．図中に赤い矢印があることから，推定報酬の平均値に対する最適方策 $\pi^*(s)$ とエキスパート方策 $\pi_{\exp}(s)$ が異なることが分かる．これは，既存手法による推定報酬の平均値が，報酬の推定に用いた環境の下での制約条件を満たしていないことを示す．一方，提案法は複数の環境の下での制約条件を満たす報酬を推定している．

### 提案法の推定報酬がエキスパートの報酬と一致するための必要条件を満たす要因

表 4.1 の 6 行目の図が示すように，提案法は，複数の環境のエキスパート方策が最適となる報酬を推定している．ここでは，提案法の推定報酬がエキスパートの報酬と一致するための必要条件を満たす理由を述べる．

6.2 で述べたように，任意の状態遷移確率の環境において，エキスパートの報酬に対する最適方策は，エキスパート方策と一致する．そのため，複数の環境のエキスパート方策が推定報酬に対して最適となることは，推定報酬がエキスパートの報酬と一致する必要条件である．提案法は，各環境の線形計画問題の制約条件式 (4.16) を満たす実行可能領域の共通集合から報酬を推定する．したがって，提案法の線形計画問題に実行可能領域が存在するとき，推定報酬はエキスパートの報酬と一致する必要条件を満たす．

## 4.6 まとめ

本章では,「状態遷移確率の異なる複数の環境下で得られるエキスパート方策」と，推定報酬との無矛盾性を保証する逆強化学習手法を提案した．具体的には，状態遷移確率の異なる複数の環境下のエキスパートが持つ報酬を推定する問題を線形計画問題として定式化し，「状態遷移確率のみ異なる複数の環境のモデル」と「各環境で同じ報酬に従うエキスパート方策」から，各環境におけるエキスパート方策が最適となる報酬を推定する方法を示した．

提案法の貢献として,「状態遷移確率の異なる複数の環境下で得られるエキスパート方策」と，推定報酬との無矛盾性を保証したことを挙げる．風向き（状態遷移確率）が異なる複数の Windy grid world 環境を用いた実験を通じて，既存手法の推定報酬は，状態遷移確率が異なる環境におけるエキスパート方策と矛盾するが，提案法の推定報酬は，各環境におけるエキスパート方策に無矛盾であることを確認した．

# 第5章　複数のマルコフ決定過程におけるベイジアン逆強化学習

本章では，複数の環境のエキスパートの軌跡から報酬の事後分布を推定するベイジアン逆強化学習問題を定式化し，報酬の事後分布からのサンプリングを実現するマルコフ連鎖モンテカルロアルゴリズムを提案する．そして，提案したアルゴリズムが報酬の事後分布の実現に要するステップ数が，環境の数に依存しないことを示す．

実験では，二つの実験設定で既存のベイジアン逆強化学習法と提案法を比較する．逆強化学習手法の比較に一般的に用いられる指標である Expected Value Difference(EVD)[Levine 11] を用いて比較した結果，どちらの実験設定においても，提案法の推定報酬が既存手法の推定報酬と比較してエキスパートの報酬に近いことが確認された．また，ハイパーパラメータの変化に対しても頑健であることを実験を通して確認する．

## 5.1　準備

### 5.1.1　ベイジアン逆強化学習

ベイジアン逆強化学習（Bayesian Inverse Reinforcement Learning; BIRL）は，ある報酬に従うエキスパートが生成した状態行動対 $(s,a)$ のデータセット $D=\{(s_i,a_i)\}_{i=1}^{N}$ と，報酬に対する事前知識を反映した報酬の事前分布 $P(R)$ を所与として，報酬の事後分布 $P(R|D)$ を推定する枠組みである [Ramachandran 07]．報酬の事後分布は，ベイズの定理を用いて，

$$P(R|D)=\frac{P(D|R)}{P(D)}P(R), \tag{5.1}$$

と表される．尤度 $P(D|R)$ は，報酬 $R$ に対する最適方策の行動価値である $Q^*(s,a,R)$ 値を用いてモデル化される．尤度 $P(D|R)$ の定式化を式 (5.2) に示す．

$$P(D|R)=\frac{1}{Z}\exp\left(\frac{1}{\kappa}\sum_{(s,a)\in D}Q^*(s,a,R)\right) \tag{5.2}$$

ここで，$\kappa>0$ はエキスパートが最適行動を取る確率を調節するパラメータで，ボルツマン方策 [Sutton 18] における温度パラメータに相当する．報酬の事後分布 $P(R|D)$ は，式 (5.1) と式 (5.2) を用いて式 (5.3) と表すことができる．

$$P(R|D)=\frac{1}{Z'}\exp\left(\frac{1}{\kappa}\sum_{(s,a)\in D}Q^*(s,a,R)\right)P(R) \tag{5.3}$$

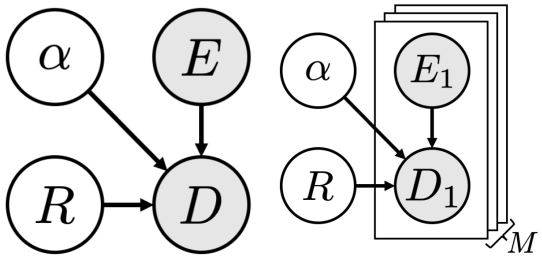
式 (5.3) の分配関数 $Z'$ は計算困難だが，指数は報酬 $R$ に対する最適方策の行動価値 $Q^*$ を求めることによって計算できる．また，

$$P(R|D) \propto \exp\left(\frac{1}{\kappa}\sum_{(s,a)\in D} Q^*(s,a,R)\right) P(R), \tag{5.4}$$

が成り立つため，マルコフ連鎖モンテカルロ法（MCMC）を用いて事後分布 $P(R|D)$ から報酬をサンプルできる．BIRL では PolicyWalk と呼ばれる MCMC アルゴリズムを用いて事後分布 $P(R|D)$ から報酬 $R$ をサンプルする．PolicyWalk とは，代表的な MCMC アルゴリズムである GridWalk[Vempala 05] を，不要な強化学習を省くことによって効率化した MCMC アルゴリズムである．

## 5.2 対象問題

本章では，ある報酬に従うエキスパートが，複数の環境で生成した軌跡からエキスパートの報酬の分布を推定する問題を定式化する．BIRL と対象問題のグラフィカルモデルの比較を図 5.1 に示す．




(a) BIRL　　(b) BIRL-MD

図 5.1: ベイジアン逆強化学習と提案法のグラフィカルモデルの比較

図 5.1 の影がついたノードは観測可能なデータを，矢印はデータ間の依存関係を指す．BIRL は，エキスパートがある環境 $E$ において生成したデータセット $D$ を所与として，報酬の分布 $P(R|D) = P(R|D,E)$ を推定する問題を定式化した．一方，本論文が定式化する複数環境におけるベイジアン逆強化学習（BIRL for Multiple Dynamics; BIRL-MD）は，状態遷移確率が異なる環境 $E_m = \langle \mathcal{S}, \mathcal{A}, T_m, \gamma \rangle$ と，エキスパートが各環境 $E_m$ で生成したデータセット $D_m$ の組からなる集合 $\{(E_m, D_m)\}_{m=1}^M$ を所与として，報酬の事後分布 $P(R|\{(D_m, E_m)\}_{m=1}^M)$ を推定する定式化である．また，図 5.1 の矢印から分かるように，報酬 $R$ は環境 $E$ と独立，つまりエキスパートの報酬は環境の状態遷移確率に依存しないものとする．

複数の環境におけるエキスパートのデータセット $\{(E_m, D_m)\}_{m=1}^M$ の下での報酬の事後分布を式 (6.1) に示す．

$$P(R|\{(D_m, E_m)\}_{m=1}^M) = \frac{\Pi_{m=1}^M P(D_m|R, E_m)}{\Pi_{m=1}^M P(D_m|E_m)} P(R), \tag{5.5}$$

式 (5.2) に示した BIRL のエキスパートのモデルを，環境 $E$，報酬 $R$ の下での最適方策の行動価値 $Q^*(s,a,R,E)$ へと拡張すれば，報酬の事後分布は

$$P(R|\{(D_m,E_m)\}_{m=1}^M) = \frac{1}{Z}\exp\left(\frac{1}{\kappa}\sum_{m=1}^{M}\sum_{(s,a)\in D_m} Q^*(s,a,R,E_m)\right)P(R), \tag{5.6}$$

と表される．

## 5.3 アプローチ

提案法は，環境 $E_m = \langle \mathcal{S}, \mathcal{A}, T_m, \gamma \rangle$ と，エキスパートが各環境において生成したデータセット $D_m$ の組からなる集合 $\{(E_m, D_m)\}_{m=1}^M$ を所与として，報酬の事後分布 $P(R|\{(D_m,E_m)\}_{m=1}^M)$ からの報酬 $R$ のサンプリングを実現する．式 (6.2) の指数は，各環境 $E$ における報酬 $R$ に対する最適方策の行動価値 $Q^*(s,a,R,E)$ を計算することで求まる．そのため，BIRL-MD においても BIRL と同様に，MCMC を用いて事後分布 $P(R|\{(D_m,E_m)\}_{m=1}^M)$ から報酬をサンプルできる．提案法の MCMC アルゴリズム，PolicyWalk for Multiple Dynamics (PolicyWalk-MD) を Algorithm 4 に示す．ここで，Algorithm 4 の 7 行目の PolicyIteration は方策反復と呼ばれる動的計画法の一種である [Sutton 18]．

**Algorithm 4** PolicyWalk for Multiple Dynamics

**INPUT:** Environments $\{E_m\}_{m=1}^M$, Demonstrations $\{D_m\}_{m=1}^M$, Prior $P(R)$, Step Size $\delta$
**OUTPUT:** Sampled Rewards $\{R_i\}_{i=1}^N$

1: Pick a random vector $R_0 \in \mathbb{R}^{|S|}/\delta$
2: $\{\pi_m\}_{m=1}^M \leftarrow \{\text{PolicyIteration}(E_m, R_0)\}_{m=1}^M$
3: **for** $i = 1$ **do** N
4: Pick a reward vector $\tilde{R}$ uniformly at random from the neighbours of $R_{i-1} \in \mathbb{R}^{|S|}/\delta$
5: Compute $Q^\pi(s,a,\tilde{R},E) \quad \forall\{s,a,(E_m,\pi_m)\} \in \mathcal{S}\times\mathcal{A}\times\{(E_m,\pi_m)\}_{m=1}^M$
6: **if** $\exists\{s,a,(E,\pi)\} \in \mathcal{S}\times\mathcal{A}\times\{(E_m,\pi_m)\}_{m=1}^M, Q^\pi(s,\pi(s),\tilde{R},E) < Q^\pi(s,a,\tilde{R},E)$ **then** ▷ If any policy is not optimal
7: $\{\tilde{\pi}_m\}_{m=1}^M \leftarrow \left\{\text{PolicyIteration}(E_m,\tilde{R})\right\}_{m=1}^M$
8: $R_i \leftarrow \tilde{R}$ and $\{\pi_m\}_{m=1}^M \leftarrow \{\tilde{\pi}_m\}_{m=1}^M$ with probability $\min\left\{1, \frac{P(\tilde{R},\{(D_m,E_m)\}_{m=1}^M)}{P(R_{i-1},\{(D_m,E_m)\}_{m=1}^M)}\right\}$
9: **else**
10: $R_i \leftarrow \tilde{R}$ with probability $\min\left\{1, \frac{P(\tilde{R},\{(D_m,E_m)\}_{m=1}^M)}{P(R_{i-1},\{(D_m,E_m)\}_{m=1}^M)}\right\}$
11: **end if**
12: **end for**

BIRL の MCMC アルゴリズムである PolicyWalk は，$P(R|D,E)$ から誤差 $\epsilon$ 以下のサンプリングを $O(|\mathcal{S}|^2 \log\frac{1}{\epsilon})$ のステップ数で実現する [Ramachandran 07]．PolicyWalk-MD の

サンプリングの計算量に関して次の定理が成り立ち，提案法は PolicyWalk と同じオーダーで事後分布からのサンプリングを実現する．

**補題 1.** [Ramachandran 07]

$F$ を，定義域が $\{x \in \mathbb{R}^n | -d \leq x_i \leq d\}$ の正の実数関数とする．ここで，$d$ は任意の正の実数である．また，任意の $\lambda \in [0,1]$ の下で，ある $\alpha, \beta$ が存在し，以下の二つの式

$$|f(x) - f(y)| \leq \alpha \|x - y\|_\infty, \tag{5.7}$$

$$f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda) f(y) - \beta, \tag{5.8}$$

が成り立つと仮定する．ここで，$f(\cdot) = \log F(\cdot)$ である．この時，関数 $F$ に対する GridWalk（PolicyWalk，PolicyWalk-MD も同様に）は誤差 $\epsilon$ 以下のサンプリングを，$O(n^2 d^2 \alpha^2 e^{2\beta} \log \frac{1}{\epsilon})$ ステップで実現する [Applegate 91]．

**定理 1.** 環境 $E = \langle \mathcal{S}, \mathcal{A}, T, \gamma \rangle$，事前分布 $P(R)$ が一様分布 $\mathcal{U}(-R_{\max}, R_{\max})$ で，事後分布 $P(R|\{(D_m, E_m)\}_{m=1}^M)$ が式 (6.2) に示す式で表されるとする．$R_{\max} = O(\frac{1}{\sum_{m=1}^M |D_m|})$ と仮定すると，PolicyWalk-MD は $P(R|\{(D_m, E_m)\}_{m=1}^M)$ からの誤差 $\epsilon$ 以下のサンプリングを，$O(|\mathcal{S}|^2 \log \frac{1}{\epsilon})$ のステップ数で実現する．

*Proof.*

$$f(R) = \frac{1}{\kappa} \sum_{m=1}^{M} \sum_{(s,a) \in D_m} Q^*(s, a, R, E_m) \tag{5.9}$$

$$f_\pi(R) = \frac{1}{\kappa} \sum_{m=1}^{M} \sum_{(s,a) \in D_m} Q^\pi(s, a, R, E_m) \tag{5.10}$$

ここで，$f_\pi$ はベクトル化した $R$ について線形で，$f(R) \geq f_\pi(R)$ が成り立つ．また，$Q$ 値について，

$$\max_{s,a} Q^*(s,a) \leq \sum_{t=0}^{\infty} \gamma^t R_{\max} = \frac{R_{\max}}{1-\gamma} \tag{5.11}$$

$$\min_{s,a} Q^*(s,a) \geq -\frac{R_{\max}}{1-\gamma} \tag{5.12}$$

が成り立つ．Q 値の下界を与える不等式 (5.12) から，$f_\pi(R)$ の下界

$$f_\pi(R) \geq -\sum_{m=1}^{M} |D_m| \frac{R_{\max}}{\kappa(1-\gamma)}, \tag{5.13}$$

が得られる．また，式 (5.11) から，

$$\frac{\alpha \sum_{m=1}^M |D_m| R_{\max}}{1-\gamma} \geq f(R), \tag{5.14}$$

$$0 \geq f(R) - \sum_{m=1}^{M} |D_m| \frac{R_{\max}}{\kappa(1-\gamma)}, \tag{5.15}$$

と言える．式 (5.13) と (5.15) を足すことによって $f(R)$ と $f_\pi(R)$ の関係式，

$$f_\pi(R) \geq f(R) - 2\sum_{m=1}^{M} |D_m| \frac{R_{\max}}{\kappa(1-\gamma)}. \tag{5.16}$$

が得られる．式 (5.16) を用いれば

$$\begin{aligned} f\left(\lambda R_1 + (1-\lambda)R_1\right) \geq \quad & f_\pi\left(\lambda R_1 + (1-\lambda)R_2\right) && (5.17)\\ = \quad & \lambda f_\pi(R_1) + (1-\lambda) f_\pi(R_2) && (5.18)\\ \geq \quad & \lambda f(R_1) + (1-\lambda) f(R_2) \\ & - \frac{2\sum_{m=1}^{M} |D_m| R_{\max}}{\kappa(1-\gamma)} && (5.19) \end{aligned}$$

となる．ここで，$f_\pi$ はベクトル化した $R$ について線形であることを用いた．補題の変数 $\alpha, \beta$ はそれぞれ，

$$\begin{aligned} \alpha &= \frac{|f(R_1) - f(R_2)|}{\|R_1 - R_2\|_\infty} \\ &\leq \frac{2\sum_{m=1}^{M} |D_m| R_{\max}}{\kappa(1-\gamma) O\left(\frac{1}{\sum_{m=1}^{M} |D_m|}\right)} = O(\sum_{m=1}^{M} |D_m|) \end{aligned} \tag{5.20}$$

$$\begin{aligned} \beta &= \frac{2\frac{1}{\kappa}\sum_{m=1}^{M} |D_m| R_{\max}}{1-\gamma} \\ &= 2\sum_{m=1}^{M} |D_m| \frac{O\left(\frac{1}{\sum_{m=1}^{M} |D_m|}\right)}{\kappa(1-\gamma)} = O(1) \end{aligned} \tag{5.21}$$

である．よって PolicyWalk-MD は $P\left(R|\{(D_m, E_m)\}_{m=1}^{M}\right)$ からの誤差 $\epsilon$ 以下のサンプリングを，

$$O\left(|\mathcal{S}|^2 \frac{1}{\left(\sum_{m=1}^{M} |D_m|\right)^2} \left(\sum_{m=1}^{M} |D_m|\right)^2 \log\frac{1}{\epsilon}\right) = O\left(|\mathcal{S}|^2 \log\frac{1}{\epsilon}\right) \tag{5.22}$$

のステップ数で実現する． □

報酬のスケールに関する仮定 $R_{\max} = O(\frac{1}{\sum_{m=1}^{M} |D_m|})$ を設けることによって推定報酬の有用性が低下することはない．なぜなら，推定報酬を定数倍しても，その最適方策は変わらないからである．PolicyWalk-MD が要するステップ数のオーダーは既存手法と等しいが，各ステップで要する方策反復の数は環境の数に依存する．提案法の計算量は環境の数に対してたかだか線形である．

ここでは報酬を状態に依存する関数 $R : \mathcal{S} \to \mathbb{R}$ として定義したが，提案法は状態と行動に依存する報酬関数 $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ や，報酬関数が特徴量 $\phi : \mathcal{S} \to \mathbb{R}^n$ の線形和 $R(s) = w \cdot \phi(s)$ で表される場合にも適用できる．状態と行動に依存する報酬関数 $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ を用いた場合，PolicyWalk-MD は $P(R|\{(D_m, E_m)\}_{m=1}^{M})$ からの誤差 $\epsilon$ 以下のサンプリングを，

$O\left((|\mathcal{S}||\mathcal{A}|)^2 \log \frac{1}{\epsilon}\right)$ のステップ数で実現する．また，報酬関数が特徴量 $\phi : \mathcal{S} \to \mathbb{R}^n$ の線形和 $R(s) = w \cdot \phi(s)$ で表される場合，PolicyWalk-MD は $P(R|\{(D_m, E_m)\}_{m=1}^M)$ からの誤差 $\epsilon$ 以下のサンプリングを，$O\left(n^2 \log \frac{1}{\epsilon}\right)$ のステップ数で実現する．転移可能な報酬を推定するためには，報酬を状態に依存する関数 $R : \mathcal{S} \to \mathbb{R}$ として定義すると良いと知られている [Fu 18]．そのため，本研究では状態に依存する報酬関数 $R(s)$ を用いる．

## 5.4 実験

実験では二つの環境で実験を行う．一つ目は状態数と行動数が少なく，エキスパートが最適となる報酬が複数存在する環境で，複数の環境の軌跡を用いることの有用性を確認する．もう一つは，逆強化学習のベンチマークとして用いられる Grid world 環境で，一つ目の環境と比較して状態数，行動数が多い環境における提案法の有用性を確認する．

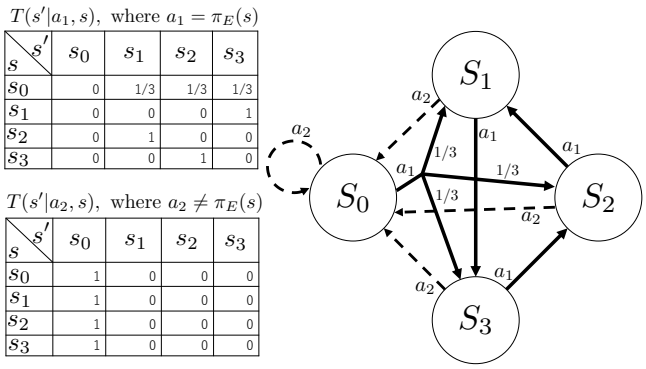
本実験では，提案法の推定報酬，既存法（BIRL）の推定報酬と，エキスパートの報酬の近さを Expected Value Difference（EVD）[Levine 11] を用いて評価する．EVD の定義式を (5.23) に示す．

$$\mathrm{EVD} = E_{\pi_{\mathrm{exp}}}\left[\sum_{t=0}^{\infty} \gamma^t R_{\mathrm{exp}}(s_t)\right] - E_{\hat{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t R_{\mathrm{exp}}(s_t)\right] \tag{5.23}$$

ここで，$R_{\mathrm{exp}}$ はエキスパートの報酬，$\pi_{\mathrm{exp}}$ はエキスパートの方策，$\hat{\pi}$ は推定報酬 $\hat{R}$ に対する最適方策を指す．EVD の値が小さいほど推定報酬がエキスパートの報酬と類似している．エキスパートの方策 $\pi_{\mathrm{exp}}$ はエキスパートの報酬 $R_{\mathrm{exp}}$ に対する最適方策であるため EVD は非負で，その最小値は 0 である．以下で，各実験の設定と，その結果について述べる．

### Small Environments with Uncertainty in Reward Estimation

ここでは，エキスパートが最適となる報酬が複数存在することが明らかな環境を用いて，提案法と既存手法を比較する．図 5.2 に，報酬の推定に用いる二つの環境を示す．



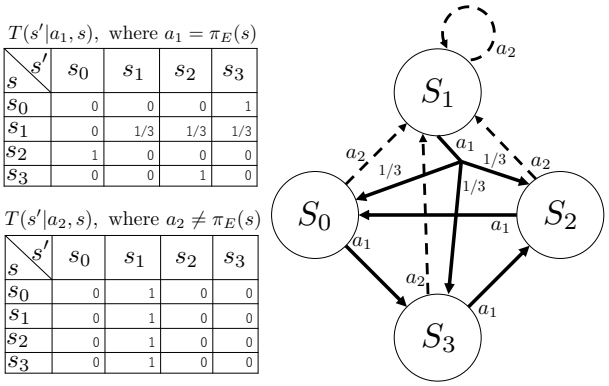
$T(s'|a_1, s)$, where $a_1 = \pi_E(s)$

| $s$ \ $s'$ | $s_0$ | $s_1$ | $s_2$ | $s_3$ |
|---|---|---|---|---|
| $s_0$ | 0 | 1/3 | 1/3 | 1/3 |
| $s_1$ | 0 | 0 | 0 | 1 |
| $s_2$ | 0 | 1 | 0 | 0 |
| $s_3$ | 0 | 0 | 1 | 0 |

$T(s'|a_2, s)$, where $a_2 \neq \pi_E(s)$

| $s$ \ $s'$ | $s_0$ | $s_1$ | $s_2$ | $s_3$ |
|---|---|---|---|---|
| $s_0$ | 1 | 0 | 0 | 0 |
| $s_1$ | 1 | 0 | 0 | 0 |
| $s_2$ | 1 | 0 | 0 | 0 |
| $s_3$ | 1 | 0 | 0 | 0 |

(a) Environment A



$T(s'|a_1, s)$, where $a_1 = \pi_E(s)$

| $s$ \ $s'$ | $s_0$ | $s_1$ | $s_2$ | $s_3$ |
|---|---|---|---|---|
| $s_0$ | 0 | 0 | 0 | 1 |
| $s_1$ | 0 | 1/3 | 1/3 | 1/3 |
| $s_2$ | 1 | 0 | 0 | 0 |
| $s_3$ | 0 | 0 | 1 | 0 |

$T(s'|a_2, s)$, where $a_2 \neq \pi_E(s)$

| $s$ \ $s'$ | $s_0$ | $s_1$ | $s_2$ | $s_3$ |
|---|---|---|---|---|
| $s_0$ | 0 | 1 | 0 | 0 |
| $s_1$ | 0 | 1 | 0 | 0 |
| $s_2$ | 0 | 1 | 0 | 0 |
| $s_3$ | 0 | 1 | 0 | 0 |

(b) Environment B

図 5.2: エキスパートが最適となる報酬が複数存在する複数環境の例．$a_1$ はエキスパートの行動を, $a_2$ はその他の行動を示す．

ここで，エキスパートの報酬を $R(s_0) = R(s_1) = 0$，$R(s_2) = R(s_3) = 0.7$ とした．各環境におけるエキスパートの方策（最適方策）は，全状態のうち三つの状態を巡回する方策であり，エキスパートによって巡回されるいずれか，または複数の状態の報酬が大きい場合にエキスパートが最適となる．したがって，これらの環境にはエキスパートが最適となる報酬が複数存在する．例えば，環境 A では，$s_1, s_3, s_2$ という順に状態を巡回する方策がエキスパートの方策で，$s_1, s_3, s_2$ のいずれか，または複数の状態の報酬が大きい場合に，エキスパートの方策が最適となる．一方，環境 B では，$s_0, s_3, s_2$ という順に状態を巡回する方策がエキスパートの方策で，$s_0, s_3, s_2$ のいずれかの状態，または複数の状態の報酬が大きい場合に，エキスパートの方策が最適となる．

EVD の評価には，状態遷移確率を一様分布によって定めた 100 個の環境を用いる．EVD の評価に用いる環境の状態数は 4，行動数は 2 で，報酬の推定に用いた環境と同じである．次に，エキスパートの軌跡について説明する．報酬の推定には，エキスパートの方策から生成された 5 ステップの軌跡を合計 8 個用いる．具体的には，BIRL で報酬を推定する際には，一つの環境でエキスパートの軌跡を 8 個，BIRL-MD で報酬を推定する際には，環境 A と環境 B からそれぞれ 4 個，計 8 個の軌跡を用いて報酬を推定する．エキスパートの軌跡を生成

する方策には，$\gamma = 0.95$ の最適行動価値 $Q^*$ に基づく $\epsilon = 0.1$ の $\epsilon$-greedy 方策を用いた．

BIRL と BIRL-MD で報酬を推定する際には，式 (5.2), (6.2) のパラメータ $\kappa$，事前分布 $P(R)$，MCMC の各種パラメータの値を定める必要がある．ここでは，式 (6.2) のハイパーパラメータ $\kappa$ を 1.0，事前分布 $P(R)$ を一様分布 $\mathcal{U}(-1, 1)$ とした．また，MCMC のステップ数を 1000，バーンインを 200，ステップサイズ $\delta$ を 0.05 とし，収束判定は標本系列の時系列プロットによる可視化に基づいて行なった．また，EVD の計算にはサンプルした報酬の平均値を用いた．

既存手法，提案法の推定報酬の EVD の平均値と最大値を 5.1 に示す．

表 5.1: BIRL（単一環境）と BIRL-MD（環境数 2）の比較

| Method | BIRL | BIRL-MD |
|---|---|---|
| Number of environments | 1 | 2 |
| Total number of trajectories | 8 | 8 |
| EVD(Mean) | 0.559 | **0.172** |
| EVD(Max) | 4.540 | **1.188** |

5.1 の一行目は報酬の推定に用いた手法を，二行目は報酬の推定に用いた環境の数を，三行目は報酬の推定に用いた軌跡の合計の数を示す．また，四行目は転移先の 100 個の環境における EVD の平均値を，五行目は EVD の最大値を示す．5.1 の二列目に示す BIRL の EVD の平均値は，図 5.2 の各環境 A，B で推定した二つの報酬の EVD の平均値で，三列目に示す BIRL-MD の EVD は，二つの環境におけるエキスパートの軌跡から推定した報酬の EVD である．5.1 から分かるように，エキスパートの方策が最適となる報酬が複数存在することが明らかな環境において，複数の環境におけるエキスパートの軌跡を用いることによって，EVD の小さい報酬が推定された．EVD が小さいほど推定報酬がエキスパートの報酬に近いことから，ある報酬に従うエキスパートが複数の環境で生成した軌跡を用いて報酬を推定できる BIRL-MD が有用であると言える．

### Windy Grid World Environments

本実験では，風が吹く Grid world 環境である Windy grid world 環境を用いる [Sutton 18]. Windy grid world 環境におけるエージェントは，風が吹く状態ではエージェントの行動に関わらず，ある一定の確率で風が吹く方向に遷移する．そのため，各状態の風向きを変えることによって，状態遷移確率が異なる環境を生成できる．本実験では，エージェントの行動に関わらず風が吹く方向に遷移する確率を 30%とする．風が吹く方向は上下左右の 4 方向で，無風の状態も存在し，各状態の風向きは独立とする．

本実験で用いる $5 \times 5$ マスの Windy grid world の例を図 6.1 に示す．

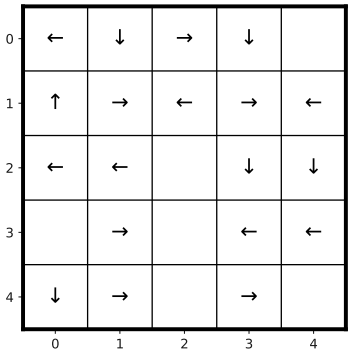


図 5.3: Windy grid world 環境の例

図 6.1 の各マス目の矢印は，矢印のある状態で吹く風の方向を示す．矢印がない状態は無風で，エージェントが選択した行動に従い確率 1 で遷移する．風向きの数は 4 方向と無風を合わせた 5 パターンであることから，環境の数は $5^{|S|} = 5^{25}$ パターン存在する．本節では，Windy grid world 環境を用いて，二つの実験を行う．一つ目はエキスパートのデータ数を変える実験で，データ数に対する BIRL と BIRL-MD の性能を比較する．二つ目はモデルのパラメータ $\kappa$ を変える実験で，異なる $\kappa$ を推定に用いた際の既存法と提案法の性能を比較する．

エキスパートの報酬を図 6.2 に示す．

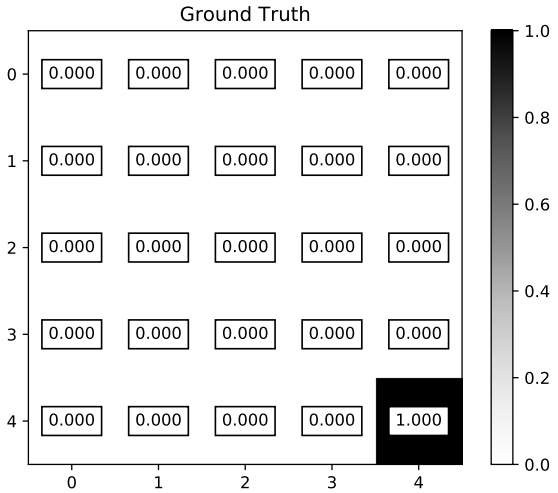



図 5.4: Windy grid world 環境のエキスパート報酬

エキスパートの報酬は最も右下の状態，座標 (4,4) の報酬が 1.0 で，それ以外の状態における報酬が 0 の一般的な Grid world 環境の報酬設計である．この報酬の下での最適方策は，右下のゴールと反対向きに風が吹く状態を避けつつゴールに向かう方策である．そのため，エキスパートの方策は，その環境の風向きによって異なる．EVD の評価には，各状態の風向きを一様分布で生成した 100 個の環境を用いる，また，各実験の評価値には 10 試行の平

均値を用いた．エキスパートの軌跡は，割引率 $\gamma = 0.7$，$\epsilon = 0.1$ の $\epsilon$-greedy 方策を用いて生成した．各軌跡のステップ数は 15 ステップとした．MCMC の各パラメータは，ステップ数を 20000，バーンインを 5000，ステップサイズ $\delta$ を 0.05 とし，収束判定は標本系列の時系列プロットによる可視化に基づいて行なった．EVD の計算には，サンプルした報酬の平均値を用いた．BIRL と BIRL-MD で報酬の推定に用いる事前分布 $P(R)$ は一様分布 $\mathcal{U}(-1, 1)$ とした．なお，報酬の事前分布を標準正規分布にして実験を行なったが，一様分布の場合と同様の結果が得られた．

報酬に用いるエキスパートの軌跡の合計数を変更した計算機実験の結果を図 6.3 に示す．

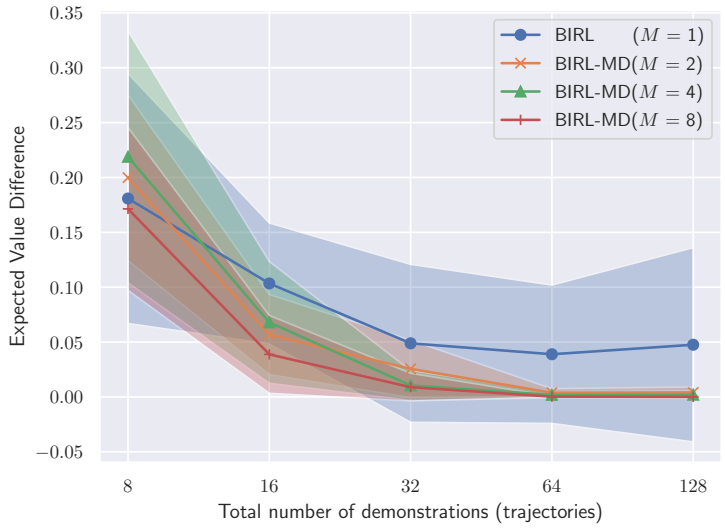



図 5.5: エキスパートの軌跡の合計数に対する BIRL と BIRL-MD の推定報酬の EVD の比較（10 試行平均）．$M$ はエキスパートの報酬推定に用いた環境の数．

ここで，式 (6.2) の $\kappa$ は 1.0 とし，エキスパートの軌跡の数は，報酬の推定に用いる各環境で均等に得られるものとした．図 6.3 のプロットの値は，EVD の平均値で，プロットと同じ色の領域は標準偏差を示す．BIRL，BIRL-MD のどちらもエキスパートのデータ数の増加に従い EVD が減少している．軌跡の数が 8 個の場合，報酬の推定に用いた環境の数が 2 個や 4 個の時の EVD の平均値が，環境の数が 1 個の時の EVD よりも大きな値を取っている．しかし，軌跡の数が 16 個以上の場合には，複数の環境を用いた方が EVD が小さく，複数の環境のエキスパートのデータを用いることによって，よりエキスパートに近い報酬が推定されていることが確認できる．このように，エキスパートの軌跡の数が一定数より多い場合は，複数の環境におけるエキスパートの軌跡を用いることによって，より EVD が小さい報酬が推定できると言える．

次に，式 (5.2), (6.2) のエキスパートが最適行動を取る確率を調節するパラメータ $\kappa$ の値を変更した実験の結果を図 5.6 に示す．

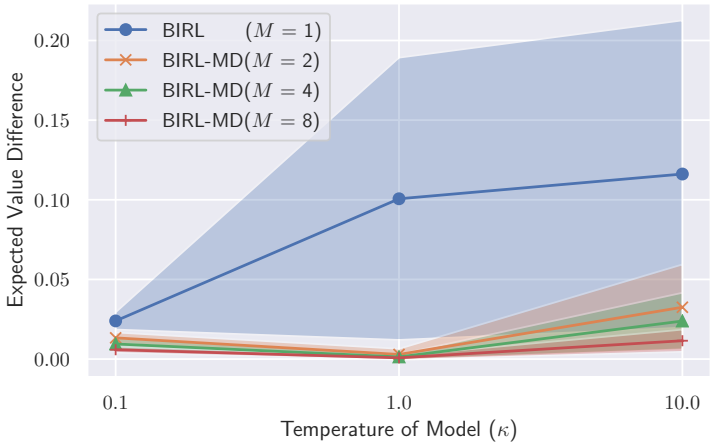
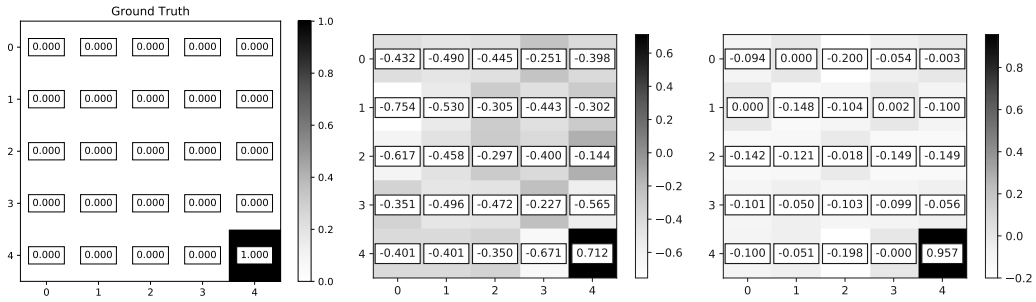


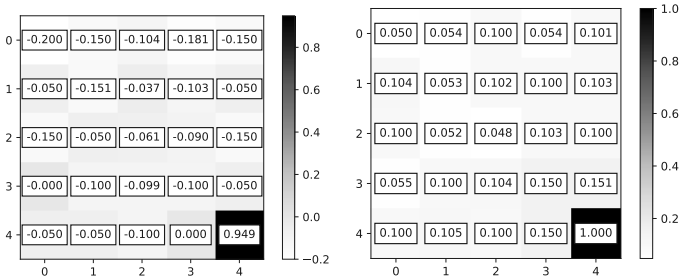

図 5.6: ボルツマン分布のパラーメータ $\kappa$ に対する BIRL と BIRL-MD の感度分析（10 試行平均）.

ここで，報酬の推定にはエキスパートの軌跡を合計 128 個用いた．図 5.6 の結果から，推定に用いるモデルのパラメータの値に依らず BIRL-MD が BIRL よりも小さい EVD の報酬を推定している．また，報酬の推定に用いる環境の数の増加に従い EVD が減少していることも確認できる．よって，複数の環境におけるエキスパートの軌跡から報酬を推定する提案法は，報酬推定のモデルのパラメータの変化に対して頑健だと言える．

次に，エキスパートの報酬と，各環境の数の下での推定報酬（サンプルした報酬の平均値）を図 5.7 に示す．




(a) Expert's rewards (b) BIRL-MD (M=1) (c) BIRL-MD (M=2)




(d) BIRL-MD (M=4) (e) BIRL-MD (M=8)

図 5.7: エキスパートの報酬と各環境数の下での推定報酬の比較

いずれの推定報酬も，右下の状態の報酬が高く推定されている．しかし，報酬の推定に用いた環境の数が 1 個，2 個の場合には推定報酬がまだらで，環境を 4 個，8 個用いて推定報酬の方がエキスパートの報酬に類似している．よって，定性的にも複数の環境のデータを用いることが有用であることが確認できる．

## 5.5 考察

複数の環境で得られたエキスパートの軌跡データを用いることによって，EVD が小さい報酬が推定された理由は，一つの環境から得られるエキスパートの報酬に関する情報には限りがあり，複数の環境におけるエキスパートの軌跡を用いることによって，エキスパートの報酬に関する情報をより多く得られるからだと考えられる．例えば，図 5.2 に示す二つの環境では，各環境においてエキスパートの報酬に関して得られる情報が異なる．具体的には，図 5.2 のうち，どちらか片方の環境におけるエキスパートの軌跡から得られる情報は，全 4 状態のうち 3 状態（環境 A では $s_1, s_3, s_2$，環境 B では $s_0, s_3, s_2$）のいずれかの状態の報酬が高いという情報である．二つの環境の情報を統合することによって，$s_2, s_3$ のいずれかの状態，または両方の状態の報酬が高いことは明らかである．

実際，提案法は複数の環境で得られたエキスパートの軌跡データを用いて報酬を推定することによって，一つの環境を対象とする既存法よりも EVD の小さい報酬を推定している．これは Windy grid world 環境においても同様である．ある一つの環境における軌跡だけでは，エキスパートがある状態を訪れた理由が報酬が高いからなのか，風向きの影響なのか判

別することは困難である．しかし，複数の環境のデータを用いることができれば，任意の状態遷移確率の環境において，エキスパートが高い頻度で訪れる状態が明確になり，エキスパートが訪れる頻度の高い状態には大きな報酬が存在すると推測できる．式 (6.2) で得られる事後分布は，多くの環境でエキスパートに共通する状態行動対の Q 値が高くなる報酬に高い確率を置くようモデル化されているため，エキスパートが複数の環境で生成した軌跡を用いることによって，より EVD が小さい，つまりエキスパートに近い報酬が推定されたと考えられる．

## 5.6 まとめ

本研究では,「あるタスクの解法を知るエキスパートが，状態遷移確率の異なる複数の環境において生成した軌跡データ」を所与として，そのエキスパートの報酬分布を推定する問題を，BIRL の枠組みで定式化し，マルコフ連鎖モンテカルロ法を用いた解法を提案した．

実験では，逆強化学習手法の性能の評価指標 EVD[Levine 11] を用いて提案法と既存手法を比較した．その結果，二種類の実験設定において，提案法の推定報酬が，既存手法の推定報酬よりもエキスパートの報酬に近いことが確認された．

# 第6章　複数のマルコフ決定過程におけるミニバッチベイジアン逆強化学習

複数の環境におけるミニバッチベイジアン逆強化学習手法は，複数の環境におけるベイジアン逆強化学習手法の計算量を削減する．具体的には，複数環境のベイジアン逆強化学習の各マルコフ連鎖モンテカルロステップにおける動的計画法の回数を，エキスパートの軌跡が得られた環境の数よりも小さい任意のミニバッチ数に削減する．

## 6.1　対象問題

BIRL-MD は，状態遷移確率が異なる環境 $E_m = \langle \mathcal{S}, \mathcal{A}, T_m, \gamma \rangle$ と，エキスパートが各環境 $E_m$ で生成したデータセット $D_m$ の集合 $\{(E_m, D_m)\}_{m=1}^M$ を所与として，報酬の事後分布 $P(R|\{(D_m E_m)\}_{m=1}^M)$ を推定する問題である．

報酬の事後分布 $P(R|\{(D_m, E_m)\}_{m=1}^M)$ は，ベイズの定理を用いて，

$$P(R|\{(D_m, E_m)\}_{m=1}^M) = \frac{\Pi_{m=1}^M P(D_m|R, E_m)}{\Pi_{m=1}^M P(D_m|E_m)} P(R), \tag{6.1}$$

と表される．式 (5.3) の $Q$ 値を環境 $E$，報酬 $R$ の下での最適方策の行動価値 $Q^*(s, a, R, E)$ へと拡張すれば，報酬の事後分布は

$$P(R|\{(D_m, E_m)\}_{m=1}^M) = \frac{1}{Z} \exp\left( \frac{1}{\kappa} \sum_{m=1}^M \sum_{(s,a)\in D_m} Q^*(s, a, R, E_m) \right) P(R), \tag{6.2}$$

と表すことができる．BIRL-MD は式 (6.2) を学習する問題である．

文献は，式 (6.2) に示す事後分布から報酬をサンプリングする MCMC アルゴリズムを提案している．具体的には，対称性を有する確率分布 $P(\tilde{R}|R) = P(R|\tilde{R})$ に従いサンプルされた $\tilde{R}$ を確率，

$$\min\left\{ 1, \frac{P\left(\tilde{R}, \{(D_m, E_m)\}_{m=1}^M\right)}{P\left(R, \{(D_m, E_m)\}_{m=1}^M\right)} \right\} \tag{6.3}$$

で採択することによって，事後分布 $P(R|\{(D_m, E_m)\}_{m=1}^M)$ を実現する．事後分布 $P(R|\{(D_m, E_m)\}_{m=1}^M)$ からの報酬のサンプルには，各 MCMC ステップで式 (6.4) の右辺の計算を要する．

$$P(R|\{(D_m, E_m)\}_{m=1}^M) \propto \exp\left( \frac{1}{\kappa} \sum_{m=1}^M \sum_{(s,a)\in D_m} Q^*(s, a, R, E_m) \right) P(R) \tag{6.4}$$

式 (6.4) の右辺には，$M$ 個の環境における $Q$ 値が含まれているため，各 MCMC ステップで $M$ 個の環境において動的計画法や強化学習を行う必要がある．

## 6.2 アプローチ

BIRL-MD は，各 MCMC ステップで式 (6.4) を計算する必要があるため，各 MCMC ステップで $M$ 個の環境における動的計画法を要する．したがって，BIRL-MD の計算量は環境の数 $M$ に対して線形に増加し，多くの環境から得られた軌跡から報酬を推定することは困難である．

そこで本論文では，環境の数が増加しても報酬の推定に要する計算量が変わらないアルゴリズム，Mini-batch BIRL-MD を提案する．BIRL-MD が各 MCMC ステップで $M$ 回の動的計画法を要する理由は，報酬 $\tilde{R}$ を採択する確率が $\min\left\{1, \frac{P\left(\tilde{R},\{(D_m,E_m)\}_{m=1}^{M}\right)}{P\left(R,\{(D_m,E_m)\}_{m=1}^{M}\right)}\right\}$ で表され，$\frac{P\left(\tilde{R},\{(D_m,E_m)\}_{m=1}^{M}\right)}{P\left(R,\{(D_m,E_m)\}_{m=1}^{M}\right)}$ の計算に $M$ 個の環境における動的計画法を要するからである．以下で，この採択確率を $M$ 個の環境から，$N(N \leq M)$ 個のサンプル環境（これをミニバッチと呼ぶ）に対する動的計画法によって近似する方法について述べる．

報酬 $\tilde{R}$ の採択確率 $\min\left\{1, \frac{P\left(\tilde{R},\{(D_m,E_m)\}_{m=1}^{M}\right)}{P\left(R,\{(D_m,E_m)\}_{m=1}^{M}\right)}\right\}$ は，式 (6.5) と表すことができる．

$$u < \min\left\{1, \frac{P(\tilde{R})\prod_{i=1}^{M} P(D_i,E_i|\tilde{R})}{P(R)\prod_{i=1}^{M} P(D_i,E_i|R)}\right\} \tag{6.5}$$

ここで，$u$ は一様分布 $\mathcal{U}(0,1)$ からのサンプルである．式 (6.5) は，下のように展開できる [Bardenet 17]．

$$u < \frac{P(\tilde{R})\prod_{i=1}^{M} P(D_i,E_i|\tilde{R})}{P(R)\prod_{i=1}^{M} P(D_i,E_i|R)} \tag{6.6}$$

$$u\frac{P(R)}{P(\tilde{R})} < \prod_{i=1}^{M} \frac{P(D_i,E_i|\tilde{R})}{P(D_i,E_i|R)} \tag{6.7}$$

$$\log\left(u\frac{P(R)}{P(\tilde{R})}\right) < \sum_{i=1}^{M}\left\{\log\frac{P(D_i,E_i|\tilde{R})}{P(D_i,E_i|R)}\right\} \tag{6.8}$$

$$\frac{1}{M}\log\left(u\frac{P(R)}{P(\tilde{R})}\right) < \frac{1}{M}\sum_{i=1}^{M}\left\{\log\frac{P(D_i,E_i|\tilde{R})}{P(D_i,E_i|R)}\right\} \tag{6.9}$$

式 (6.9) の右辺は，$\log P(D_i,E_i|\tilde{R}) - \log P(D_i,E_i|R)$ の平均値である．そのため，式 (6.10) を用いれば，$M$ 個より少ない $N$ 個のデータで近似できる．

$$\frac{1}{M}\sum_{i=1}^{M}\left\{\log\frac{P(D_i,E_i|\tilde{R})}{P(D_i,E_i|R)}\right\} \approx \frac{1}{N}\sum_{i=1}^{N}\left\{\log\frac{P(D_i,E_i|\tilde{R})}{P(D_i,E_i|R)}\right\} \tag{6.10}$$

ここで，式 (6.10) の近似が成り立つ理由を述べる．式 (6.10) の左辺と右辺の差は，標本平均の標準誤差に相当する．具体的には，左辺の標準偏差 $\sigma$ の下での標準誤差は，式 (6.11) と表される．

$$\sqrt{\frac{M-N}{M}}\frac{\sigma}{\sqrt{N}} \tag{6.11}$$

式 (6.11) の値はミニバッチ数 $N$ の増加に従い単調に減少し，$M = N$ のときに 0 となる．標準誤差がミニバッチ数 $N$ の増加に従い単調減少する性質は，全環境数 M やミニバッチ数 N の大きさに依らないことに留意されたい．

式 (6.9) は式 (6.10) を用いて，次のように近似できる．

$$\frac{1}{M}\log\left(u\frac{P(R)}{P(\tilde{R})}\right) < \frac{1}{N}\sum_{i=1}^{N}\left\{\log\frac{P(D_i,E_i|\tilde{R})}{P(D_i,E_i|R)}\right\} \tag{6.12}$$

式 (6.12) におけるミニバッチ数 $N$ は $M$ 以下の数に定めることができ，各 MCMC ステップで要する動的計画法の数は $O(N)$ である．報酬の事後分布を実現するアルゴリズム，Mini-batch BIRL-MD を Algorithm 5 に示す．

**Algorithm 5** Mini-batch Bayesian Inverse Reinforcement Learning for Multiple Dynamics

**INPUT:** Environments $\{E_m\}_{m=1}^{M}$, Demonstrations $\{D_m\}_{m=1}^{M}$, Prior $P(R)$, Step Size $\delta$, Mini-batch Size $N$

**OUTPUT:** Sampled Rewards $\{R_i\}_{i=1}^{t}$

Pick a random vector $R \in \mathbb{R}^{|S|}/\delta$
$\{\pi_m\}_{m=1}^{M} \leftarrow \{\text{PolicyIteration}(E_m, R)\}_{m=1}^{M}$
**for** $i = 1$ **do** $t$
    Pick a reward vector $\tilde{R}$ uniformly at random from the neighbours of $R \in \mathbb{R}^{|S|}/\delta$
    $u \leftarrow$ Sample from Uniform distrbution $U(0,1)$
    $\tilde{\mathbb{N}} \leftarrow$ Sampled $N$ integers from $\{n \in \mathbb{N} |\ n \leq M\}$ without repetition
    Compute $Q^{\pi}(s,a,R,E) \quad \forall\{s,a,(E_n,\pi_n)\} \in \mathcal{S}\times\mathcal{A}\times\{(E_n,\pi_n)\}_{n\in\tilde{\mathbb{N}}}$
    **if** $\exists\{s,a,(E,\pi)\} \in \mathcal{S}\times\mathcal{A}\times\{(E_n,\pi_n)\}_{n\in\tilde{\mathbb{N}}}, Q^{\pi}(s,\pi(s),\tilde{R},E) < Q^{\pi}(s,a,\tilde{R},E)$ **then**
    ▷ If any sampled policy is not optimal
        $\{\tilde{\pi}_n\}_{n\in\tilde{\mathbb{N}}} \leftarrow \left\{\text{PolicyIteration}(E_n,\tilde{R})\right\}_{n\in\tilde{\mathbb{N}}}$
        **if** $\frac{1}{M}\log\left(u\frac{P(R)}{P(\tilde{R})}\right) < \frac{1}{N}\sum_{n\in\tilde{\mathbb{N}}}\left\{\log P\left(D_n,E_n\middle|\tilde{R}\right) - \log P\left(D_n,E_n\middle|R\right)\right\}$ **then**
            $R \leftarrow \tilde{R}$
            $\{\pi_n\}_{n\in\tilde{\mathbb{N}}} \leftarrow \{\tilde{\pi}_n\}_{n\in\tilde{\mathbb{N}}}$
        **end if**
    **else**
        **if** $\frac{1}{M}\log\left(u\frac{P(R)}{P(\tilde{R})}\right) < \frac{1}{N}\sum_{n\in\tilde{\mathbb{N}}}\left\{\log P\left(D_n,E_n\middle|\tilde{R}\right) - \log P\left(D_n,E_n\middle|R\right)\right\}$ **then**
            $R \leftarrow \tilde{R}$
        **end if**
    **end if**
    $R_i \leftarrow R$
**end for**

Mini-batch BIRL-MD は，報酬 $R$ の採択基準を $N$ 個の環境における尤度を用いて近似するため，BIRL-MD と比較して計算量が少ない．既存手法と Mini-batch BIRL-MD の比較を表 6.1 に示す．

表 6.1: 既存手法と提案法の比較．$M$ はエキスパートが軌跡を生成した環境の数，$N$ はミニバッチ数を指す．

| | ベイジアン逆強化学習 [Ramachandran 07] | ベイジアン逆強化学習（複数環境） | ミニバッチ ベイジアン逆強化学習（複数環境) |
|---|---|---|---|
| Number of Dynamic Programming for each MCMC steps | $O(1)$ | $O(M)$ | $O(N)$ |
| Multiple environments | - | ✓ | ✓ |

## 6.3 予備実験

### 実験設定

本実験では，風が吹くグリッドワールド環境である Windy grid world 環境を用いる．Windy grid world 環境におけるエージェントは，風が吹く状態ではエージェントの行動に関わらず，ある一定の確率で風が吹く方向に遷移する．そのため，各状態の風向きを変えることによって，状態遷移確率が異なる環境を生成できる．本実験では，エージェントの行動に関わらず風が吹く方向に遷移する確率を 30%とする．風が吹く方向は上下左右の 4 方向で，無風の状態も存在し，各状態の風向きは独立とする．本実験で用いる $5 \times 5$ マスの Windy grid world の例を図 6.1 に示す．

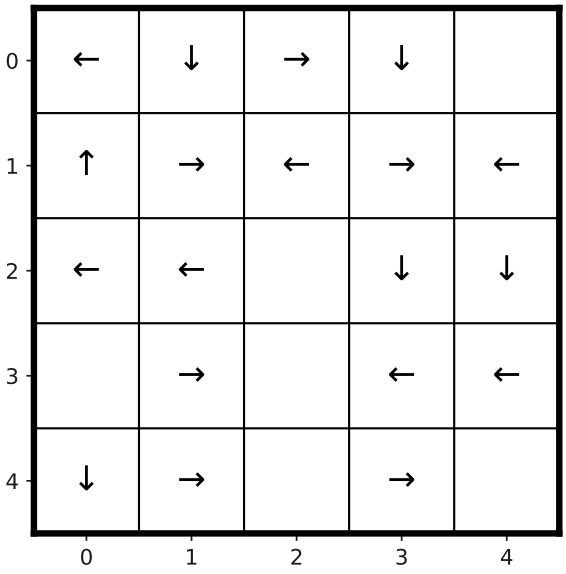



図 6.1: Windy Grid World 環境の例

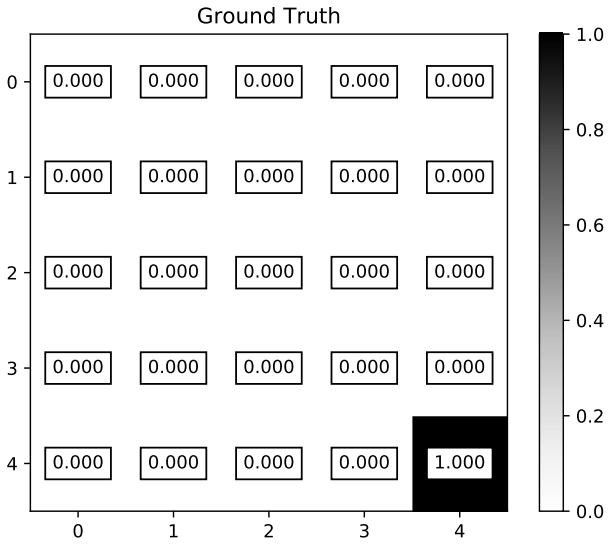



図 6.2: Windy Grid World 環境におけるエキスパートの報酬

図 6.1 の各マス目の矢印は，矢印のある状態で吹く風の方向を示す．矢印がない状態は無風で，エージェントが選択した行動に従い確率 1 で遷移する．風向きの数は 4 方向と無風を合わせた 5 パターンであることから，環境の数は $5^{|S|} = 5^{25}$ パターン存在する．エキスパートの報酬は最も右下の状態，座標 (4, 4) の報酬が 1.0 で，それ以外の状態における報酬が 0 の一般的な迷路問題の報酬とした．この報酬の下での最適方策は，右下のゴールと反対向きに風が吹く状態を避けつつゴールに向かう方策である．そのため，エキスパートの方策は，各環境の風向きによって異なる．

本実験では，提案法とベイジアン逆強化学習の推定報酬を，エキスパートの報酬との近さを表す指標である Expected Value Difference（EVD）[Levine 11] を用いて評価する．EVD の評価には，各状態の風向きを一様分布で生成した 100 個の環境を用いる，また，各実験の評価値には 10 試行の平均値を用いた．エキスパートの軌跡は，割引率 $\gamma = 0.7$，$\epsilon = 0.1$ の $\epsilon$-greedy 方策を用いて生成した．各軌跡のステップ数は 15 ステップで，エキスパートの軌跡は，報酬の推定に用いる各環境で均等に得られるものとした．ベイジアン逆強化学習と BIRL-MD で報酬の推定に用いる事前分布 $P(R)$ は一様分布 $\mathcal{U}(-1, 1)$ とした．MCMC の各パラメータは，ステップ数を 20000，バーンインを 5000，ステップサイズ $\delta$ を 0.05 とした．また，式 (6.2) の $\kappa$ は 1.0 とした．EVD の計算には，サンプルした報酬の平均値を用いた．

### 実験結果

報酬に用いるエキスパートの軌跡の合計数を変更した計算機実験の結果を図 6.3 に示す．

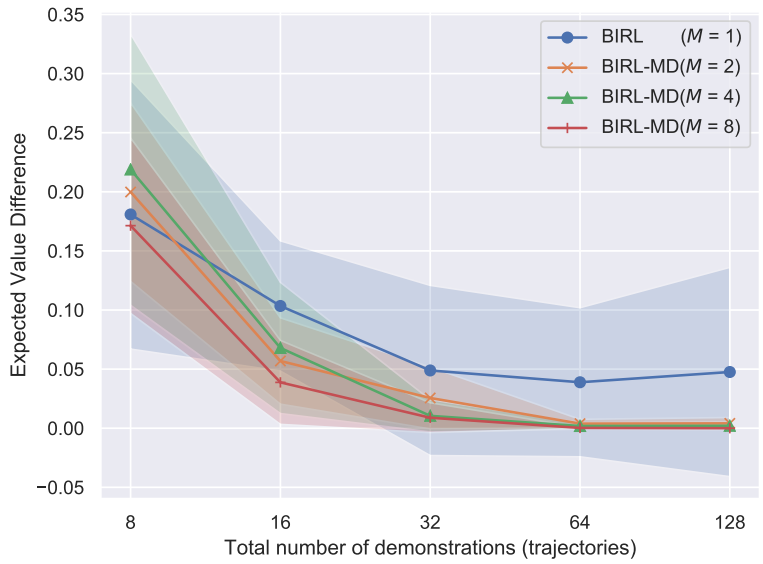



図 6.3: エキスパートが軌跡を生成した環境の数に対する EVD の評価

図 6.3 のプロットの値は，Eexpted Value Difference (EVD) の平均値で，プロットと同じ色の領域は標準偏差を示す．ベイジアン逆強化学習，BIRL-MD のどちらもエキスパートのデータ数の増加に従い EVD が減少している．エキスパートが軌跡を生成した環境の数 $M$ の増加に伴い EVD が減少していることから，複数の環境の軌跡を報酬の推定に用いることが有用だと言える．

## 6.4 実験

予備実験と同様の環境を用いて複数環境におけるベイジアン逆強化学習とミニバッチベイジアン逆強化学習を比較する．実験は二つある．一つは，軌跡が得られる環境の数 $M = 8$ の下でミニバッチ数 $N$ を変えた実験で，もう一つは，ミニバッチ数 $N = 1$ の下で軌跡が得られる環境の数 $M$ を変えた実験である．前者の実験の結果を図 6.4 と表 6.2 に示す．

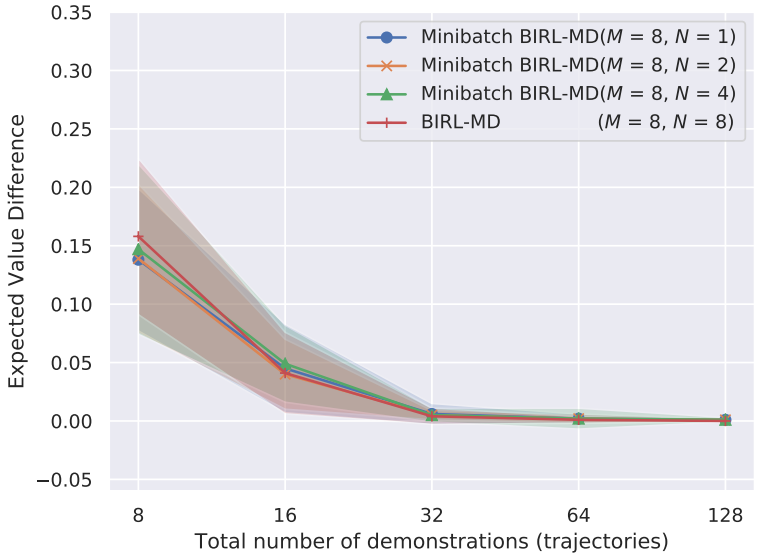



図 6.4: ミニバッチサイズ $N$ に対する Mini-batch BIRL-MD の性能評価

表 6.2: ミニバッチサイズ $N$ に対する Mini-batch BIRL-MD の EVD の評価 ($10^{-2}$)

| Mini-batch size | Total # of demonstrations | | | | |
|---|---|---|---|---|---|
| | 8 | 16 | 32 | 64 | 128 |
| 1 | 13.8±6.1 | 4.5±3.8 | 0.6±0.9 | 0.2±0.4 | 0.1±0.1 |
| 2 | 13.9±6.4 | 4.0±3.0 | 0.5±0.4 | 0.2±0.4 | 0.1±0.1 |
| 4 | 14.7±7.3 | 4.9±3.3 | 0.5±0.6 | 0.2±0.9 | 0.1±0.2 |
| 8 | 15.8±4.1 | 4.1±3.5 | 0.4±0.7 | 0.1±0.3 | 0.0±0.1 |

図 6.4 と表 6.2 から，環境の数 $M$ より小さいミニバッチ数 $N$ を用いても，推定報酬の EVD が大きく変化していないと言える．よって，ミニバッチ数 $N \leq M$ のサンプルで事後分布からのサンプリングが実現できていると言える．なお，環境の数 $M$ よりミニバッチ数 $N$ が小さい場合も，ミニバッチ近似を用いない場合と同様に，右下のゴールと反対向きに風が吹く状態を避けつつゴールに向かう方策を獲得した．

次に，ミニバッチ数 $N = 1$ の下で軌跡が得られる環境の数 $M$ を変化させた実験の結果を図 6.5 に示す．

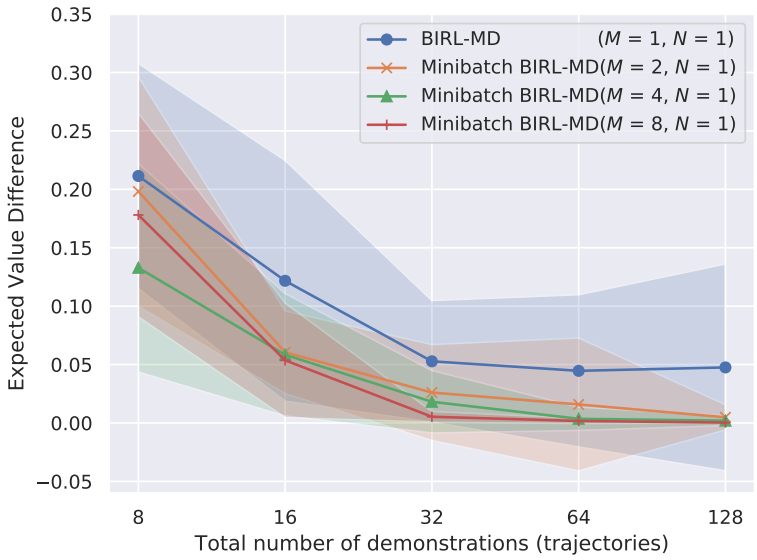
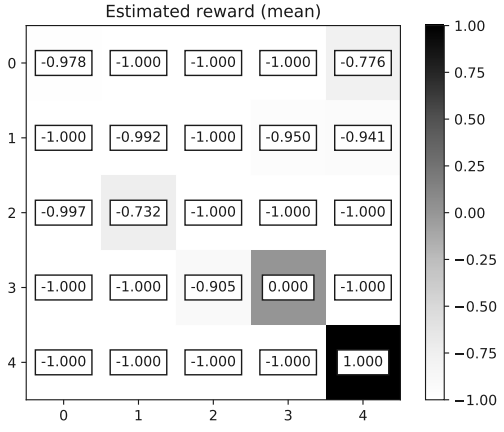


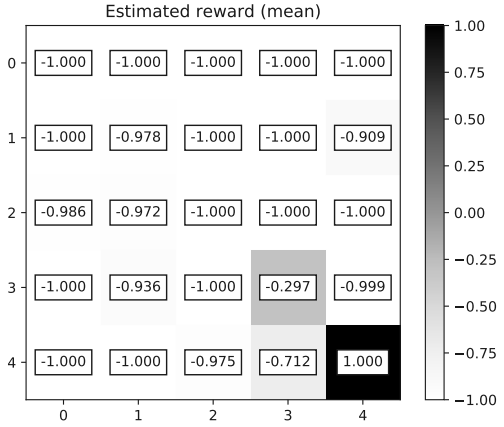

図 6.5: エキスパートが軌跡を生成した環境の数 $M$ に対する Mini-batch BIRL-MD の性能評価

図 6.5 では，環境の数の増加にしたがい EVD が小さくなっている．この結果は，あるミニバッチ数 $N$ の下で，複数の環境の軌跡を用いることの有用性を示している．

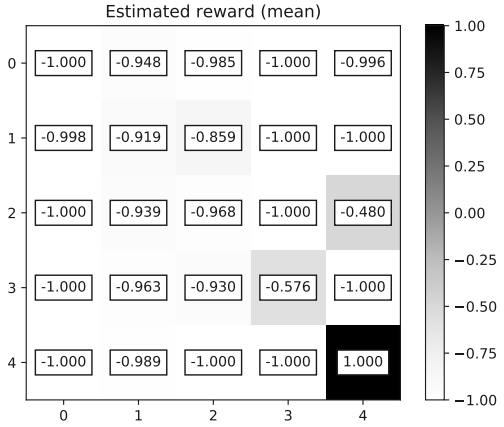
図 6.5 において得られた推定報酬のサンプルを図 6.6 に示す．



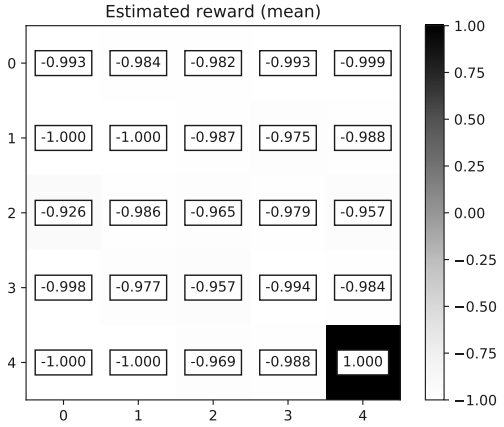

(a) 全環境数 $M = 1$，ミニバッチ数 $N = 1$（バッチ学習と等価）




(b) 全環境数 $M = 2$，ミニバッチ数 $N = 1$




(c) 全環境数 $M = 4$，ミニバッチ数 $N =$




(d) 全環境数 $M = 8$，ミニバッチ数 $N = 1$

図 6.6: 推定報酬の例．各状態の報酬の値は事後分布の下での報酬の平均値．

図 6.5 が示すように環境の数 $M$ の増加に従って推定報酬がエキスパートの報酬に近づいている．そのため，定性的にも Mini-batch BIRL-MD を用いて複数の環境の軌跡から報酬を推定することが有用であると言える．

## 6.5 考察

本章では，ミニバッチ数 $N = 1$ が推定に与える影響を考察する．本実験においては，ミニバッチ数 $N = 1$ の下で優れた報酬が推定された．これは，本実験において MCMC の採択条件が Mini-batch BIRL-MD で適切に近似できていることを示している．採択条件の近似の標準誤差は式 (6.11) で表され，式 (6.10) の左辺の分散に比例し，ミニバッチ数 $N$ の増加に従い単調減少する．したがって，式 (6.10) の左辺の分散が大きい環境やエキスパートデータに対しては，より多くのミニバッチ数 $N$ を設定する必要がある．

## 6.6 まとめ

本論文では，エキスパートが複数の環境で生成した軌跡から報酬を推定する逆強化学習問題の解法を提案した．提案法は，既存手法と比較して少ない計算量で実行可能なアルゴリズムである．具体的には，エキスパートが軌跡を生成した環境の数 $M$ に対して，既存手法の計算量が $O(M)$ のオーダーで増加するのに対し，ミニバッチ数 $N(N \leq M)$ に対して線形の計算量 $O(N)$ のアルゴリズムを提案した．実験では，一般的なベンチマーク問題であるグリッドワールドで既存手法と提案法を比較した．その結果，提案法が既存手法よりも少ない計算量で，既存手法と同等の報酬を推定できることが確認された．提案法の課題は，その適用範囲は離散状態行動空間の環境に限られてしまう点にある．つづく 7 章では，連続状態行動空間の複数環境におけるエキスパートの軌跡に適用可能な逆強化学習手法について記す．

# 第7章　複数のマルコフ決定過程における敵対的最大エントロピー逆強化学習

本章では，複数環境の軌跡から報酬を推定する敵対的最大エントロピー逆強化学習手法を提案する．提案法は，離散状態行動空間と連続状態行動空間に適用可能な敵対的学習を用いた最大エントロピー逆強化学習に基づいており，ここまで述べた定式化では扱うことのできない連続状態行動空間の複数の環境におけるエキスパートの軌跡から，報酬と方策を推定できる手法である．実験では，連続状態行動空間のプラント制御問題を対象として提案法の有用性を検証する．

## 7.1　準備

本節では提案法の基礎となる最大エントロピー逆強化学習と，敵対的最大エントロピー逆強化学習，そしてニューラルネットワークについて説明する．

### 7.1.1　最大エントロピー逆強化学習

Ziebart らが提案した最大エントロピー逆強化学習は，報酬 $R(s,a)$ を特徴量 $\phi(s,a) \mapsto \mathbb{R}^n$ を用いた線形関数でモデル化した [Ziebart 08]．線形の報酬関数の定義式を以下に示す．

$$R(s,a) = \theta \cdot \phi(s,a) \tag{7.1}$$

ここで $\theta \in \mathbb{R}^n$ は各特徴の重要度を表す報酬のパラメータである．線形の報酬関数を持つ環境の下で，エキスパート方策 $\pi_{\exp}$ と方策 $\pi$ の報酬の期待値が一致する十分条件は，特徴量の期待値が一致することである．特徴量の期待値の一致が報酬の期待値の一致の十分条件であることを以下に示す．

$$\mathbb{E}_{(s,a)\sim\pi_{\exp},T}\left[\phi(s,a)\right] = \mathbb{E}_{(s,a)\sim\pi,T}\left[\phi(s,a)\right] \tag{7.2}$$

$$\Rightarrow \theta \cdot \mathbb{E}_{(s,a)\sim\pi_{\exp},T}\left[\phi(s,a)\right] = \theta \cdot \mathbb{E}_{(s,a)\sim\pi,T}\left[\phi(s,a)\right] \tag{7.3}$$

$$\Rightarrow \mathbb{E}_{(s,a)\sim\pi_{\exp},T}\left[\theta \cdot \phi(s,a)\right] = \mathbb{E}_{(s,a)\sim\pi,T}\left[\theta \cdot \phi(s,a)\right] \tag{7.4}$$

$$\Rightarrow \mathbb{E}_{(s,a)\sim\pi_{\exp},T}\left[R(s,a)\right] = \mathbb{E}_{(s,a)\sim\pi,T}\left[R(s,a)\right] \tag{7.5}$$

したがって，最適方策とエキスパート方策の特徴期待値が一致する報酬を推定すれば，推定報酬に対する最適方策とエキスパート方策 $\pi_{\mathrm{exp}}$ の報酬の期待値が一致する．ここで問題なのは,「エキスパート方策との特徴期待値の一致」という制約を満たす方策は複数存在する点である．Ziebart らは，この制約を満たす方策の集合から一つの方策を選択する方法として最大エントロピー原理を用いた．最大エントロピー原理とは，特定の期待値を持つ確率分布から，エントロピーが最大の確率分布を選ぶアプローチである．最大エントロピー原理を用いて選択した解は以下の最適化問題の解となることが知られている [Grünwald 04b].

$$\begin{aligned} &\inf_{\mathrm{P}(X)} \sup_{\tilde{\mathrm{P}}(X)} -\sum_{X} \tilde{\mathrm{P}}(X) \log \mathrm{P}(X) \\ &\text{s.t.} \mathbb{E}_{\tilde{\mathrm{P}}(X)}\left[f(X)\right] = \mathbb{E}_{\mathrm{P}(X)}\left[f(X)\right] \end{aligned} \tag{7.6}$$

この最適化問題は，二つの確率分布 $\mathrm{P}(X)$，$\tilde{\mathrm{P}}(X)$ の期待値が一致するという制約条件の下で，一方のエージェントが $P(X)$ と距離が最大となる確率分布 $\tilde{\mathrm{P}}(X)$ を選択し，他方のエージェントが $\tilde{\mathrm{P}}(X)$ との距離が最小となる確率分布 $P(X)$ を選択する問題である．最大エントロピー原理を用いた解は式 (7.6) の最適化問題の解となるため，真の確率分布 $\tilde{\mathrm{P}}(X)$ が確率分布 $P(X)$ に対してワーストケースである可能性を考えた時には，エントロピーが最大となる確率分布を選択することが望ましいと言える [Grünwald 04b].

最大エントロピー逆強化学習問題の定式化を以下に示す.

$$\text{maximize} -\sum_{\tau \in \mathcal{Z}} P(\tau) \log P(\tau) \tag{7.7}$$

$$\text{subject to} : \sum_{\tau \in \mathcal{Z}} P(\tau)\phi(\tau) = \mathbb{E}_{P_{\pi_{\mathrm{exp}}}(\tau)}[\phi(\tau)] \tag{7.8}$$

$$\sum_{\tau \in \mathcal{Z}} P(\tau) = 1 \tag{7.9}$$

$$P(\tau) \geq 0 \quad \forall \tau \in \mathcal{Z} \tag{7.10}$$

$$\mathcal{Z} = \{\tau_1, \ldots, \tau_m\} \tag{7.11}$$

ここで，$\tau$ はエピソードの軌跡 $\tau = (s_{t=0}, a_{t=0}, s_{t=1}, a_{t=1}, \ldots)$ を表す．式 (7.7) はエントロピーの最大化を意味し，式 (7.8) は特徴期待値の一致の条件，その他の制約は $P$ が確率分布の性質を満たす条件である．上記の制約付き最適化問題をラグランジュ法を用いて $P$ について解くと，以下の式が得られる．

$$\begin{aligned} P(\tau \mid \theta) = \frac{\exp(R(\tau))}{\sum_{\tau \in \mathcal{Z}} \exp(R(\tau))} = \frac{\exp(R(\tau))}{Z} \\ \text{where } R(\tau) = \theta \cdot \sum_{s \in \tau} \phi(s) \end{aligned} \tag{7.12}$$

ここで $\theta$ はラグランジュ変数である．式 (7.12) は，ある軌跡 $\tau$ が報酬のパラメータ $\theta$ の下で得られる確率は，報酬 $R(\tau)$ に指数比例することを示している．この式に基づいてエキスパートの軌跡のデータセット $\mathcal{D} = \{\tau_1, \tau_2, \ldots, \tau_N\}$ に対して対数尤度最大化問題を解くことで，エキスパートの報酬が推定できる．尤度最大化問題の目的関数を以下に示す．

$$L(\theta) = \sum_{\tau \in \mathcal{D}} \log P(\tau \mid \theta) \tag{7.13}$$

対数尤度は報酬のパラメータ $\theta$ について微分可能であり，勾配法を用いて解くことができる．対数尤度に対する勾配を以下に示す．

$$\log P(\tau \mid \theta) = R(\tau) - \log Z \tag{7.14}$$

$$\nabla \log P(\tau \mid \theta) = \phi(\tau) - \mathbb{E}_{P(\tau|\theta)}[\phi(\tau)] \tag{7.15}$$

式 (7.15) を式 (7.13) に代入すれば，目的関数に対する勾配が得られる．

$$\nabla L(\theta) = \sum_{\tau \in \mathcal{D}} \nabla \log P(\tau \mid \theta) \tag{7.16}$$

$$= \mathbb{E}_{P_{\pi_E}(\tau)}[\phi(\tau)] - \mathbb{E}_{P(\tau|\theta)}[\phi(\tau)] \tag{7.17}$$

目的関数に関する勾配は，エキスパートの特徴期待値と推定報酬に対する最適方策の特徴期待値の差である．最大エントロピー逆強化学習では，この勾配を用いてエキスパートの報酬を推定する．最大エントロピー逆強化学習のアルゴリズムを以下に示す．

**Algorithm 6** Maximum Entropy Inverse Reinforcement Learning

**INPUT:** Environment $E$, Demonstrations $\mathcal{D} = \{\tau_1, \tau_2, \ldots, \tau_N\}$, Learning rate $\alpha$
**OUTPUT:** Parameter of reward function $\theta$

1: Initialize parameter $\theta$
2: **for** step $n$ in $\{1 \ldots N\}$ **do**
3: Compute optimal policy $\pi^*_\theta$ for $R(s, a) = \theta \cdot \phi(s, a)$ via reinforcement learning.
4: Compute gradient $\nabla L(\theta) = \mathbb{E}_{P_{\pi_E}(\tau)}[\phi(\tau)] - \mathbb{E}_{P_{\pi^*_\theta}(\tau)}[\phi(\tau)]$
5: Update parameter $\theta = \theta - \alpha \cdot \nabla L(\theta)$
6: **end for**

Ziebart の最大エントロピー逆強化学習は各ステップ $n$ で推定報酬に対する最適方策を求める必要があり，最適方策の計算がアルゴリズムのボトルネックとなっている．このボトルネックを解消した逆強化学習手法について次節で述べる．

### 7.1.2 敵対的最大エントロピー逆強化学習

敵対的学習とは，二つのモデル（例: ニューラルネットワーク）が共通の目的関数を持ち，一方のモデルが目的関数を最大化，他方のモデルが目的関数を最小化する学習法である．敵

対的逆強化学習では，識別器がエキスパートの状態遷移と行動のデータ $(s, a, s')$ と学習中のエージェントの状態遷移と行動のデータの分類精度を最大化する一方で，学習中のエージェントは識別器の分類精度を最小化するように学習する．この学習を通して，エージェントがエキスパートの方策を学習することが出来れば，識別器はエージェントとエキスパートの状態遷移と行動のデータ $(s, a, s')$ を分類不能になる．

Fu らは，この敵対的逆強化学習の推定報酬が最大エントロピー逆強化学習の推定結果と一致することを示し，これを利用した Adversarial Inverse Rienforcement Learning を提案した [Fu 18]．敵対的な学習を行う逆強化学習手法 Adversarial Inverse Reinforcement Learning（AIRL）の概要を図 7.1 に示す．

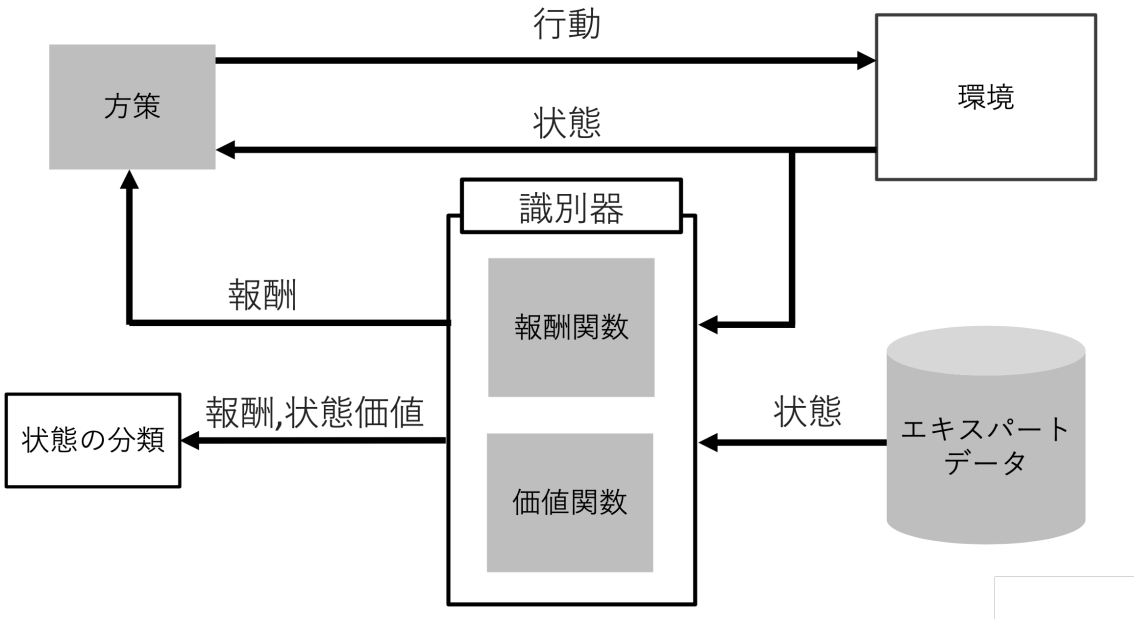



図 7.1: Adversarial Inverse Reinforcement Learing（AIRL）の概要

Adversarial Inverse Rienforcement Learning（AIRL）の識別器 $Disc$ の構成を以下に示す．

$$Disc_{\theta,\phi}\left(s, a, s'\right) = \frac{\exp\left\{f_{\theta,\phi}\left(s, a, s'\right)\right\}}{\exp\left\{f_{\theta,\phi}\left(s, a, s'\right)\right\} + \pi(a \mid s)} \tag{7.18}$$
$$\text{where } f_{\theta,\phi}\left(s, a, s'\right) = r_\theta(s) + \gamma V_\phi\left(s'\right) - V_\phi(s)$$

識別器の出力は $[0, 1]$ で，出力は入力 $(s, a, s')$ がエキスパートのデータである確率である．AIRL は式 (7.18) に示した識別器 $Disc$ を用いて，次のミニマックス問題を解くことによって，エキスパートの方策と報酬を学習する．

$$\underset{\pi}{\text{minimize}} \max_{Disc} \mathbb{E}_\pi[\log(1 - Disc(s, a, s'))] + \mathbb{E}_{\pi_E}[\log(Disc(s, a, s'))] - \lambda H(\pi) \tag{7.19}$$

ここで，$H(\pi)$ は方策 $\pi$ のエントロピーである．

**Algorithm 7** Adversarial Inverse Reinforcement Learning

**INPUT:** Environment $E$, Demonstrations $\mathcal{D} = \{\tau_1, \tau_2, \ldots, \tau_N\}$
**OUTPUT:** Reward function $r_\theta$
1: Initialize policy $\pi$ and $Disc_{\theta,\phi}$
2: **for** step $n$ in $\{1 \ldots N\}$ **do**:wherejh N
3: Collect trajexctories $\tau_i = (s_0, a_0, \ldots, s_T, a_T)$ by executing $\pi$
4: Train $Discriminator_{\theta,\phi}$ via binary logistic regression to classify expert data $\tau_{\mathrm{exp},i}$ from samples $\tau_i$.
5: Update policy $\pi$ with respect to $r_\theta$ via policy optimization method.
6: **end for**

AIRL と Ziebart の最大エントロピー逆強化学習の違いについて説明する，Ziebart の最大エントロピー逆強化学習は各ステップ $n$ で推定報酬に対する最適方策を求める必要があり，これが計算のボトルネックであった，一方 AIRL は，各ステップ $n$ で推定報酬に対する最適方策を求める必要はなく，推定報酬に対して方策を更新すれば良い．この違いから，AIRL は最大エントロピー逆強化学習と比較して計算効率に優れたアルゴリズムとなっている．

### 7.1.3 ニューラルネットワーク

ニューラルネットワークは，人間をはじめとする動物の脳神経系 (cerebral nerve system) を模した情報処理の数理的モデルである [Goodfellow 16]．ディープニューラルネットワーク (DNN) や畳み込みニューラルネットワーク (CNN) をはじめとする多層順伝搬型ネットワークは高い学習能力と汎化性能で知られる．本章では AIRL や提案法で用いるニューラルネットワークの基本事項についてまとめる．

#### 順伝搬型ネットワークの構造

順伝搬型ネットワーク（feedforward neural network）は層状に並べたユニットが隣接層間での結合した構造を持ち，情報が入力側から出力側に一方向にのみ伝搬するニューラルネットワークである．ニューラルネットワークを構成する層と，ネットワークを構成するユニットのモデルをそれぞれ図 7.2, 7.3 に示す．

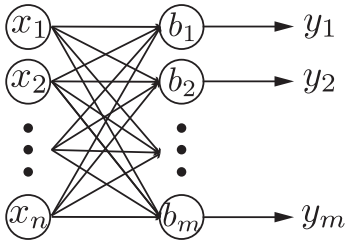



図 7.2: ニューラルネットワークの層の構造

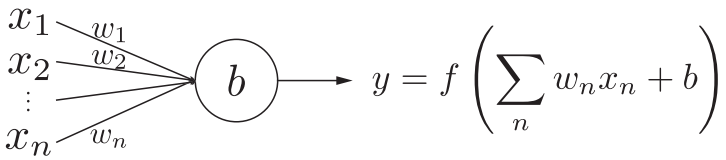



図 7.3: ニューラルネットワークのユニットの構造

ネットワークの各ユニット多次元の入力 $\mathbf{x} = (x_1, ..., x_n)$ を受け取り $y$ を出力する．ユニットが受け取る総入力は各入力に対する結合の重み $w$ を乗じたものとバイアス (bias) と呼ばれる値 $b$ の総和である．ユニットの出力 $y$ は総入力に対する活性化関数 (activation function) と呼ばれる関数 $f$ の出力である．ユニットの処理を式 (7.20) に示す．

$$y = f\left(\sum_{n=1}^{N} w_n x_n + b\right) \tag{7.20}$$

常に 1 の値をとるダミー変数 $x_0$ を導入して，$\mathbf{x} = x_0, ..., x_n$, $\mathbf{w} = w_0, ..., w_n$ とすれば，式 (7.20) は

$$y = f\left(\sum_{n=0}^{N} w_n x_n\right) = f(\mathbf{w}^{\mathrm{T}}\mathbf{x}) \tag{7.21}$$

と表すことができる．ここで $w_0 = b$ である．以下ではこの表現を用いる．$\mathbf{w}^{\mathrm{T}}\mathbf{x}$ は重みベクトルと入力ベクトルの内積である．

### ディープニューラルネットワークの構造

ディープニューラルネットワーク（DNN）はユニットが複数の層に分かれており，ある層のユニットはその層の出力側のユニットと結合している．それぞれの結合には結合の重みが割り当てられる．DNN の構造を図 7.4 に示す．

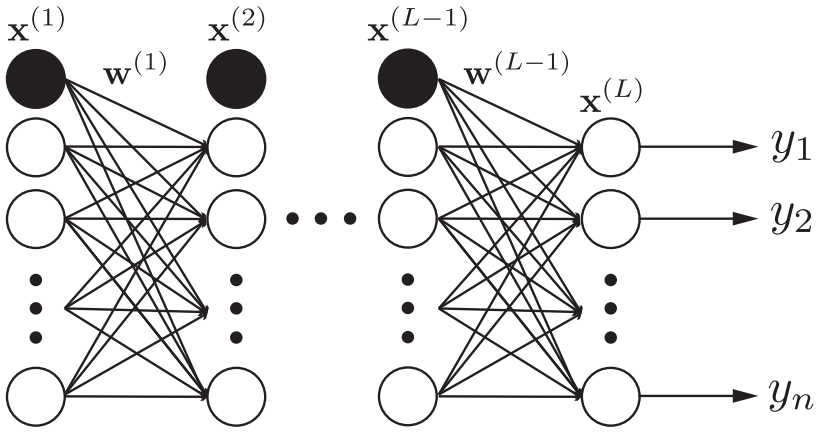



図 7.4: ディープニューラルネットワークの構造

入力層 (input layer) の信号は複数の中間層 (intermediate layer) を経て最終段の出力層 (output layer) へと一方向に伝搬されながら，式 (7.21) に従い変換されてゆく．なお，入力層のユニットは，通常，入力信号をそのまま出力する．ネットワークの中の結合の重みをすべてまとめて $\mathbf{W}$ と記述すると，DNN の出力は $\mathbf{y} = g(\mathbf{x}, \mathbf{W})$ と表され，入力ベクトルと結合の重みにより定まる．

### 確率的勾配降下法

DNN の教師あり学習の方法を，DNN で一般的な学習手法である確率的勾配降下法 (stochastic gradient descent) の基本的なアルゴリズムと共に説明する．DNN をはじめとする順伝搬型ネットワークの学習は，与えられた訓練データ

$$D = \{(\mathbf{x_1}, \mathbf{d_1}), \ldots, \mathbf{x_N}, \mathbf{d_N})\}$$

を元に計算される誤差関数 $E(\mathbf{w})$ のネットワークのパラメータ (重みとバイアス)$\mathbf{w}$ についての最小化である．ここで $\mathbf{x}, \mathbf{d}$ はそれぞれ入力，教師信号である．

多クラス分類問題における誤差関数は

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} d_{nk} \log y_k(\mathbf{x}_n; \mathbf{w}) \tag{7.22}$$

である．$\mathbf{d}_n$ は正解クラスに対応する成分のみ 1 の値をとり，他の成分の値が 0 の $K$ 次元ベクトルである．$K$ はクラス数と等しい．学習の目標は，$E(\mathbf{w})$ に対し最小値を与える $\mathbf{w} = \arg\min_{\mathbf{w}} E(\mathbf{w})$ の算出である．誤差関数は一般に凸関数ではなく，大域的な最小解を求めるのは通常，不可能である．そこで，誤差関数の局所的な極小点 $\mathbf{w}$ を求めることを考える．極小解は初期値から $\mathbf{w}$ を繰り返し更新する反復計算によって求める．反復計算により極小解を求める手法に勾配降下法がある．勾配降下法における勾配 (gradient) は

$$\Delta E = \frac{\partial E}{\partial \mathbf{w}} = \left[ \frac{\partial E}{\partial w_1} \cdots \frac{\partial E}{\partial w_M} \right]^{\mathrm{T}} \tag{7.23}$$

というベクトルである．勾配降下法では式 (7.23) の勾配を用いて現在の重み $\mathbf{w}^{(t)}$ を $\mathbf{w}^{(t+1)}$ に更新する．更新式を式 (7.24) に示す．

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \epsilon \Delta E \tag{7.24}$$

ここで，$\epsilon$ は $\mathbf{w}$ の更新量の大きさを定める定数で，学習率 (learning rate) と呼ばれる．勾配降下法は全訓練サンプル $n = 1, \ldots, N$ に対して計算される誤差関数を最小化する手法である．これに対しサンプルの一部を用いてパラーメータの更新を行う手法を確率的勾配降下法と呼ぶ．確率的勾配降下法には勾配降下法と比較して，計算効率の向上，局所的な極小解にトラップされるリスクの低減という長所がある．

### 誤差逆伝搬法

勾配降下法の実行には誤差関数 $E(\mathbf{w})$ の勾配 $\Delta E$ の計算を要するが，多層のニューラルネットワークの中間層における勾配の計算は困難であった．この問題を非常に簡単に定式化できる変化則で解決したのが誤差逆伝搬法（backpropagation）[Rumelhart 86] と呼ばれる学習則である．ここでは，誤差逆伝搬法について説明する．ネットワークの層数を $L$, 第 $l$ 層の第 $k$ ユニットから第 $l+1$ 層の第 $j$ ユニットへの結合の重みを $w_{k,j}^{(l+1)}$ とするときの，確率的勾配降下法による結合の重みの更新式を式 (7.25) に示す．

$$w_{k,j}^{(l+1)} = w_{k,j}^{(l+1)} - \epsilon \frac{\partial E(\mathbf{W})}{\partial w_{k,j}^{(l+1)}} \tag{7.25}$$

例えば，出力層での誤差関数 $E(\mathbf{W})$ が二乗誤差 $\frac{1}{2}\sum_j \left(d_j - y_j^{(L)}\right)^2$ であるとき，具体的な更新式は

$$w_{k,j}^{(l+1)} = w_{k,j}^{(l+1)} - \epsilon \delta_j^{(l+1)} y_k^{(l)} \tag{7.26}$$

と記述される．ここで $y_k(l)$ は第 $l$ 層の第 $k$ ユニットの出力である．式 (7.26) における $\delta_j^{(l+1)}$ は，合成関数の微分法に従い，出力層での二乗誤差 $\frac{1}{2}\sum_j \left(d_j - y_j^{(L)}\right)^2$ から求まる $\delta_j^{(L)}$ から始まり，以下のように再帰的に計算できる．

$$\begin{aligned} \delta_j^{(L)} &= -(d_j - y_j^{(L)}) y_j^{(L)} (1 - y_j^{(L)}) \\ \delta_j^{(l)} &= -\left\{ \sum_{k=1}^{K^{(l+1)}} \delta_j^{(l+1)} w_{k,j}^{(l+1)} \right\} y_j^{(l)} (1 - y_j^{(l)}) \end{aligned} \tag{7.27}$$

ここで，$K^{(l+1)}$ は，第 $l+1$ 層のユニット数である．式 (7.27) に示す計算過程が，出力層における誤差を一つ前の層に伝搬させていくことから誤差逆伝搬法と呼ばれている．

## 7.2 対象問題

複数の環境における敵対的最大エントロピー逆強化学習問題 (Adversarial Inverse Reinforcement Learning for Multiple Dynamics）を定式化する．Adversarial Inverse Reinforcement Learning for Multiple Dynamics（AIRL-MD）は，状態遷移確率が異なる環境 $E_m = \langle \mathcal{S}, \mathcal{A}, T_m, \gamma \rangle$ と，エキスパートが各環境 $E_m$ で生成したデータセット $D_m$ の集合 $\{(E_m, D_m)\}_{m=1}^{M}$ を所与として，報酬の $R$ を推定する問題である．AIRL-MD の目的関数を以下に示す．

$$\begin{aligned}\underset{\pi_1,\ldots,\pi_M}{\text{minimize}} \max_{Disc_1,\ldots,Disc_M} \frac{1}{M}\sum_{m=1}^{M} & \mathbb{E}_{\pi_m,E_m}[\log(1-Disc_m(s,a,s'))] \\ & +\mathbb{E}_{D_m}[\log(Disc_m(s,a,s'))] - \lambda H(\pi_m)\end{aligned} \tag{7.28}$$

ここで $\pi_m$ は環境 $E_m$ における方策を，$Disc_m$ は環境 $E_m$ における識別器を示す．各識別器は，共通の報酬関数を持ち，異なる価値関数を持つ．

## 7.3 アプローチ

AIRL-MD を解くアプローチについて説明する．敵対的学習は計算効率に優れている一方で，学習が不安定であることが知られている．これは，敵対的学習においては複数のモデル（逆強化学習における方策と識別器）が均衡を保ちながら学習をすることが困難であるからである．例えば，AIRL の方策と識別器が均衡を保つことができず，識別器の分類精度が高くなり過ぎると方策が探索時に獲得する報酬が一様に低くなるため，方策の改善が望めなくなる．式 (7.28) に示す定式化では，環境数の増加に伴って方策や識別器の数が増加し，学習が困難となることが予想される．

そこで本研究では，各方策や識別器に入力する特徴量を変換することによって，複数の方策と識別器をそれぞれ統一して学習可能にする．特徴量の変換の目的は各環境の違いを緩和することである．いま，各環境 $E_m$ における目標状態 $s_{g,m}$ の特徴量を $\phi(s_{g,m})$ と表す．提案するヒューリスティクスは特徴量の変換を用いて，各環境 $E_m$ の目標状態の特徴量 $\phi(s_{g,m})$ を一致させる．特徴量変換 $\phi'$ が満たす制約を以下に示す．

$$\phi'(s_{g,1}) = \phi'(s_{g,2}) = \cdots = \phi'(s_{g,M}) \tag{7.29}$$

対象問題によっては，各環境におけるゴール $s_{g,m}$ が自明でない問題も存在する．このような場合には，エキスパートの各軌跡の終端状態がゴールの確率分布 $P(S_{g,m})$ に従うと仮定

し，エキスパートの各軌跡の終端状態のヒストグラムに基づく最頻値の算出や，ガウス分布を用いた最尤推定から $\phi(s_{g,m})$ を推定する必要がある．

特徴量変換の具体例として，プラントのバルブを制御し，プラント内部の温度を目標値まで昇温する問題を考える．環境間の違いはプラントに供給される流体の温度で，目標値も供給温度の条件に比例して変化する．具体的には，環境 $E_1$ では，供給温度 $505[^\circ C]$ に対して目標温度が $490[^\circ C]$ ，環境 $E_2$ では供給温度 $460[^\circ C]$ に対して目標温度が $445[^\circ C]$ であるとする．この時，提案するヒューリスティクスは，供給温度 $x$ を用いて温度 $y$ を特徴量 $\phi(x,y) = \frac{y+15}{x}$ に変換する．変換後の特徴 $\phi(x,y) = \frac{y+15}{x}$ は，環境 $E_1$ と環境 $E_2$ が目標状態に到達した時に 1 の値を取り，各環境の違いを緩和していることが確認できる．

複数の方策と識別器をそれぞれ統一した目的関数を以下に示す．

$$\underset{\pi}{\text{minimize}} \max_{Disc} \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{\pi,E_m}[\log(1 - Disc(s,a,s'))] + \mathbb{E}_{D_m}[\log(Disc(s,a,s'))] - \lambda H(\pi) \tag{7.30}$$

式 (7.30) は，式 (7.28) の方策，識別器を統一した式で，複数環境全てに適用可能な報酬や方策を学習する問題である．

提案法のアルゴリズムをアルゴリズム 8 に示す．

**Algorithm 8** Adversarial Inverse Reinforcement Learning for Multiple Dynamics

**INPUT:** Environments $\{E_m\}_{m=1}^M$, Demonstrations $\{D_m\}_{m=1}^M$, Prior $P(R)$, Step Size $\delta$, Mini-batch Size $N$

**OUTPUT:** Reward function $r_\theta$

1: Initialize policy $\pi$ and $Disc_{\theta,\phi}$
2: **for** step $n$ in $\{1 \dots N\}$ **do**:wherejh N
3: Collect trajexctories $\tau_i = (s_0, a_0, \dots, s_T, a_T)$ by executing $\pi$ in each environment $E_m$.
4: Train $Disc_{\theta,\phi}$ via binary logistic regression to classify expert data $\tau_{\text{exp},m,i}$ from samples $\tau_{m,i}$.
5: Update policy $\pi$ with respect to $r_\theta$ via policy optimization method.
6: **end for**

## 7.4 実験

本章では，連続状態行動空間のプラント制御問題を対象に実験を行い提案法の有用性を示す．実験は三つある．一つは，対象とする環境に対する逆強化学習の適用可能性を検証する実験で，単一の供給条件の下での逆強化学習の性能を示す．二つ目は，提案法 AIRL-MD の性能を検証する実験で，供給条件が異なる複数の環境のエキスパートの方策を AIRL-MD を用いて学習可能であること確認する．三つ目は，AIRL-MD の転移性能を検証する実験で，AIRL で獲得した報酬と方策を転移した結果と AIRL-MD を比較し，AIRL-MD の性能を検証する．

### 実験設定

はじめに実験対象となるプラントについて説明し，提案法である AIRL-MD の設定について説明する．本実験で対象とするプラントのプロセスは，蒸気タービンに繋がる大気状態の配管を高圧高温蒸気を用いて昇温・昇圧するプロセスである．操作プロセスの概要を図 7.5 に示す．

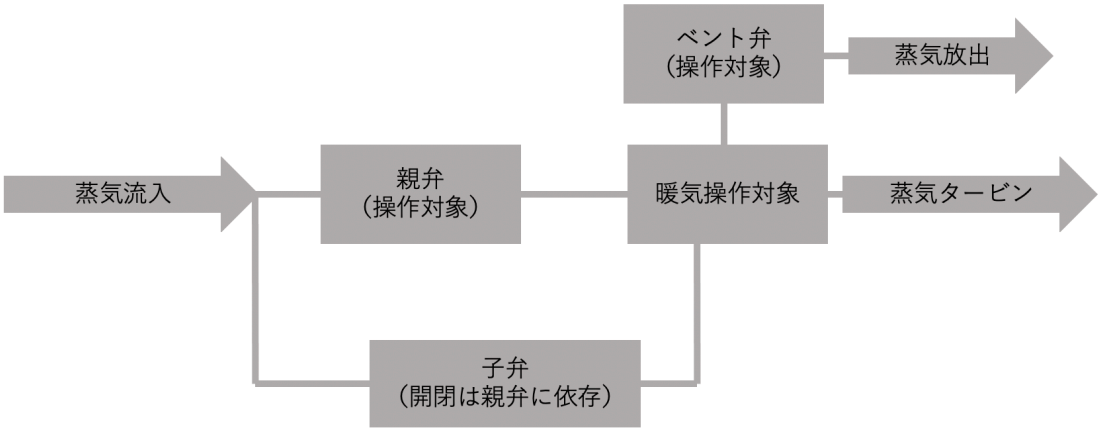



図 7.5: プラント制御問題の概要

操作対象は二つで，暖気操作対象への蒸気の流入量を制御する親弁と，暖気操作対象から大気への流出量を制御するベント弁である．子弁は親弁と同じく暖気操作対象への蒸気の流入量を制御し，子弁の開閉は，子弁に依存して自動的に制御される．

目標状態は，供給条件（供給される蒸気の温度と圧力）に依存する．具体的には，目標温度は供給温度 $-15[^{\circ}C]$ で，目標圧力は供給圧力 $-5[kg/cm^2]$ の圧力である．エキスパートデータは供給条件が異なる四つの環境から取得する．エキスパートデータを取得した供給条件を表 7.1 に示す．

表 7.1: プラント制御問題の供給条件

| 供給条件 | 供給圧力 $[kg/cm^2]$ | 供給温度 $[^{\circ}C]$ |
|---|---|---|
| 1 | 温度 1 | 圧力 1 |
| 2 | 温度 1 | 圧力 2 |
| 3 | 温度 2 | 圧力 1 |
| 4 | 温度 2 | 圧力 2 |

各供給条件におけるエキスパート数は 100 で，プラントシミュレータでルールベースを用いて作成した．

供給条件 1 におけるエキスパートの操作軌跡を以下に示す．

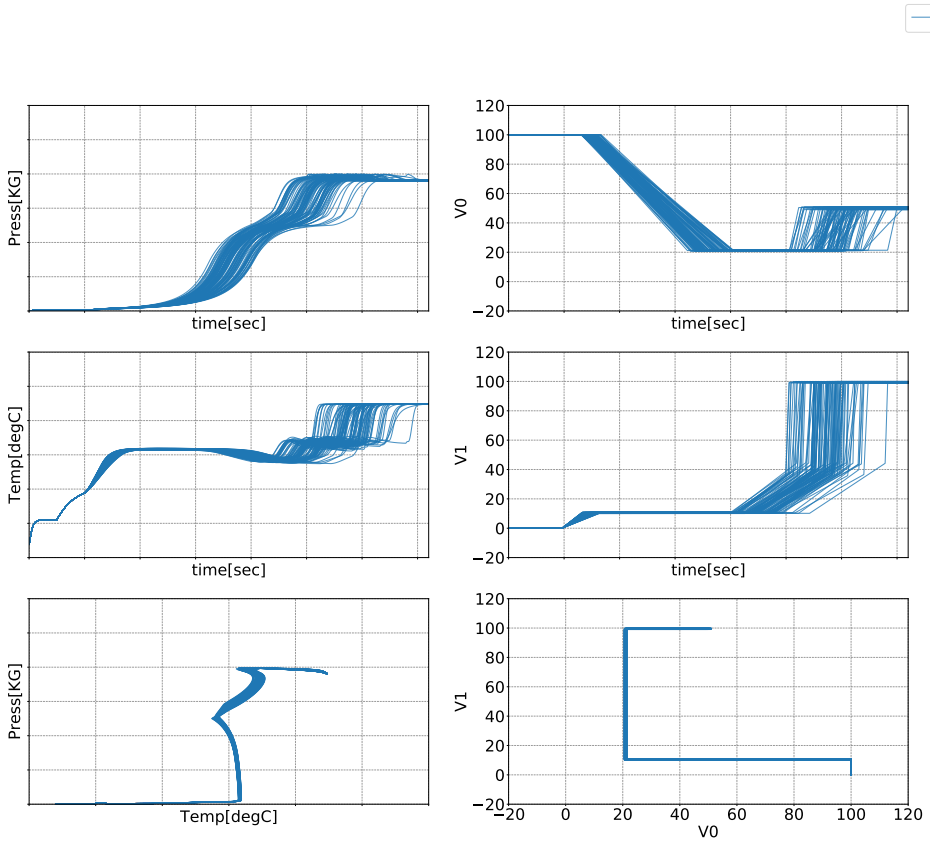



図 7.6: プラント制御問題におけるエキスパートの軌跡

図 7.6 の左側，圧力と温度の時間変化を表した図で，右側はベント弁（V0）と親弁（V1）の制御を表している．エキスパートは目標状態を実現するだけでなく，暖気制御対象の内部を飽和状態や，昇圧速度を一定値以下に保つ必要がある．これらの制約を満たす方策を学習可能な報酬の設計は困難であり，逆強化学習の適用が必要となる．

本実験の状態入力を表 7.3 に示す．

表 7.2: プラント制御問題の状態入力

| | |
|---|---|
| 1 | 圧力 |
| 2 | 温度 |
| 3 | 圧力変化量 |
| 4 | 温度変化量 |
| 5 | 親弁開度 |
| 6 | ベント弁開度 |
| 7 | 子弁開度 |
| 8 | 親弁変化量 |
| 9 | ベント弁変化量 |
| 10 | 蒸気流量 |
| 11 | 飽和曲線との乖離度 |
| 12 | 昇圧速度違反のフラグ |

状態入力の全ての値は，$[0, 1]$ の値をとるよう正規化し，目標状態を定める温度と圧力は 7.3 で述べた方法で正規化を行う．行動出力は，ベント弁（V0）と親弁（V1）の変化量である．

次に，提案法 AIRL と AIRL-MD の設定について述べる．学習ステップ数は 1000 万ステップで 1 ステップは 2 秒間である．1 エピソードは 3600 ステップで，7200 秒に相当する．学習結果は 5 万ステップ毎に保存する．保存した学習済みモデルのうち目標状態に到達可能で，制約違反がないモデルを成功モデルと呼ぶ．

AIRL と AIRL-MD の内部の強化学習には Proximal Policy Optimization（PPO）を用いる．PPO は方策に加えて状態価値関数を使用する強化学習手法であるため，AIRL と AIRL-MD は PPO の方策と状態価値関数，識別器の報酬と状態価値関数（識別器）の四つのニューラルネットワークを持つ．PPO のニューラルネットワークの活性化関数は Relu で，中間層の数は二つ，中間層のユニット数は 64 である．識別器のニューラルネットワークの活性化関数は tanh で，中間層の数は三つ，中間層のユニット数は 32 である．

AIRL のハイパーパラメータを以下に示す．

表 7.3: AIRL と AIRL-MD のハイパーパラメータの設定

| パラメータ名 | 値 |
|---|---|
| 学習率（PPO） | 0.0001 |
| 学習率（識別器） | 0.0001 |
| パラメータ更新頻度 | 16384 ステップ |
| バッチサイズ | 2048 |
| 割引率 | 0.995 |

各実験において，ランダムシードを変更した実験を 10 回行った．

### 実験結果

#### 単一の供給条件の下での逆強化学習

ここでは，ある供給条件の環境に対する AIRL の適用可能性を検証する．以下に各供給条件に対する実験の結果を示す．

表 7.4: プラント制御問題に対する AIRL の実験結果

| 供給温度 [°C] | 供給圧力 [kg/cm$^2$] | 成功実験数/全実験数 | 成功モデル数/全保存数 |
|---|---|---|---|
| 温度 1 | 圧力 1 | 1/10 | 63/2000 |
| 温度 1 | 圧力 2 | 2/10 | 62/2000 |
| 温度 2 | 圧力 1 | 3/10 | 191/2000 |
| 温度 2 | 圧力 2 | 3/10 | 94/2000 |

表 7.4 の最左列はプラントに供給される蒸気の供給温度を，左から 2 列目は供給圧力を示す．左から 3 列目は成功モデルが得られた実験数と全実験数を示す．3 列目の値から，各供給条件において，成功条件を満たす方策を学習できていることが確認できる．AIRL が各供給条件において学習した軌跡を以下に示す．

逆強化学習の学習過程を図 7.8 に示す．横軸は学習ステップ数，縦軸は各指標の値を，プロット中の黒い点は成功モデルが得られたことを示している．

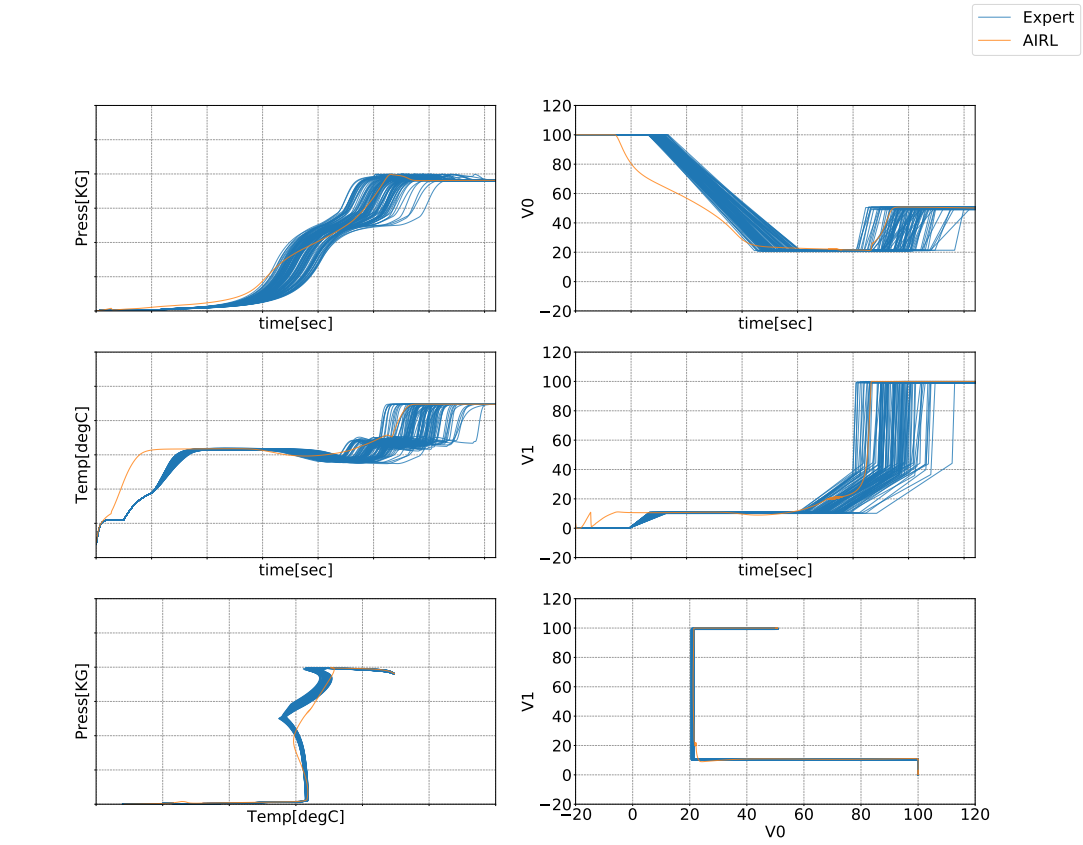



図 7.7: プラント制御問題の学習例（供給温度 1，供給圧力 1）

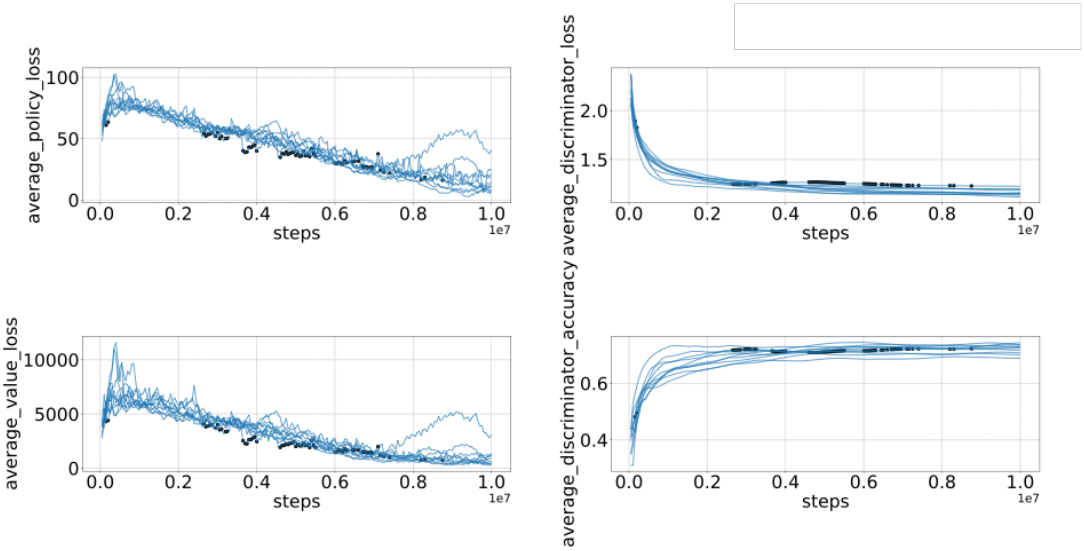



図 7.8: プラント制御問題の学習経過（供給温度 1，供給圧力 1）

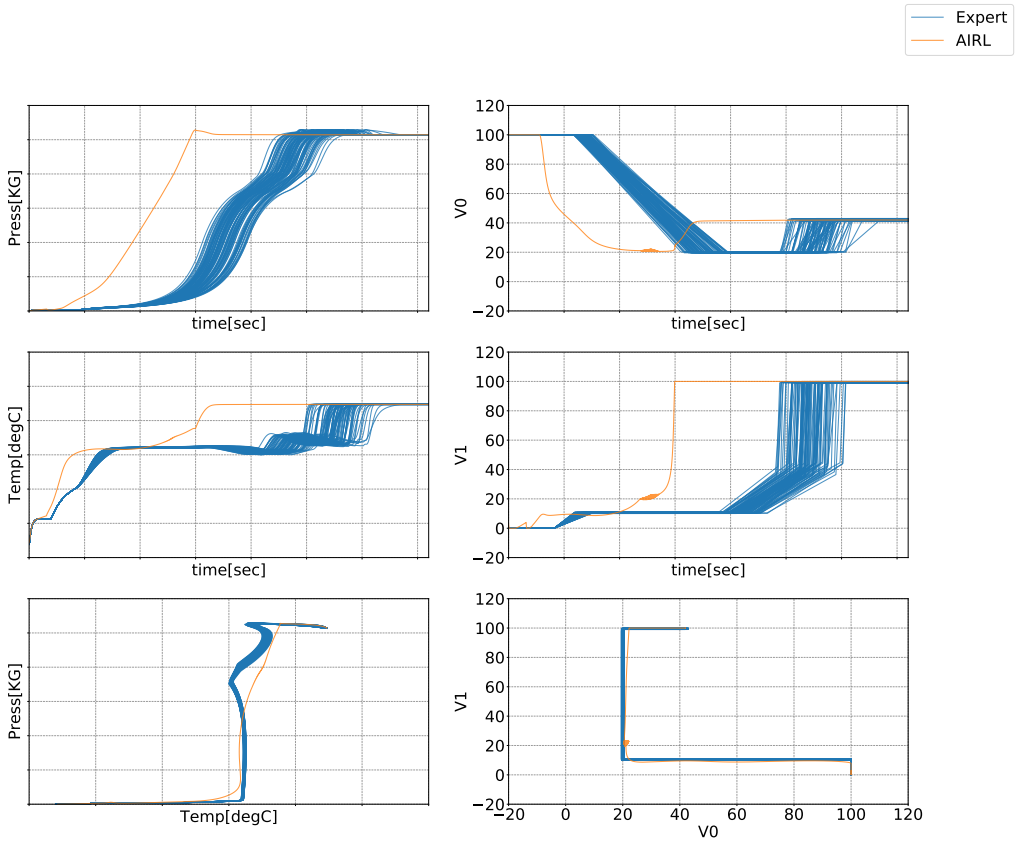



図 7.9: プラント制御問題の学習例（供給温度 1，供給圧力 2）

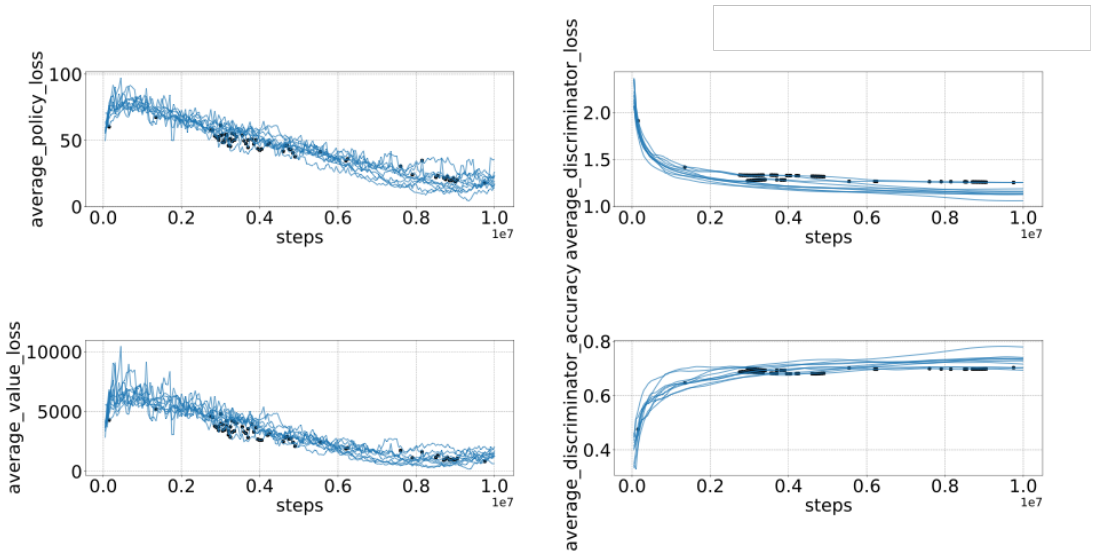



図 7.10: プラント制御問題の学習経過（供給温度 1，供給圧力 2）

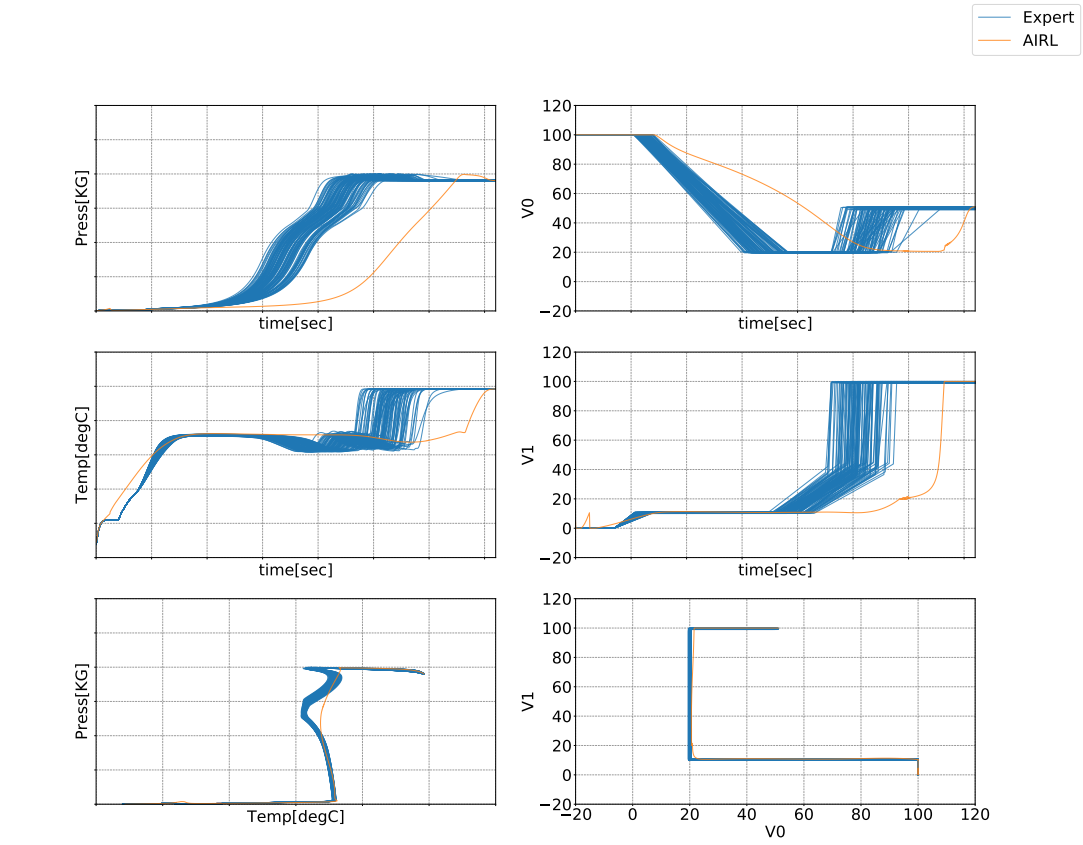



図 7.11: プラント制御問題の学習例（供給温度 2，供給圧力 1）

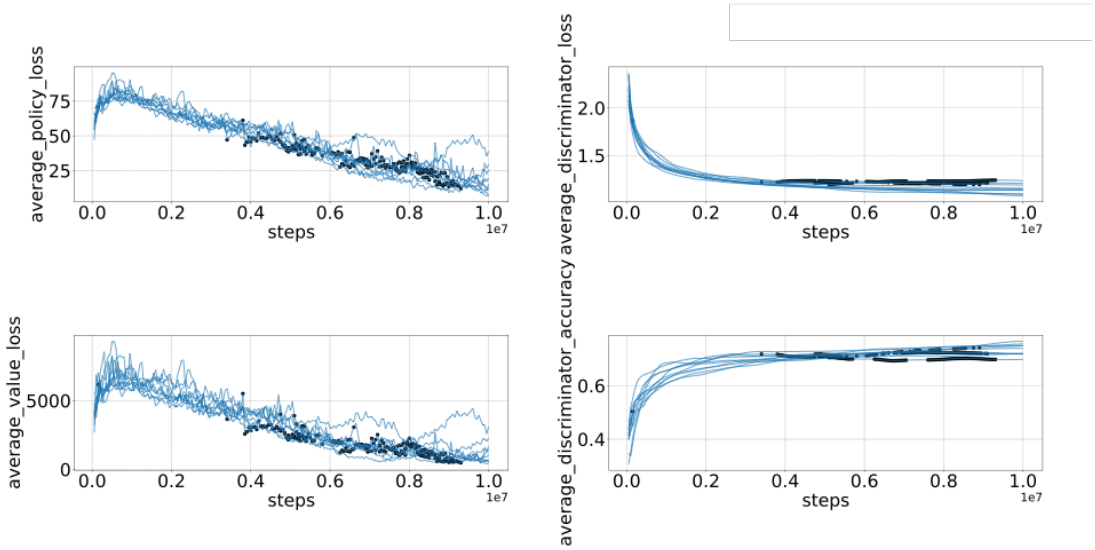



図 7.12: プラント制御問題の学習経過（供給温度 2，供給圧力 1）

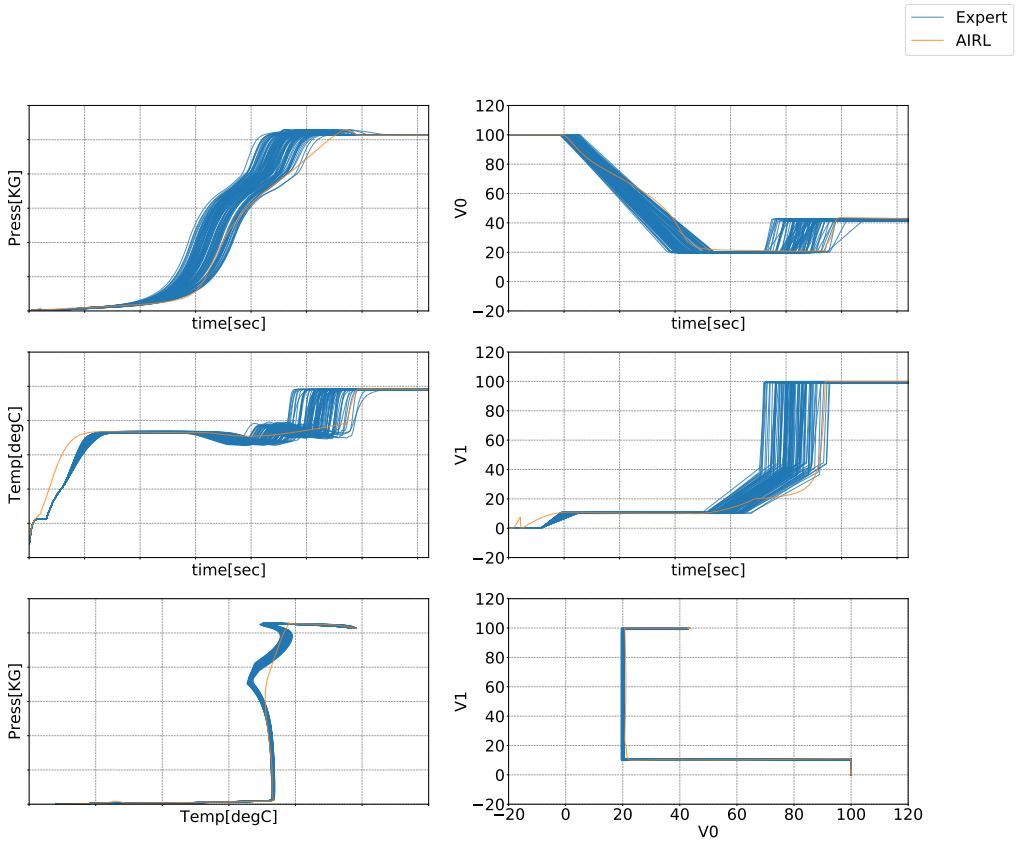



図 7.13: プラント制御問題の学習例（供給温度 2，供給圧力 2）

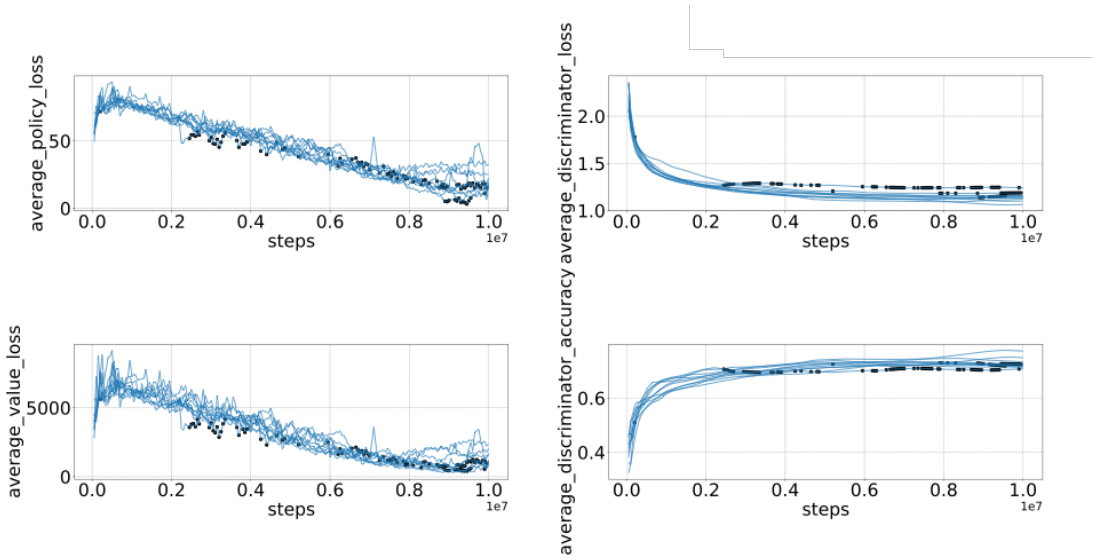



図 7.14: プラント制御問題の学習経過（供給温度 2，供給圧力 1）

### 複数の供給条件の下での逆強化学習

ここでは，ある供給条件の環境に対する AIRL-MD の適用可能性を検証する．以下に各供給条件に対する実験の結果を示す．

表 7.5: 複数環境における AIRL-MD の実験結果．全 4 環境で目標状態に到達し，制約を満たしたモデルを成功モデルと定義．

| 供給温度 [°C] | 供給圧力 [kg/cm$^2$] | 成功実験数/全実験数 | 成功モデル数/全保存数 |
|---|---|---|---|
| 温度 1, 2 | 圧力 1, 2 | 2/10 | 52/2000 |

表 7.6 の最左列はプラントに供給される蒸気の供給温度を，左から 2 列目は供給圧力を示す．左から 3 列目は成功モデルが得られた実験数と全実験数を示す．3 列目の値から，各供給条件において，成功条件を満たす方策を学習できていることが確認できる．また，提案法の実験の成功率は複数環境において学習しているのにも関わらず既存手法の AIRL と同等であり，提案法が与える学習の不安定性への影響が大きくないことを示している．

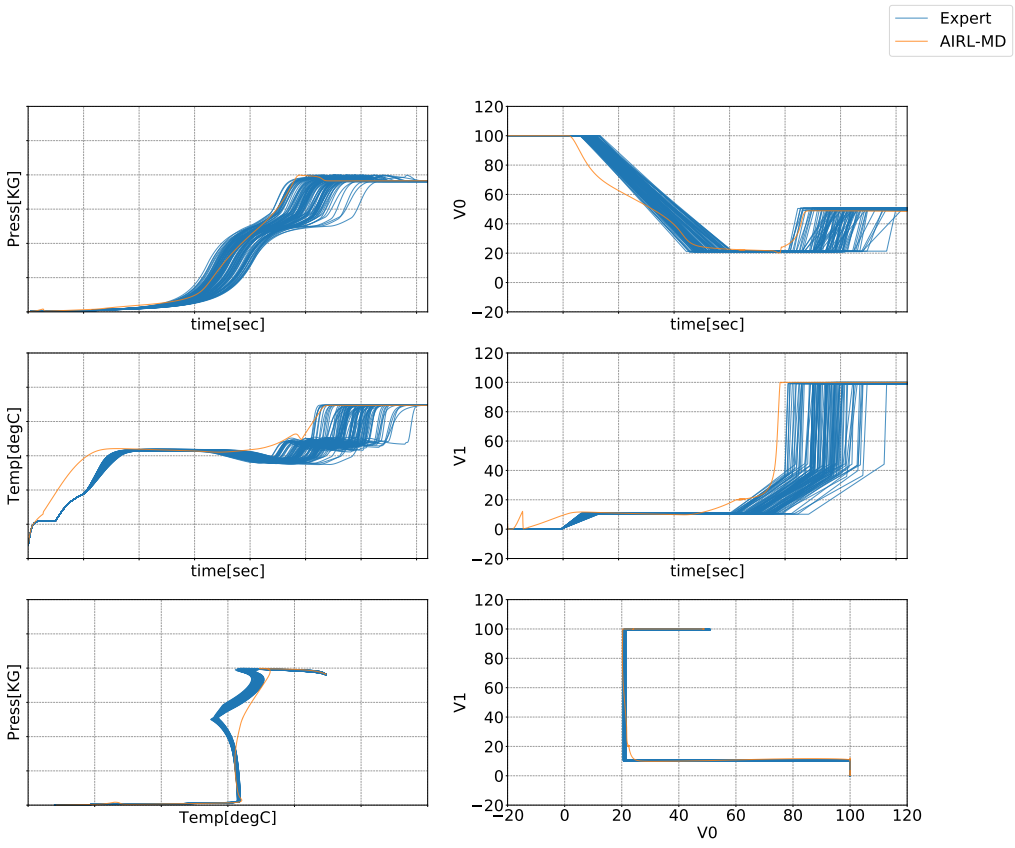



図 7.15: プラント制御問題の学習例（供給温度 1，供給圧力 1）

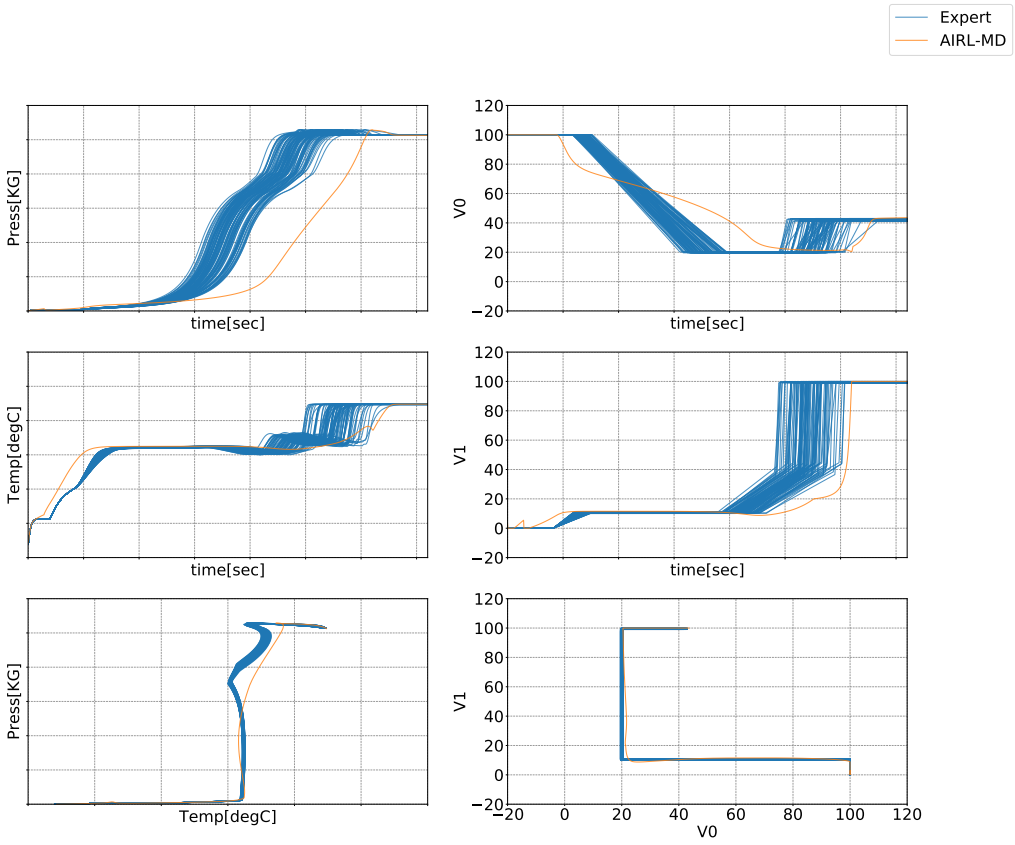



図 7.16: プラント制御問題の学習例（供給温度 1，供給圧力 2）

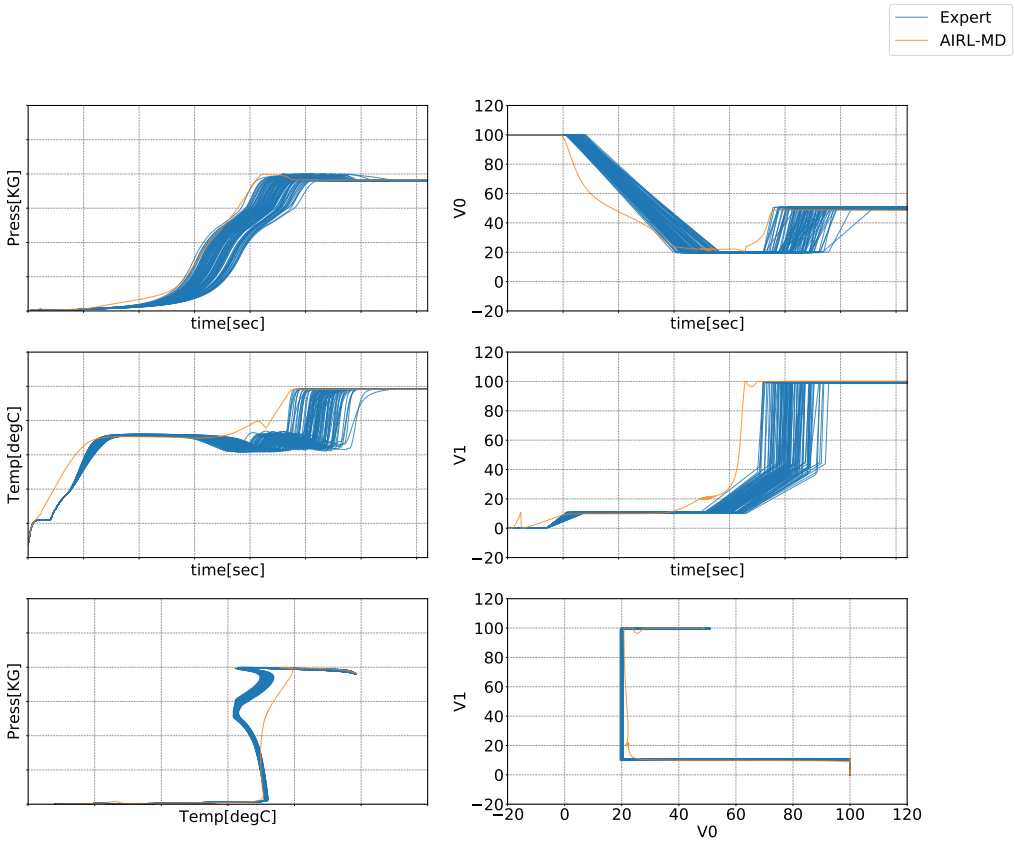



図 7.17: プラント制御問題の学習例（供給温度 2，供給圧力 1）

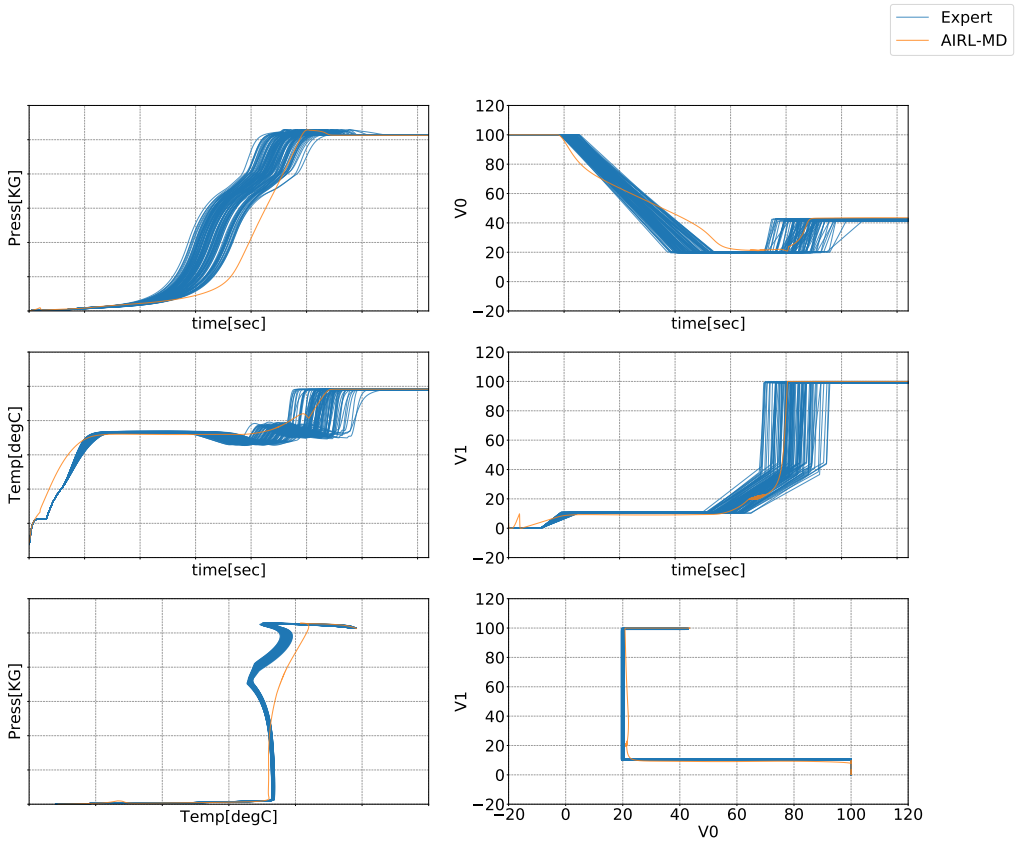



図 7.18: プラント制御問題の学習例（供給温度 2，供給圧力 2）

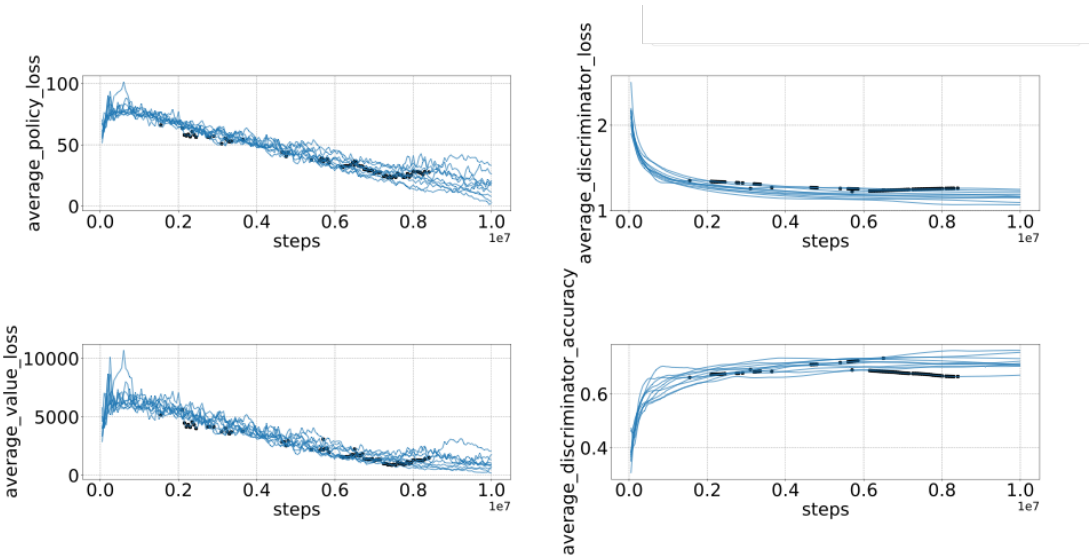



図 7.19: プラント制御問題の学習経過（複数環境における学習）

### 学習結果の転移

提案法を用いて学習した前節の学習結果を，供給温度 1∼ 温度を等間隔で 10 分割した 10 条件．供給圧力が 80 ∼ 106[kg/cm$^2$] を 2[kg/cm$^2$] 刻みで 14 条件の計 140 条件の環境に転移し，昇温昇圧操作を確認した．比較対象は，7.4 の学習結果で，複数環境の情報を学習済み方策に反映するため，複数環境の報酬の平均値に対して供給温度 490°$C$，供給圧力 92[kg/cm$^2$] で方策を 30 万ステップ追加で訓練した方策である．

表 7.6: 学習済みの方策の転移結果．

| 供給温度 [°C] | 供給圧力 [kg/cm$^2$] | 成功条件数/全条件数 |
|---|---|---|
| 温度 1∼ 温度 2，10 分割 | 圧力 1∼ 圧力 2，14 分割 | 70/140 |
| 温度 1∼ 温度 2，10 分割 | 圧力 1∼ 圧力 2，14 分割 | 140/140 |

表 7.6 の結果は，既存手法の転移の成功数が 140 条件のうち 70 条件であり，提案法の成功数が 140 条件のうち 140 条件であることが確認できる．したがって，提案法の転移の成功確率が既存手法より優れていることが確認できる．以下に，既存手法と提案法の方策を転移し獲得された軌跡を示す．オレンジ色の線は失敗した供給条件の下での軌跡を，青色の線は成功した供給条件の下での軌跡を示す．

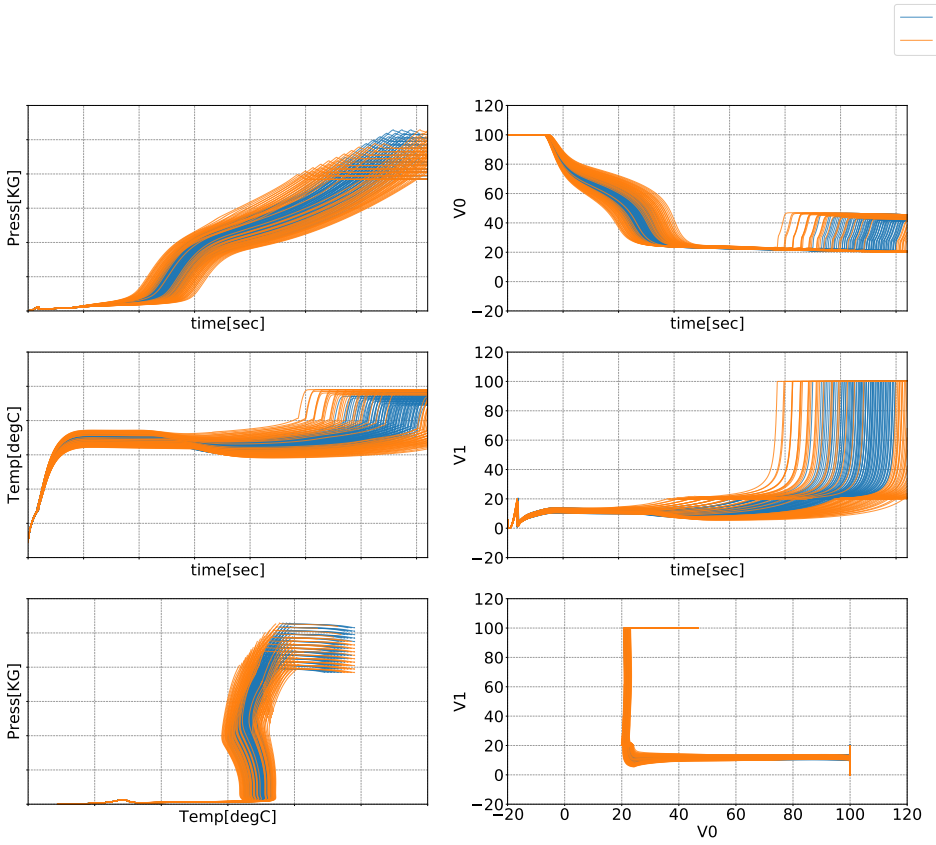



図 7.20: AIRL の転移結果（4 環境の推定報酬の平均値による追加学習）

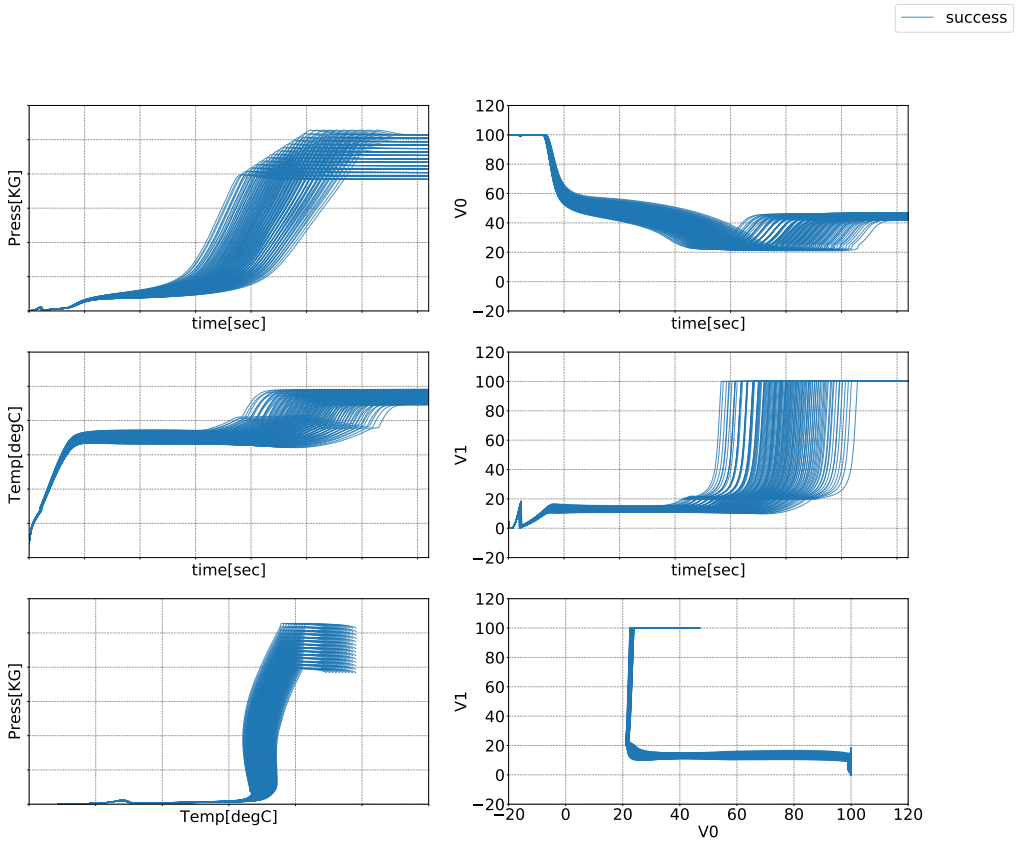



図 7.21: AIRL-MD の転移結果

## 7.5 考察

提案法 AIRL-MD は，対象とする環境が複数になることによって生じる学習の不安定性の増大を抑制するために，特徴量変換を導入し，方策などのモデルの数を一つに削減する．ここでは，特徴量変換の適用範囲について議論する．この特徴量変換は，各環境のゴールとなる状態の環境間の差異を緩和するアプローチであるが，各環境におけるゴールが定義困難な問題も存在する．例えば，エキスパートが複数の状態を循環するような問題や，エキスパートの軌跡のしゅたん状態の分布が複数のモードを持つ問題においては，ゴールとなり得る状態が複数存在する．この問題のエキスパート軌跡に対してヒストグラムによる分析や，ガウス分布による推定でゴールを算出し，特徴量変換を行ったとしても，エキスパートと同様の報酬や方策を学習することは困難であると考えられる．このように複数のゴールを持つ問題に対しては，エキスパートデータを単一のモードを持つデータセット群に分割し，各問題に対して提案法を適用するなどの対策が考えられる．

次に，環境の数 $M$ に対する提案法のスケール性について議論する．提案法が複数の環境において安定的に学習するためには，複数の環境のデータに対して同時に学習を行う必要が

ある．したがって，環境の数 $M$ の増加に従い，学習時間が長くなってしまうという課題がある．この課題に対しては，学習中の方策を用いて環境で軌跡を生成する処理の並列化や，非同期的な学習方法を導入することによって学習時間を改善できると考えられる．

## 7.6 まとめ

本章では，複数環境の軌跡から報酬を推定する敵対的最大エントロピー逆強化学習手法を提案した．提案法は，敵対的学習を用いた最大エントロピー逆強化学習を複数環境に基づいており，連続状態行動空間に適用可能な手法である．実験では連続状態行動空間のプラント制御問題を対象として既存法と提案法を比較した．実験の結果，複数環境において逆強化学習を行う提案法が既存法に比べて多くの条件に転移可能な方策を学習できることを確認した．また，複数の環境において敵対的学習を行っても，敵対的学習における重要な課題である学習の不安定性が既存法に比べて悪化しないことを確認した．

# 第8章　結論

本論文では逆強化学習の推定報酬，方策の改善を目的として，状態遷移確率が異なる複数のマルコフ決定過程のエキスパートデータを対象とする逆強化学習問題を定式化し，解法を示した．具体的には定式化した三つの問題，線形計画問題，ベイズ推定，敵対的学習問題に対し，四つの解法，線形計画逆強化学習，ベイジアン逆強化学習，ミニバッチベイジアン逆強化学習，敵対的最大エントロピー逆強化学習を提案した．以下で，各提案法について整理し，各手法の住み分けと今後の課題について述べる．

### 複数のマルコフ決定過程における線形逆強化学習

複数のマルコフ決定過程における線形逆強化学習は，複数の環境におけるエキスパートの方策から報酬を推定する方法である．この方法の利点は，複数の環境におけるエキスパート方策が最適となる報酬の推定を保証できる点にある．

実験では，ベンチマーク問題である風向きが異なる複数の Windy grid world 環境に対して提案法を適用し，複数の環境におけるエキスパートの方策が最適となる報酬を推定できることを確認した．また，推定報酬を学習時と異なる環境に転移し，転移先の環境におけるエキスパート方策を学習する実験においては，複数の環境のエキスパート方策を扱う提案法のエキスパート方策の再現率が，一環境から報酬を推定する既存法と比較して優れていることを確認した．このように複数のマルコフ決定過程における線形計画逆強化学習は複数の環境におけるエキスパート方策が与えられた下で優れた性能を示したが，エキスパートの状態行動系列（軌跡）から報酬を推定することはできないという課題が残った．

### 複数のマルコフ決定過程におけるベイジアン逆強化学習

エキスパートの軌跡を扱えないという線形計画逆強化学習の課題を克服するために，状態遷移確率が異なる複数のマルコフ決定過程においてエキスパートが生成した軌跡から報酬を推定するベイジアン逆強化学習問題を定式化した．また，定式化した問題に対してマルコフ連鎖モンテカルロ法を用いた解法を示した．ベイズ推定の利点の一つは，設計者が持つ報酬に関する事前知識を事前分布として導入できる点にある．適切な事前分布を導入することによって，エキスパートの軌跡の数が少ない場合にも，エキスパートに近い報酬が推定できることが期待される．

実験では，ベンチマーク問題である風向きが異なる複数の Windy grid world 環境に対して提案法を適用し，複数環境におけるエキスパートの軌跡から報酬を推定できることを確認した．また，提案法が，単一環境から報酬を推定する既存法と比較して，エキスパート方策の再現率が高いことを Expected Value Difference[Levine 11] を用いて確認した．実験から得られた重要な知見としては，単一環境と複数環境におけるエキスパートの軌跡の合計数が同じ場合（提案法においては各環境におけるエキスパートの軌跡の数が少ない）でも複数環境におけるエキスパート軌跡から報酬を推定することが有用性が示されたことである．これはエキスパートの方策を所与とする線形計画逆強化学習では確認することができなかった知見である．このように，複数環境におけるベイジアン逆強化学習のエキスパートの軌跡数に対する優れた性能が確認されたものの，提案法の計算量がエキスパートの軌跡が得られた環境の数に応じて増大するという課題が明らかになった．

### 複数のマルコフ決定過程におけるミニバッチベイジアン逆強化学習

提案したベイジアン逆強化学習手法の計算量がエキスパートの軌跡が得られた環境の数に応じて増大するという課題に対し，ミニバッチ近似を用いて計算量を改善したミニバッチベイジアン逆強化学習手法を提案した．具体的には，複数環境のベイジアン逆強化学習の各マルコフ連鎖モンテカルロステップにおける動的計画法の回数が環境数と同じなのに対して，ミニバッチベイジアン逆強化学習手法の動的計画法の回数は環境の数よりも小さい任意のミニバッチ数に定めることができる．

Windy grid world 環境を対象とした実験では，ミニバッチベイジアン逆強化学習手法に関する二つのことを確認した．一つ目は，環境数よりも小さいミニバッチ数を設定しても，適切にエキスパートの報酬が推定できることである．エキスパートの軌跡を使用する環境の数を固定し，ミニバッチ数を変化させ，各ミニバッチ数における推定報酬の性能を比較することによって，環境数よりも小さいミニバッチ数を設定しても，適切にエキスパートの報酬が推定できることを確認した．もう一つは，ミニバッチ数を固定した下で，エキスパートの環境数の増加に応じて推定報酬が改善することである．この実験では，ミニバッチ数を 1 に固定し，エキスパートの環境数を増加させる実験を行い，推定報酬がエキスパートの環境数の増加に応じて改善することを確認した．

ミニバッチベイジアン逆強化学習を用いることによって，推定報酬に対する最適方策を求める動的計画法の回数が減少し，計算量が削減できた．しかし，各マルコフ連鎖モンテカルロステップで推定報酬に対する最適方策を求める必要性は依然存在し，最適方策の学習が不安定になる連続状態行動空間の問題を扱うことはできないという課題がある．

### 複数のマルコフ決定過程における敵対的最大エントロピー逆強化学習

敵対的な学習を行う最大エントロピー逆強化学習 [Fu 18] は，推定報酬の更新に推定報酬に対する最適方策を必要とせず，最適方策の学習が不安定になる連続状態行動空間の問題にも適用可能な逆強化学習手法である．本論文では，この Adversarial Inverse Reinforcement Learning[Fu 18] の適用範囲を単一の環境から複数の環境へと拡張した．

実験では，プラントの内部状態を昇温，昇圧する問題を対象として提案法の有用性を示した．具体的には，供給条件が異なる複数のプラントにおけるエキスパートの軌跡に提案法を適用し，複数の供給条件のプラントにおいて適用可能な方策が獲得できることを確認した．また，複数の環境における学習によって，学習の安定性が悪化しないことも確認した．提案法によって学習した方策を，学習時と異なる供給条件の環境へと転移する実験においては，提案法で獲得した方策の昇温，昇圧の成功確率が既存手法で獲得した方策の成功確率よりも高いことを確認した．

### 各定式化の位置づけ

本論文で提案した三つの定式化の違いは大きく二つある．一つ目は対象とするエキスパートデータの違いで，線形計画逆強化学習はエキスパートの方策を対象とし，ベイジアン逆強化学習と敵対的最大エントロピー逆強化学習はエキスパートの軌跡を対象とする．エキスパートの方策が実問題で得られることは少ないが，エキスパートの方策が得られる離散状態行動空間の問題が対象である場合には，推定報酬が制約を満たす保証がある線形計画逆強化学習を用いることが望ましい．二つ目は環境の状態行動空間の違いである．線形計画逆強化学習と，ベイジアン逆強化学習の適用範囲が離散状態行動空間に限定されるのに対し，敵対的最大エントロピー逆強化学習は離散状態行動空間と連続状態行動空間の両方に適用可能である．したがって，対象問題が連続状態行動空間である場合には敵対的最大エントロピー逆強化学習を用いる必要がある．先述の通り敵対的最大エントロピー逆強化学習は，離散状態行動空間の問題にも適用することができる．しかし，離散状態行動空間において得られるエキスパートの軌跡の数が少ない場合には，報酬に関する事前知識を導入できるベイジアン逆強化学習を用いるべきだと考えられる．

### 今後の課題

今後の課題は二つある．一つ目は状態遷移確率だけでなく，状態行動空間が異なる複数の環境を対象とする逆強化学習手法の構築である．状態行動空間の差異を緩和するための方法として，複数の環境における状態行動空間の潜在表現を学習するアプローチが考えられる．もう一つの課題は，複数の環境の組み合わせの最適化である．本論文では，複数の環境のエキスパートのデータがあらかじめ与えられている問題に対する解法を示した．しかし，実問題においてはエキスパートのデータを取得する環境を選べる（例: データを取得するプラン

トを選べる）場合がある．このような問題設定においては，報酬を推定するのに有用な環境の組み合わせの最適化によって，提案法の推定性能を高めることができる可能性がある．

# 付 録A　プラント制御環境の数理モデル

## 流量計算

流量は蒸気配管圧力の差によって決まるものとする．

$$F_{\mathrm{in}}V(t) = \{CV_1 \cdot LIFT_1(t) + CV_2 \cdot LIFT_2(t)\} \cdot \{P_{in} - P(t)\} \tag{A.1}$$

$$F_{\mathrm{out}}V(t) = \{CV_0 \cdot LIFT_0(t)\} \cdot \{P(t) - P_{\mathrm{air}}\} \tag{A.2}$$

表 A.1: 流量計算式の変数の定義

| 変数名 | 説明 |
|---|---|
| $F_{\mathrm{in}}V(t)$ | 流入量（質量流量）[kg/s] |
| $F_{\mathrm{out}}V(t)$ | 流出量）[kg/s] |
| $CV_0$ | バルブ流量係数 [kg/s/Pa](ベント弁) |
| $CV_1$ | バルブ流量係数 [kg/s/Pa](入口親弁) |
| $CV_2$ | バルブ流量係数 [kg/s/Pa](入口子弁) |
| $LIFT_0$ | ベント弁，バルブ開度 [-] |
| $LIFT_1$ | 入口親弁，バルブ開度 [-] |
| $LIFT_2$ | 入口子弁，バルブ開度 [-] |
| $P(t)$ | 蒸気配管圧力 [Pa] |
| $P_{in}$ | 蒸気供給圧力 [Pa] |
| $P_{air}$ | 大気圧力 [Pa] |

## 放熱量計算

放熱量は蒸気配管温度と大気温度の差に比例する．

$$Q(t) = UA \cdot \{T(t) - T_{\mathrm{air}}\} \tag{A.3}$$

表 A.2: 放熱量計算式の変数の定義

| 変数名 | 説明 |
|---|---|
| $Q(t)$ | 放熱量 [J/s] |
| $UT$ | 放熱軽量 [J/s/K] |
| $T(t)$ | 蒸気配管温度 [K] |
| $T_{air}$ | 大気温度 [K] |

### 配管圧力計算

配管内は飽和状態にあると仮定し，蒸発量（凝縮量）を計算する．ただし，ドレンがない場合は，過熱状態にあるものとし，蒸発量はゼロとする．飽和状態は以下の基礎式を連立して求める．

#### ガス（空気，蒸気）ホールドアップ

$$H(t) = H(t-1) + \left\{ \frac{F_{in}V(t)}{18.02} - \frac{F_{out}V(t)}{MW(t)} - \frac{W(t)}{18.02} \right\} \cdot \Delta t \tag{A.4}$$

$$MW(t) = 18.02 \cdot \frac{H_{H2O}(t)}{H(t)} + 28.8 \cdot \left\{ 1 - \frac{H_{H2O}(t)}{H(t)} \right\} \tag{A.5}$$

表 A.3: 配管圧力計算式の変数の定義

| 変数名 | 説明 |
|---|---|
| $H(t)$ | ガスホールドアップ量 [kg-mol] |
| $H_{H2O}(t)$ | ガス（スチーム成分のみ）ホールドアップ量 [kg-mol] |
| $W(t)$ | 凝縮量 [kg/s] |
| $MW(t)$ | ガス分子量 [kg/kg-mol] |

#### ガス（スチーム成分のみ）ホールドアップ

$$H_{H2O}(t) = H_{H2O}(t-1) + \left\{ \frac{F_{in}V(t)}{18.02} - \frac{F_{\mathrm{out}}V(t)}{MW(t-1)} \cdot \frac{H_{H2O}(t-1)}{H(t-1)} - \frac{W(t)}{18.02} \right\} \cdot \Delta t \tag{A.6}$$

#### 飽和水蒸気圧

$$P_{H2O}(t) = A \cdot T(t) + B \tag{A.7}$$

表 A.4: 飽和水蒸気圧計算式の変数の定義

| 変数名 | 説明 |
|---|---|
| $P_{H2O}(t)$ | 飽和水蒸気圧 [Pa] |
| $A$ | 飽和蒸気圧線形近似係数 |
| $B$ | 飽和蒸気温度線形近似係数 |

飽和水蒸気圧とスチーム分圧の関係式

$$\frac{P_{H2O}(t)}{P(t)} = \frac{H_{H2O}(t)}{H(t)} \tag{A.8}$$

ヒートバランス

$$T(t) = T(t-1) + \frac{\Delta t}{MCp} \cdot \left\{ \frac{F_{in}V(t)}{18.02} \cdot Cp \cdot [T_{in} - T(t-1)] + \gamma \cdot W(t) - Q(t) \right\} \tag{A.9}$$

表 A.5: 飽和水蒸気圧計算式の変数の定義

| 変数名 | 説明 |
|---|---|
| $MCp$ | 熱容量 [J/K] |
| $Cp$ | 蒸気比熱 [J/kg -mol K] |
| $\gamma$ | 蒸気凝縮潜熱 [J/kg] |

ドレンホールドアップ

$$HL(t) = HL(t-1) + W(t) \cdot \Delta t \tag{A.10}$$

# 参考文献


[Abbeel 04] Abbeel, P. and Ng, A. Y.: Apprenticeship learning via inverse reinforcement learning, in *Proceedings of the twenty-first international conference on Machine learning*, p. 1ACM (2004)

[Amin 17] Amin, K., Jiang, N., and Singh, S.: Repeated Inverse Reinforcement Learning, in *Advances in Neural Information Processing Systems 30*, pp. 1815–1824, Curran Associates, Inc. (2017)

[Amodei 16] Amodei, D., Olah, C., Steinhardt, J., Christiano, P. F., Schulman, J., and Mané, D.: Concrete Problems in AI Safety (2016)

[Andrychowicz 18] Andrychowicz, M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., et al.: Learning dexterous in-hand manipulation, *arXiv preprint arXiv:1808.00177* (2018)

[Applegate 91] Applegate, D. and Kannan, R.: Sampling and integration of near log-concave functions, in *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, pp. 156–163ACM (1991)

[Babes 11] Babes, M., Marivate, V., Subramanian, K., and Littman, M. L.: Apprenticeship learning about multiple intentions, in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 897–904 (2011)

[Bardenet 17] Bardenet, R., Doucet, A., and Holmes, C.: On Markov chain Monte Carlo methods for tall data, *The Journal of Machine Learning Research*, Vol. 18, No. 1, pp. 1515–1557 (2017)

[Barto 89] Barto, A. G., Sutton, R. S., and Watkins, C.: *Learning and sequential decision making*, University of Massachusetts Amherst, MA (1989)

[Bengio 12] Bengio, Y.: Practical Recommendations for Gradient-Based Training of Deep Architectures, in *Neural Networks: Tricks of The Trade*, pp. 437–478, Springer (2012)

[Choi 11] Choi, J. and Kim, K.-E.: Map inference for bayesian inverse reinforcement learning, in *Advances in Neural Information Processing Systems*, pp. 1989–1997 (2011)

[Choi 12] Choi, J. and Kim, K.-E.: Nonparametric Bayesian inverse reinforcement learning for multiple reward functions, in *Advances in Neural Information Processing Systems*, pp. 305–313 (2012)

[Choi 13] Choi, J. and Kim, K.-E.: Bayesian nonparametric feature construction for inverse reinforcement learning, in *Twenty-Third International Joint Conference on Artificial Intelligence* (2013)

[Clavera 19] Clavera, I., Nagabandi, A., Liu, S., Fearing, R. S., Abbeel, P., Levine, S., and Finn, C.: Learning to Adapt in Dynamic, Real-World Environments through Meta-Reinforcement Learning, in *International Conference on Learning Representations* (2019)

[Crisci 12] Crisci, C., Ghattas, B., and Perera, G.: A review of supervised machine learning algorithms and their applications to ecological data, *Ecological Modelling*, Vol. 240, pp. 113–122 (2012)

[Daume III 06] Daume III, H. and Marcu, D.: Domain adaptation for statistical classifiers, *Journal of artificial Intelligence research*, Vol. 26, pp. 101–126 (2006)

[Duan 17] Duan, Y., Andrychowicz, M., Stadie, B., Ho, O. J., Schneider, J., Sutskever, I., Abbeel, P., and Zaremba, W.: One-shot imitation learning, in *Advances in neural information processing systems*, pp. 1087–1098 (2017)

[Finn 16] Finn, C., Levine, S., and Abbeel, P.: Guided cost learning: Deep inverse optimal control via policy optimization, in *International Conference on Machine Learning*, pp. 49–58 (2016)

[Finn 17] Finn, C., Abbeel, P., and Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks, in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1126–1135JMLR. org (2017)

[Fu 18] Fu, J., Luo, K., and Levine, S.: Learning Robust Rewards with Adverserial Inverse Reinforcement Learning, in *International Conference on Learning Representations*, pp. https://openreview.net/forum?id=rkHywl–A– (2018)

[Goodfellow 14] Goodfellow, I., 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*, pp. 2672–2680 (2014)

[Goodfellow 16] Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y.: *Deep learning*, Vol. 1, MIT press Cambridge (2016)

[Grünwald 04a] Grünwald, P. D. and Dawid, A. P.: Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory, *Annals of Statistics*, pp. 1367–1433 (2004)

[Grünwald 04b] Grünwald, P. D., Dawid, A. P., et al.: Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory, *the Annals of Statistics*, Vol. 32, No. 4, pp. 1367–1433 (2004)

[Hirakawa 18] Hirakawa, T., Yamashita, T., Tamaki, T., Fujiyoshi, H., Umezu, Y., Takeuchi, I., Matsumoto, S., and Yoda, K.: Can AI predict animal movements? Filling gaps in animal trajectories using inverse reinforcement learning, *Ecosphere*, Vol. 9, p. e02447 (2018)

[Ho 16] Ho, J. and Ermon, S.: Generative adversarial imitation learning, in *Advances in neural information processing systems*, pp. 4565–4573 (2016)

[Huth 04] Huth, M. and Ryan, M.: *Logic in Computer Science: Modelling and reasoning about systems*, Cambridge university press (2004)

[Jiang 08] Jiang, J.: A literature survey on domain adaptation of statistical classifiers, *URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey*, Vol. 3, pp. 1–12 (2008)

[Kitahara 13] Kitahara, T. and Mizuno, S.: A bound for the number of different basic solutions generated by the simplex method, *Mathematical Programming*, Vol. 137, No. 1-2, pp. 579–586 (2013)

[Kitani 12] Kitani, K. M., Ziebart, B. D., Bagnell, J. A., and Hebert, M.: Activity forecasting, in *European Conference on Computer Vision*, pp. 201–214Springer (2012)

[Kuderer 15] Kuderer, M., Gulati, S., and Burgard, W.: Learning driving styles for autonomous vehicles from demonstration, in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pp. 2641–2646IEEE (2015)

[Lamblin 10] Lamblin, P. and Bengio, Y.: Important Gains From Supervised Fine-Tuning of Deep Architectures on Large Labeled Sets, in *NIPS* 2010 Deep Learning and Unsupervised Feature Learning Workshop*, pp. 1–8 (2010)

[Levine 11] Levine, S., Popovic, Z., and Koltun, V.: Nonlinear inverse reinforcement learning with gaussian processes, in *Advances in Neural Information Processing Systems*, pp. 19–27 (2011)

[Li 17] Li, K. and Burdick, J. W.: Meta Inverse Reinforcement Learning via Maximum Reward Sharing for Human Motion Analysis, *CoRR*, Vol. abs/1710.03592, (2017)

[Mnih 13] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M.: Playing atari with deep reinforcement learning, *arXiv preprint arXiv:1312.5602* (2013)

[Ng 00] Ng, A. Y., Russell, S. J., et al.: Algorithms for inverse reinforcement learning., in *Icml*, pp. 663–670 (2000)

[Pan 17] Pan, X., You, Y., Wang, Z., and Lu, C.: Virtual to real reinforcement learning for autonomous driving, *arXiv preprint arXiv:1704.03952* (2017)

[Peng 18] Peng, X. B., Andrychowicz, M., Zaremba, W., and Abbeel, P.: Sim-to-real transfer of robotic control with dynamics randomization, in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8IEEE (2018)

[Pomerleau 89] Pomerleau, D. A.: Alvinn: An autonomous land vehicle in a neural network, in *Advances in neural information processing systems*, pp. 305–313 (1989)

[Ramachandran 07] Ramachandran, D. and Amir, E.: Bayesian inverse reinforcement learning, *IJCAI International Joint Conference on Artificial Intelligence*, pp. 2586–2591 (2007)

[Ratliff 06] Ratliff, N. D., Bagnell, J. A., and Zinkevich, M. A.: Maximum margin planning, in *Proceedings of the 23rd international conference on Machine learning*, pp. 729–736 (2006)

[Rockafellar 70] Rockafellar, R. T.: *Convex analysis*, No. 28, Princeton university press (1970)

[Ross 11] Ross, S., Gordon, G., and Bagnell, D.: A reduction of imitation learning and structured prediction to no-regret online learning, in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635 (2011)

[Rumelhart 86] Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: Learning representations by back-propagating errors, *nature*, Vol. 323, No. 6088, pp. 533–536 (1986)

[Russell 98] Russell, S. J.: Learning agents for uncertain environments, in *COLT*, Vol. 98, pp. 101–103 (1998)

[Shimodaira 00] Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function, *Journal of statistical planning and inference*, Vol. 90, No. 2, pp. 227–244 (2000)

[Sigaud 13] Sigaud, O. and Buffet, O.: *Markov decision processes in artificial intelligence*, John Wiley & Sons (2013)

[Surana 14a] Surana, A. and Srivastava, K.: Bayesian nonparametric inverse reinforcement learning for switched markov decision processes, in *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*, pp. 47–54IEEE (2014)

[Surana 14b] Surana, A. and Srivastava, K.: Bayesian nonparametric inverse reinforcement learning for switched Markov decision processes, in *2014 13th International Conference on Machine Learning and Applications*, pp. 47–54IEEE (2014)

[Sutton 18] Sutton, R. S. and Barto, A. G.: *Reinforcement learning: An introduction*, MIT press (2018)

[Thrun 98] Thrun, S. and Pratt, L.: Learning to learn: Introduction and overview, in *Learning to learn*, pp. 3–17, Springer (1998)

[Tobin 17] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world, in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30IEEE (2017)

[Vempala 05] Vempala, S.: Geometric random walks: a survey, *Combinatorial and computational geometry*, Vol. 52, No. 573-612, p. 2 (2005)

[Whitley 94] Whitley, D.: A genetic algorithm tutorial, *Statistics and computing*, Vol. 4, No. 2, pp. 65–85 (1994)

[Wulfmeier 15] Wulfmeier, M., Ondruska, P., and Posner, I.: Maximum entropy deep inverse reinforcement learning, *arXiv preprint arXiv:1507.04888* (2015)

[Yan 17] Yan, M., Frosio, I., Tyree, S., and Kautz, J.: Sim-to-real transfer of accurate grasping with eye-in-hand observations and continuous control, *arXiv preprint arXiv:1712.03303* (2017)

[Yosinski 14] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H.: How transferable are features in deep neural networks?, in *Advances in neural information processing systems*, pp. 3320–3328 (2014)

[Ziebart 08] Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K.: Maximum Entropy Inverse Reinforcement Learning., in *AAAI*, Vol. 8, pp. 1433–1438Chicago, IL, USA (2008)

[Ziebart 10] Ziebart, B. D.: *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*, Carnegie Mellon University (2010)

# 謝辞

本論文をまとめるにあたり，終始あたたかくご指導ご鞭撻いただいた荒井幸代教授に心より感謝いたします．投稿論文が初めて採録に至るまでの間は悩むこともありましたが，荒井先生のおかげで乗り越えることができました．ありがとうございました．また，学部 4 年生から過ごした 5 年間の研究室生活でたくさんのアドバイスを下さった先輩，後輩，同級生の皆様に感謝いたします．

# 研究業績

【学術雑誌に発表した論文】

1. 中田勇介，荒井幸代， 状態遷移確率の異なる MDP 環境間で無矛盾な報酬の推定法， 人工知能学会論文誌，2019.11.https://doi.org/10.1527/tjsai.B-J23 【査読 2 名】
2. 中田勇介，荒井幸代， 複数環境におけるエキスパート軌跡を用いたベイジアン逆強化学習 人工知能学会論文誌，2020.1. https://doi.org/10.1527/tjsai.G-J73 【査読 2 名】

【国際会議における発表】

1. Yusuke Nakata, Yuki Kitazato, Sachiyo Arai, Detection of Features Affording a Certain Action via Analysis of CNN The 3rd IEEE International Conference on Agents, pp. 105 – 108, 2018.8, Singapore.10.1109/AGENTS.2018.8460062. 【口頭発表，査読あり】
2. Yusuke Nakata, Sachiyo Arai, Bayesian Inverse Reinforcement Learning for Expert's Demonstrations in Multiple Dynamics, Adaptive and Learning Agents Workshop 2019 ,2019.5, Montreal. 【ポスター発表，査読あり】
3. Yusuke Nakata, Sachiyo Arai, Mini-batch Bayesian Inverse Reinforcement Learning for Multiple Dynamics, 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), ex1330, 2020.5. 【口頭発表，査読あり】

【国内学会・シンポジウムにおける発表】

1. 中田勇介，荒井幸代， 深層学習の中間層解析に基づくアフォーダンスの設計に有用な特徴の抽出，Joint Agent Workshops & Symposium， Number : 37，2016.9, 岐阜.【ポスター発表, 査読なし】
2. 中田勇介，荒井幸代， 設計支援に向けたアフォーダンスを誘発する特徴の識別， 2017 年度人工知能学会全国大会, 2M4-OS-32a-4，2017.5, 愛知.【口頭発表, 査読なし】
3. 中田勇介, 荒井幸代 エキスパートが複数の環境で生成した軌跡から報酬を推定するベイジアン逆強化学習, 2019 年度人工知能学会全国大会, 2Q5-J-2-01, 2019.6，新潟. 【口頭発表, 査読あり】
4. 中田勇介, 荒井幸代 複数環境におけるエキスパート軌跡を用いたミニバッチベイジアン逆強化学習, Joint Agent Workshops & Symposium，2019.9, 大分.【口頭発表, 査読あり】

【受賞歴】

1. 深層学習の中間層解析に基づくアフォーダンスの設計に有用な特徴の抽出，Joint Agent Workshops & Symposium 2016，優秀発表賞（2016.9 岐阜）
2. CNN の解析によるアフォーダンスを誘発する特徴の識別，千葉大学工学部都市環境システムコース，卒業研究奨励賞（2017.3）

3. 複数環境におけるエキスパート軌跡を用いたミニバッチベイジアン逆強化学習，Joint Agent Workshops & Symposium 2019, 研究奨励賞（2019.9 大分）