別紙２

令和３年度　防衛装備庁
安全保障技術研究推進制度

# 研究成果報告書
# 潜在脳ダイナミクス推定法の開発と精神状態推移の解明と制御

令和４年５月

株式会社国際電気通信基礎技術研究所

本報告書は、防衛装備庁の安全保障技術研究推進制度による委託業務として、株式会社国際電気通信基礎技術研究所が実施した令和３年度「潜在脳ダイナミクス推定法の開発と精神状態推移の解明と制御」の成果を取りまとめたものです。

## １．委託業務の目的

本研究では、熟練者の作業中の意思決定過程や、人の精神状態の時空間ダイナミクスの理解と制御を達成するために、（１）アルゴリズムの開発と工学応用、（２）ダイナミクスを記述する特徴量・学習データの抽出と医療応用、（３）アルゴリズムの医療応用、を通して人の精神状態の時空間ダイナミクスの解明を目指す。

### （１）アルゴリズムの開発と工学応用

本項目では、（１－１）目標となる高次元の時空間ダイナミクスからプランニング可能な低次元の潜在脳ダイナミクスを推定するアルゴリズムと、（１－２）現在のシステムの潜在状態を目標状態に近づける順・逆強化学習アルゴリズムを開発する。開発した二つのアルゴリズムを統合して、（１－３）人の動作をロボットに移植するスキル伝達（模倣学習）の問題に適用し、これまで導入に膨大なコストを要したロボットの実問題への適用を目指す。特に検証する課題として、シミュレーションが困難な衣類の折り畳みといった柔軟物の操り課題を外部からの詳細な報酬無しで人からの行動事例から模倣する。

### （１－１）潜在脳ダイナミクス推定アルゴリズムの目標仕様

高次元な時空間データを少量のデータで制御するためにはデータの背後に潜在するダイナミクスをマルコフ決定過程(Markov Decision Process; MDP)として推定することが不可欠である。ロボットによる模倣実験と精神疾患の解析に要求される仕様は以下の通りとなる。

- 入力となる観測情報の次元：320×240 の RGB 画像（ロボット実験の場合）、10 万次元以上のボクセルデータ (fMRI の場合)
- 出力となる潜在脳ダイナミクスの表現 : 10 個程度の離散状態（fMRI の場合）、もしくは 100 次元程度の連続状態（ロボット実験の場合）。

<u>中間目標：</u>

(i) 連続ベンチマーク課題を用いた検証

連続ベンチマーク課題において、潜在状態次元数を入力データ次元の 2 分の 1 から 10 分の 1 程度圧縮した連続 MDP として記述する。（１－２）の模倣学習アルゴリズムと統合する。MDP 推定アルゴリズムを利用しない強化学習アルゴリズムと比較し、2 分の 1 から 5 分の 1 程度の学習データ数で同程度の性能を目指す。

<u>最終目標：</u>

(i) 離散ベンチマーク課題を用いた検証

離散ベンチマーク課題において、潜在状態数 10 から 1000 程度の離散 MDP として記述する。（１－２）の模倣学習アルゴリズムと統合する。MDP 推定アルゴリズムを利用しない強化学習アルゴリズムと比較し、5 分の 1 から 10 分の 1 程度の学習データ数で同程度の性能を目指す。

(ii) ニューロフィードバック課題を用いた検証

精神疾患チームから得られるデータに対し、潜在状態数 2 から 10 個程度の離散連続 MDP として記述する。ニューロフィードバック技術と統合する。従来のダイナミクスを考慮しない場合と比較し、統計的に有意なレベルでデータの予測を目指す。

### （１－２）システムの潜在状態を目標状態に近づける模倣学習アルゴリズムの目標仕様

システムの状態を現在の潜在状態から目標状態へ近づけるアルゴリズムはヒトの意思決定の計算モデルとして広く用いられ、ロボット制御の分野でも近年注目されている順強化学習を用い、目標状態との差異は逆強化学習によって推定する。

<u>中間目標：</u>

(i) 連続ベンチマーク課題を用いた検証

連続ベンチマーク課題において、モデルフリー強化学習と比較し 2 分の 1 から 5 分の 1 程度の学習データ数で同程度の性能を目指す。

(ii) 離散ベンチマーク課題を用いた検証

離散ベンチマーク課題において、モデルフリー強化学習と比較し 2 分の 1 から 5 分の 1 程度の学習データ数で同程度の性能を目指す。
(iii) 連続システムに対するモデル学習法の提案
連続状態行動システムに適用可能なモデル学習法を開発し、標準的な深層学習を用いたモデル学習よりも 2 分の 1 から 5 分の 1 程度の学習データ数で同程度のモデル学習性能を目指す。

最終目標:
(i) 連続ベンチマーク課題を用いた検証
連続ベンチマーク課題において、モデルフリー強化学習と比較し 5 分の 1 から 10 分の 1 程度の学習データ数で同程度の性能を目指す。
(ii) 離散ベンチマーク課題を用いた検証
離散ベンチマーク課題において、モデルフリー強化学習と比較し 5 分の 1 から 10 分の 1 程度の学習データ数で同程度の性能を目指す。
(iii) 連続システムに対するモデル学習法の提案
連続状態行動システムに適用可能なモデル学習法を開発し、標準的な深層学習を用いたモデル学習よりも 5 分の 1 から 10 分の 1 程度の学習データ数で同程度のモデル学習性能を目指す。
(iv) 離散状態行動システムに適用可能なモデル学習法を開発し、標準的な深層学習を用いたモデル学習よりも 5 分の 1 から 10 分の 1 程度の学習データ数で同程度のモデル学習性能を目指す。

**（１－３）潜在脳ダイナミクスを利用した模倣学習の実ロボットによる検証**
本研究は基礎研究であり、特に近年は標準的なベンチマーク課題を用いて検証して他の手法と広く比較することが世界的に要求されている。そのため、これまでに多くの従来研究でなされてきた下記の課題で検証する。

最終目標
(i) 室内環境における移動ロボットのナビゲーション
室内環境における小型移動ロボットを用いたナビゲーション課題に対して、人がロボットを操作することで上記課題を達成した際の行動データから、ロボットの方策を学習する。モデルフリー強化学習と比較し 5 分の 1 から 10 分の 1 程度の学習データ数で同程度の学習性能を目指す。
(ii) アーム型ロボットを用いた物体の操り動作（柔軟物の操りなど）
ハンカチの裏返し、T シャツの折り畳みなどのアーム型ロボットを用いた柔軟物の操り課題に対して、人が上記課題を達成する際の行動データをモーションキャプチャ等により計測する。この行動データからロボットが上記課題を達成する方策を移植する。モデルフリー強化学習と比較し 5 分の 1 から 10 分の 1 程度の学習データ数で同程度の学習性能を目指す。

**（２）ダイナミクスを記述する特徴量・学習データの抽出と医療応用**
本項目では脳活動の空間パターンの医療応用と因果的解釈を通じて、精神疾患患者の脳活動パターンの時系列ダイナミクスを記述する特徴量を抽出するために、（２－１）脳活動の空間パターンと疾患状態の因果的関係の解析、（２－２）空間パターンのみを考慮したニューロフィードバックによる治療法の開発を目指す。本研究では後述（（３）アルゴリズムの医療応用など）のとおり、精神疾患の状態変動と外的要因との関連を目標の一つとして置いている。そのため、当初はこれらの関連の解析が容易な PTSD をモデルケースとして対象の中心とする。PTSD は疾患発症の原因が特定のトラウマ体験であり、外的要因との関連も他の疾患と比較して関連付けやすい。また、行動嗜癖や依存など疾患の増悪因子が関連付けしやすい疾患群、また統合失調症やうつ病などの有病率が高い疾患のサブクリニカル群（健常者集団の中で特定疾患の傾向を示す群）なども対象に含める。

**（２－１）脳活動の空間パターンと疾患状態の因果的関係の解析**
時系列ダイナミクスを抽出するためにはまず、固定した時点での空間パターンの因果的解釈が必要となる。空間パターンの因果的解釈には、先進的なニューロフィードバックを用いる。古典

的なニューロフィードバックが局所的な脳領域の平均活動レベルを目標状態に据えたのに対し、先進的なニューロフィードバックでは局所内､あるいは全脳の脳活動パターンを目標状態とする。したがって、先進的なニューロフィードバック前後での疾患状態の変化を観察することで、脳活動の空間パターンと疾患状態の因果的関係を解析する｡ニューロフィードバックの実施に先立ち、ニューロフィードバックで誘導の目標とする脳活動を定義するために、①脳活動からの健常者と精神疾患の判別、②外的要因（例：治療や精神的ストレス）曝露時の脳活動の同定、③外的要因曝露前後の脳活動変化の解明を実施する。

中間目標:

(i)　PTSD のデータ収集（60－100 例）、脳活動による健常者と PTSD の判別
(ii)　ストレス曝露により変化する脳活動パターンの同定
(iii) ストレス耐性を予測する脳活動の同定
(IV)　EEG/fMRI を用いた脳ダイナミクスの可視化
　　※可視化されたダイナミクスの統計的な有意性

最終目標:

(i) マインドフルネスにより変化する脳活動パターンの同定
(ii)ストレス耐性を予測する脳活動の同定

**（２－２）空間パターンのみを考慮したニューロフィードバックによる治療法の開発**

（２－１）にもとづき、空間パターンへのニューロフィードバックによる疾患治療法の開発を目指す。また、治療効果が十分でない疾患については本項目の目標を脳活動の因果的解釈と位置づけ、得られた結果を、時系列ダイナミクスを考慮したニューロフィードバック（最終目標（３））に活用する。つまり、同定した疾患の原因となる脳活動の空間パターンは時系列ダイナミクスを記述する特徴量として（３）の研究で活用する。例えば、PTSD においては恐怖条件付けの消去学習（PTSD の既存治療の枠組みとして知られる）に影響するような脳活動パターン（具体的には腹内側前頭前野と扁桃体の活動のバランスやトラウマ刺激を見ている時の脳活動など）に介入し、消去学習の効率との因果的解釈を行う。これにより、消去学習効率の高い空間パターン（具体的には腹内側前頭前野の活動亢進と扁桃体の活動抑制など）や時系列ダイナミクスを記述する特徴量として扱うことが可能となる。

一例として、恐怖刺激に介入した我々の研究では、消去学習を模した手順により恐怖刺激と扁桃体活動の連合を弱めることに成功している。PTSD への治療法を開発とした研究でも同様の手順（下記）で症状を改善できる可能性がある。

(i) トラウマ刺激と非トラウマ刺激を視覚野の脳活動で判別する
(ii) DecNef によりトラウマ刺激に対応する脳活動を自助努力により誘導する
(iii)トラウマ刺激の脳活動を誘導できた試行に報酬を与える
(iv) ii～iii をスキャナー内で反復することで被験者は特定脳活動の誘導方法を学ぶ
(v)これにより、トラウマ刺激と恐怖の連合を報酬との連合で上書きする

中間目標:

(i) うつ度軽減ニューロフィードバックのプロトコル確立：
　3 種類のプロトコル（先行研究の手法、異なる時間窓、異なる教示方法）について 10 人ずつニューロフィードバックを行い、効果を比較する
(ii) PTSD 患者 5 例を対象としたニューロフィードバックによる治療効果の確認：
　介入前後での群内比較により効果検証

最終目標:

(i) うつ度軽減ニューロフィードバックのプロトコル決定：
　より良いプロトコルで効果検証を行い、最適なプロトコルの確立
(ii) PTSD 患者へのニューロフィードバックによる治療効果の確認：

ダブルブラインドで群間比較により効果検証
(iii) フィードバック法とマインドフルネス介入の併用効果の検証：
介入前後での抑うつ傾向、クリエイティブ等機能の統計的有意な改善

**（３）アルゴリズムの医療応用**

（１）で開発したアルゴリズムと（２）で得られた知見をもとに（３－１）各種精神状態の時空間ダイナミクスとその外的要因との関係を明らかにし、（３－２）時空間ダイナミクスを考慮したニューロフィードバック技法による新規治療法の開発を目指す。各疾患の治療効果の尺度としては、研究や臨床でゴールドスタンダードとして用いられている臨床スコアを用いる。例えば、自閉症では Autism Diagnostic Observation Schedule Second Edition (ADOS-2)、 PTSD では Clinician-Administered PTSD Scale (CAPS) 、うつ病では Hamilton Rating Scale for Depression (HAM-D)を用いる。

○治療を目指した目標においては、下記 2 点をもって最終目標の達成とする：
1. 臨床的に意義があるとされる以上の臨床スコアの改善（または改善効果への影響）
2. 対照群と比較して統計的に有意な臨床スコアの改善（または改善効果への影響）

○脳活動の同定においては、下記 1 点をもって最終目標の達成とする：
1. 独立データ（※1）情報の統計的に有意な予測性能
例：治療前の脳活動からその治療効果を統計的に有意な精度で予測できる
※1 判別器作成に用いた訓練データと独立したテスト用データ

**（３－１）各種精神状態の時系列ダイナミクスとその外的要因との関係**

- 開発したアルゴリズムをもとに、大規模な時系列データから潜在ダイナミクスを同定する
- 疾患の発症や増悪と関係する脳活動を同定する
- 治療の効果を最適化する潜在脳ダイナミクスを同定する
- 疾患の増悪因子に脆弱な潜在脳ダイナミクスを同定する
- 治療時期の制御や、特定の疾患増悪因子の予防による症状への影響を明らかとする

中間目標:
(i) レトロスペクティブデータからの特徴量候補の同定：
受診時の症状プロファイルから病気の進行を有意に予測する
(ii) 定期的に脳活動取得する PTSD 患者１０例のリクルート及び取得開始

最終目標：
(i) 潜在脳ダイナミクスの可視化：
一か月単位での潜在脳ダイナミクスの変化を有意に予測可能とする

**（３－２）潜在脳ダイナミクスを考慮したニューロフィードバック**

- 精神疾患の症状改善効果を検証する
- 潜在脳ダイナミクスへの介入による治療反応性の変化を検証する
- 潜在脳ダイナミクスへの介入による外的要因への反応性の変化を検証する
- 健常者や患者への各種認知機能への介入効果を検証する

最終目標:
(i) ダイナミクスを考慮したニューロフィードバックによる認知機能の改善：
介入前後での症状の改善を評価
(ii) ダイナミクスを考慮したニューロフィードバックによる PTSD 症状の改善：
介入前後での症状の改善を評価

**２．研究開始時に設定した研究目標の達成度**

本プロジェクトは令和3年度までとなったため、中間目標ごとの達成度を記載する。

実施項目(1) アルゴリズムの開発と工学応用

設定した全ての中間目標は 100%達成しており、実ロボットを用いた検証を計画よりも早く実施できた。1-1(i) 潜在脳ダイナミクス推定アルゴリズムの連続ベンチマーク課題を用いた検証は達成度 100%である。1-2(i) システムの潜在状態を目標状態に近づける模倣学習アルゴリズムの連続ベンチマーク課題を用いた検証では、開発したモデルベース順・逆強化学習アルゴリズムはモデルを明示的に推定しないモデルフリーと比較し、1/10 程度のデータ数で同程度の性能を実現し、当初目標を達成している。1-2(ii) 離散ベンチマーク課題を用いた検証では、移動ロボット Turtlebot3 waffle pi を用いたシミュレーションにより、3432 次元の画像・Lidar 情報を 616 個の離散潜在表現に変換（1/5 程度に圧縮）し、モデルフリーと比べて 1/2 程度のデータ数で同程度の性能を実現し、当初目標を達成している。1-2(iii) 連続システムに対するモデル学習法についても 100%達成済みである。

実施項目(2) ダイナミクスを記述する特徴量・学習データの抽出と医療応用

設定した中間目標の多くは達成済みである。2-1(i) コロナ禍の状況においてデータ収集に遅れがあったが、脳活動の空間パターンと疾患状態の因果関係の解析について最低限のデータは収集できた。予備解析において、脳活動から疾患の有無を判別するアルゴリズムも構築できた。新規のコホートデータはテストデータとして用い、構築したアルゴリズムの有用性を評価した。2-1(ii)のストレス暴露による変化する脳活動パターンの同定および(iii)のストレス耐性を予測する脳活動の同定は達成済みである。2-1(iv)の EEG/fMRI を用いた脳ダイナミクスの可視化も達成済みである。2-2(i) うつ度軽減ニューロフィードバックのプロトコル開発の達成度は 80%である。現在までに、3 種類のプロトコル（先行研究の手法、異なる時間窓、異なる教示方法）についてニューロフィードバックを実施した。先行研究の手法についてはニューロフィードバックの効果・再現性・長期効果を確認した。また対象脳機能結合に関わる症状のみの改善を認めた。(ii) PTSD 患者 5 例を対象としたニューロフィードバックによる治療効果は当初計画よりも早く確認できた。実験条件 6 例での効果検証に加えて、対照条件を 4 例に対し実施し、実験条件でえられた効果が単なるプラセボでない可能性が高いことを確認した。

実施項目(3) アルゴリズムの医療応用

3-1(i)レトロスペクティブデータからの特徴量候補の同定は 100%達成済みである。3-1(ii)定期的に脳活動取得する PTSD 患者 10 例のリクルート及び取得開始はコロナ禍の影響による遅延のため、達成度は 50%程度となった。これは、データ収集開始後に実験中断となることを避けるための措置であり、遅延の間に緊急事態宣言中も実験を継続できる態勢を構築した。また、遅延のための代替として横断データを拡充し、短期間での潜在脳の変化に対応すると考えられる行動指標の変動の確認に成功した。

## ３．委託業務における研究の方法及び成果

### 3.0 準備

強化学習の問題設定

本項目では、基礎技術となる（順）強化学習について概説する。強化学習は所望の動作を達成する方策（制御則）を取得するための計算論的な枠組みであり、制御工学や神経科学、心理学など多くの意思決定に関する分野に関連する。標準的なマルコフ決定過程(Markov Decision Process; MDP)を仮定する。状態空間を$\mathcal{X}$、行動空間を$\mathcal{U}$、状態$x \in \mathcal{X}$で行動$u \in \mathcal{U}$を実行した時に状態$x' \in \mathcal{X}$に遷移する確率を$p_e(x' \mid x, u)$とする。状態行動対に対する即時報酬を$r(x, u)$とする。この即時報酬は制御工学における即時コストの正負を反転したものである。状態$x$で行動$u$を選択する確率を方策と呼び、$\pi(u \mid x)$と記す。強化学習の目的は積算報酬を最大にする方策を求めることである。

強化学習において最も重要な関数は状態価値関数と呼ばれる、方策$\pi$のもとで将来にわたって得られる総報酬の期待値で、次のように定義される。

$$V^{\pi}(x) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, u_t) \mid x_0 = x\right].$$

ここで$\gamma \in [0,1]$は将来得られる報酬を割り引くための定数である。$\mathbb{E}_{\pi}$は方策と状態遷移確率のも

とで得られる状態行動系列に関する期待値である。$V(x)$は方策$\pi$の状態$x$における価値を表し、価値の大きいほどその状態は良いとみなすことができ、各状態において状態価値を最大化する方策を最適方策と呼び、対応する状態価値関数を最適状態価値関数と呼び、

$$V^*(x) = \max_{\pi} V^{\pi}(x), \forall x$$

と表す。この式の右辺を1ステップだけ展開するとベルマン最適方程式

$$V^*(x) = \max_{u} \left\{ r(x,u) + \gamma \int p_e(x' \mid x,u) V^*(x) \mathrm{d}x \right\},$$

が得られる。右辺のmax演算子のため、ベルマン最適方程式は非線形である。

近年の主要な手法であるエントロピ正則化強化学習は、報酬関数に方策のエントロピを追加したもので、max演算子をソフト化し微分可能な演算子に置き換える効果を持つ。9]。具体的には報酬関数が

$$r(x,u) \leftarrow r(x,u) + \kappa^{-1}\mathcal{H}\big(\pi(\cdot \mid x)\big) - \eta^{-1}\mathrm{KL}\big(\pi(\cdot \mid x) \,\|\, b(\cdot \mid x)\big)$$

のように正則化される。ここで$\mathcal{H}(\pi(\cdot \mid x))$は方策$\pi$のShannonエントロピ、KL$(\pi(\cdot \mid x) \,\|\, b(\cdot \mid x))$は$\pi$とベースライン方策$b(u \mid x)$の間のKullback-Leibler (KL) ダイバージェンス、$\kappa, \eta$は実験者の定めるメタパラメータである。エントロピの役割は最適方策が決定論的になることを防ぎ、探索を促進する。KL ダイバージェンスの役割はベースライン方策からあまり逸脱しないように方策改善ステップを保守的にする。またベルマン最適性方程式は$p_e$に依存する。

$p_e$を明示的に推定しベルマン最適方程式の求解に積極的に利用する方法をモデルベース強化学習と呼ぶ。一方、ベルマン最適方程式の積分計算をモンテカルロ推定に置き換えることで$p_e$を明示的に推定しない方法をモデルフリーと呼ぶ。

順・逆強化学習を組み合わせた模倣学習

一般に強化学習を適用するためには報酬を事前に準備する必要がある。強化学習の成功例の一つである囲碁への応用の場合、対戦の結果勝利すれば+1、敗北すれば−1、それ以外は0といった非常にスパースな報酬関数を使えば原理的には学習できる。スパースな報酬は容易に設計できるが、代償として実世界では不可能な試行回数を必要とするため、ロボット学習に用いるのは適切ではない。そこで注目されるのが逆強化学習である。逆強化学習では報酬関数を設計する代わりに、タスクを達成しているエキスパートからの行動データが利用できると仮定し、エキスパートの報酬関数をデータから推定する。ロボット制御の問題では正解行動をデモンストレーションとして準備できる場合も多く、人や動物の意思決定の解析も可能となるため、非常に多くの研究がなされている。特に逆強化学習と通常の順強化学習の組み合わせが敵対的生成ネットワーク (Generative Adversarial Networks; GAN) として定式化できることが示された。逆強化学習は GAN における識別器に対応し、エキスパートからのデータとロボット自身が生成したデータを区別するよう識別器を訓練する。順強化学習はGAN における生成器に対応し、逆強化学習から導出される報酬の期待積算を最大にするように方策を更新する。デモンストレーションから教師あり学習として求める行動クローニング (Behavioral Cloning; BC) と異なり、事前に与えられていない状況に対しても適切な行動が順強化学習で学習される。

順・逆強化学習とニューロフィードバックの関係

ロボット制御における順・逆強化学習を用いた模倣学習の場合、目標となるエキスパートの行動とロボット自身が生成した行動の「距離」を逆強化学習によって推定する（図 1(a)参照）。具体的には両者のデータを識別する識別器を構成する。識別器はディープロジスティック回帰によって実装されるため、エキスパートの行動分布とロボットの行動分布の尤度比が重要な役割を果たす。次にロボットは逆強化学習によって推定された尤度比をもとに、行動を生成する制御方策を順強化学習によって改善する。このとき、エキスパートの行動の尤度を報酬とするエントロピ正則化強化学習を行っていると解釈できる。

一方ニューロフィードバックでは、目標となる脳活動パターンと現在の脳活動パターンを区別するデコーダを事前に訓練し、現在の脳活動パターンが目標にどれだけ近いかを尤度として計算

する。デコーダはスパースロジステック回帰によって実装される。被験者は提示された尤度を最大化するように脳内の活動を変化させる（図 1(b)参照）。

目標と脳活動のパターンを比較にロジステック回帰を用いた分類を行っていることから、逆強化学習の処理は両者で共通している。一方で人の意思決定モデルとして強化学習は広く採用されており、ロボットの行動学習アルゴリズムとしても主要なものである。このことから、一見すると両者は全く異なる問題にとらえられがちであるが、本質的に同じ数理構造を持つ。一方で、模倣学習は高次元時空間パターンを扱うことができず、ニューロフィードバックでは多段の意思決定問題を取り扱うのが困難である。そこで両者を統一的に研究することで互いの研究を加速するとともに、新しい学問分野を創設することが本研究の目的である。

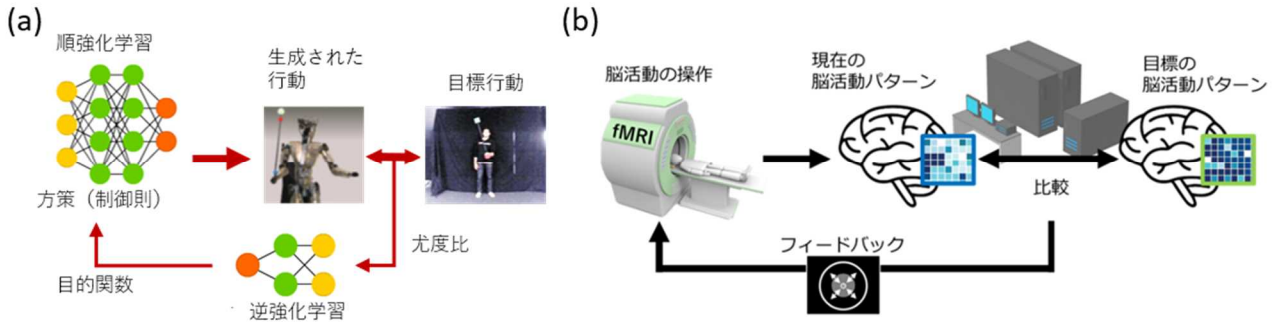


**図 1 ロボットの行動学習とニューロフィードバックが共有する構造。**

### 3.1 (1-1) 離散状態・行動MDPを推定するアルゴリズムの開発

エキスパート方策が実際の環境から生成された状態行動遷移対と学習者の方策とモデルから生成された状態行動対の違いをKLダイバージェンスで測るために、学習者の方策だけでなくモデルのエントロピとKLダイバージェンスを考慮した新しいエントロピ正則化順・逆強化学習を定式化した。また高次元の観測値を扱うために決定論的正則化オートエンコーダを導入し、潜在状態を離散化するためのGumbel-Softmaxを導入した。移動ロボットTurtlebot 3 waffle piを使ったナビゲーション課題によって開発手法を評価した。

### 3.2 (1-2) 連続状態・行動MDPを推定するアルゴリズムの開発

3.0節で述べたエントロピ正則化を導入した離散時間の無限期間MDPを考える。このときベルマン方程式を解くと以下の関係式が得られる。

$$Q(x,u) = r(x,u) + \eta^{-1} \ln b(u|x) + \gamma \mathbb{E}_{x' \sim p_e(\cdot|x,u)}[V(x')]$$

$$V(x) = \beta^{-1} \ln \exp \int \exp\big(\beta Q(x,u)\big)\, du$$

$$\pi(u|x) = \exp\big[\beta\big(Q(x,u) - V(x)\big)\big]$$

$V(x)$は状態価値関数、$Q(x,u)$は状態行動価値関数と呼ばれる。また$\beta = 1/(\kappa + \eta)$と定義している。

モデルフリー模倣学習（順・逆強化学習）であるModel-Free Entropy-Regularized Imitation Learning (MF-ERIL) (Uchibe and Doya, 2021)は、エキスパート方策$\pi^E(u \mid x)$から収集された状態行動遷移対

$$\mathcal{D}^E = \{(x_i, u_i, x_i')\}_{i=1}^{N^E}, \qquad u_i \sim \pi^E(\cdot\,|x_i), \qquad x_i' \sim p_e(\cdot\,|x_i, u_i)$$

と、学習者の方策$\pi^L$から同様に生成された状態行動遷移対

$$\mathcal{D}^L = \{(x_i, u_i, x_i')\}_{i=1}^{N^L}, \qquad u_i \sim \pi^L(\cdot\,|x_i), \qquad x_i' \sim p_e(\cdot\,|x_i, u_i)$$

からエキスパート方策とそれを復元する報酬を推定する。ただし初期状態確率や状態遷移確率は未知である。MF-ERILの目的関数はReverse KLダイバージェンス

$$J^{MF-ERIL}(\theta) = \mathbb{E}_{\pi^L(x,u,x')}\left[\ln \frac{\pi^L(x,u,x')}{\pi^E(x,u,x')}\right]$$

と与えられる。ここで同時分布はマルコフ性の仮定の下

$$\pi^L(x,u,x') = p_e(x' \mid x,u)\pi^L(u \mid x)\pi^L(x)$$
$$\pi^E(x,u,x') = p_e(x' \mid x,u)\pi^E(u \mid x)\pi^E(x)$$

と表される。ここで$\pi^L(x), \pi^E(x)$は状態にのみ依存する分布である。MF-ERILの逆強化学習はReverse KLダイバージェンスの推定に対応し、二つの識別器を用いた密度比推定によってなされる。一つ目の識別器は対数密度比$\ln \pi^L(x)/\pi^E(x)$の推定に対応し、

$$D^{(1)}(x) = \frac{1}{1+\exp(-g(x))}$$

と与えられる。$g(x)$は通常の二値分類問題として推定できる。もう一つの識別器は対数密度比$\ln \pi^L(x,u,x')/\pi^E(x,u,x')$の推定に対応するが、状態遷移確率がキャンセルアウトされること、$\ln \pi^L(x)/\pi^E(x)$のかわりに$g(x)$を用いること、さらにエントロピ正則化強化学習の関係式を用いることで、識別器が構造化できることを利用する。結果として

$$D^{(2)}(x,u,x') = \frac{\exp(\beta\kappa^{-1}\, ln\, \pi^L(u \mid x))}{\exp(\beta f(x,x')) + \exp(\beta\kappa^{-1}\, ln\, \pi^L(u \mid x))}$$

と構築する。ここで

$$f(x,x') = r(x) - \beta^{-1}g(x) + \gamma V(x') - V(x)$$

と定義している。この識別器は(Uchibe, 2018)の拡張となっている。順強化学習はSoft Actor-Critic (Haarnoja et al., 2018)と同様の関係式を用いることで実現できる。

次に環境$p_e$を（1-5）のモデル学習アルゴリズムによってモデル化し、モデルから生成されるデータを使った順・逆強化学習を実施する。最大の特徴はモデル学習を通常の教師あり学習とするのではなく、得られる報酬系列に応じて重みづけする点である。さらに順・逆強化学習の損失関数を実データとモデルのギャップを補正する重点サンプリングを導入する。これは従来のモデルベース模倣学習では無視されてきた点である。なお従来の強化学習は、過去の方策から得られたデータを再利用するために重点サンプリングを用いることが多かったが、本研究では実環境とモデル環境の違いを補正するために用いられる点に注意する。環境のモデルを$q(x' \mid x,u)$とする。状態行動対$(x,u)$において，次の状態$x'$がモデル$q$から生成されたものか，実環境$p_e$から生成されたものかを判定する識別器

$$D^{(3)}(x' \mid x,u) = \begin{cases} 1 & x' \sim q(x' \mid x,u) \\ 0 & x' \sim p_e(x' \mid x,u) \end{cases}$$

を導入する。この識別器の学習も二値分類問題として、以下の目的関数

$$J_D^{(3)} = \mathbb{E}_{x' \sim q(\cdot \mid x,u),(x,u)\sim p^{E/L}}\left[\ln D^{(3)}(x' \mid x,u)\right] + \mathbb{E}_{x' \sim p_e(\cdot \mid x,u),(x,u)\sim p^{E/L}}\left[\ln\left(1 - D^{(3)}(x' \mid x,u)\right)\right]$$

を使って学習できる。ここで第1項の期待値を計算する分布は、$(x,u)$は実際の環境からサンプルされているが、$x'$のみがモデルから生成されているのに対し、第2項は$(x,u,x')$が実際の環境からサンプルされていることに注意されたい。$D^{(3)}$が得られると密度比$q(x' \mid x,u)/p_e(x' \mid x,u)$の推定量が計算される。これをISW($D^{(3)}$)とする。

逆強化学習の損失関数は重点サンプリングを用いて

$$J_D^{(2)} = \mathbb{E}_{(x,u,x')\sim q^L}\left[ISW\left(D^{(3)}\right)\ln D^{(2)}(x,u,x')\right] + \mathbb{E}_{(x,u,x')\sim q^L}\left[\ln\left(1 - D^{(2)}(x,u,x')\right)\right]$$

と修正される。なお、右辺第1項の期待値計算に用いる確率分布を$(x,u,x') \sim p^L$として実環境から得られる状態遷移を用いるように設定したものがMF-ERILの逆強化学習の目的関数となる。同様に順強化学習における状態行動価値関数の損失関数は

$$J_Q = \mathbb{E}_{(x,u,x')\sim q^L}\left[ISW\left(D^{(3)}\right)\left(Q(x,u) - Q(x,u,x')\right)^2\right]$$

と修正される。ここで

$$Q(x,u,x') = r(x) + \eta^{-1} \ln \pi^L(u \mid x) + \gamma \hat{V}(x')$$

であり、$\hat{V}(x')$は深層強化学習で用いられるターゲットネットワークである。以上を重点サンプリングを用いたモデルベースERIL (Model-Based ERIL; MB-ERIL)と呼ぶ。

MB-ERILで用いるモデルは3.5.1節で述べる方法で推定する。またMB-ERILはモデルベース手法であるため、モデルと実環境とのギャップによる学習のバイアスは存在し、学習後期での漸近性能はモデルフリーに劣ることが想定される。そこで3.4.1節で述べるモデルベースとモデルフリーのハイブリッド学習法を導入したMBMF-ERILも開発した。

提案手法の有効性を検証するために、OpenAI gym (Brockman et al., 2016a)で提供されているAnt, Humanoidという2種類のロボット制御課題に適用する。これらのタスクの目的はできるだけ速く移動することである。まず本来設定されている報酬関数をもとに最適方策をTRPO (Schulman et al., 2015)によって学習し、そこから得られるデータを$\mathcal{D}^E$として用いる。対数密度$g(x)$、報酬$r(x)$、状態価値関数$V(x)$、行動価値関数$Q(x,u)$は2層のニューラルネットワークを用い, 活性度関数はReLU、ユニット数はそれぞれ400, 300とした。また方策$\pi^L(u \mid x)$はガウス分布によって構成し，その平均値を同じ構成のニューラルネットワークで表現した。1エポックあたり学習方策によって生成される軌跡は100とし、各軌跡は50個の状態行動遷移対を含むとする。

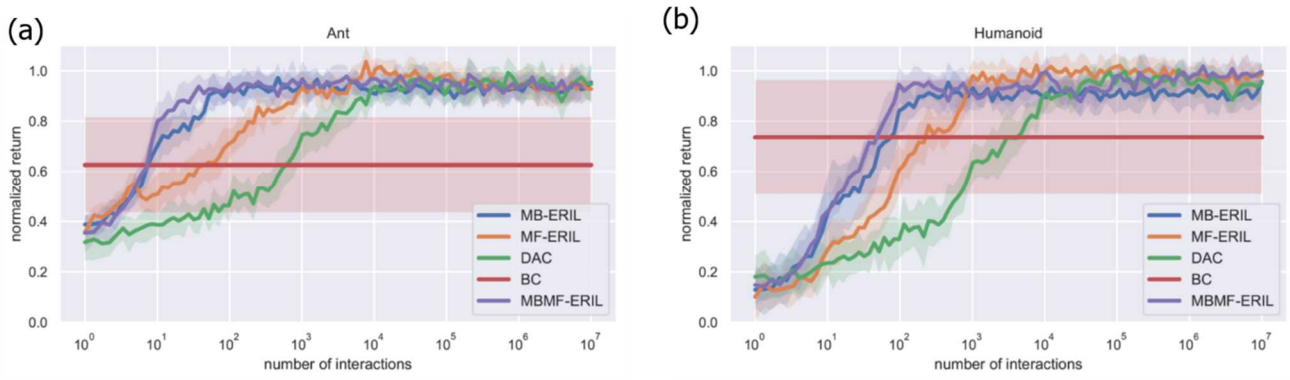


**図 2 順強化学習の性能評価。(a) Ant 環境。(b) Humanoid 環境**

まずエキスパート方策からの軌跡の数をAnt環境では30、Humanoid環境では350と設定したときの順強化学習の性能を比較した。図 2にMB-ERIL、MBMF-ERILとMF-ERIL、DAC、BCを比較した結果を示す。ただし模倣学習は不良設定問題であるため、各手法で推定された報酬は直接比較できない。そこで最終的に獲得された学習方策をシミュレータで提供される本来の報酬を使ったエピソード当たりの総報酬を正規化したもので評価する。Ant環境ではMBMF-ERIL、MB-ERIL、MF-ERIL、DACの順で正規化総報酬の最大値に到達している。BCは従来研究の報告通り、エキスパート方策の性能には到達しなかった。Ant環境ではモデル学習により、モデルフリーのMF-ERIL、DACの学習効率を10倍程度改善することができた。Humanoid環境でも同様の結果が得られたが、最終的な制御性能を確認するとMB-ERILはMF-ERILよりも劣っていることがわかった。Humanoid環境はAnt環境よりも複雑でモデル化誤差の影響が大きいことも一つの原因と考えられ、ハイブリッドにしたMBMF-ERILは漸近性能も改善していることがわかる。実際に最終的に獲得された方策について、負の対数尤度をエキスパート方策から生成したテストデータを用いて評価した。結果を図 3に示す。MBMF-ERILとMF-ERILの方策には統計的な優位差はなかったが、MB-ERILとは統計的な優位差が確認された。

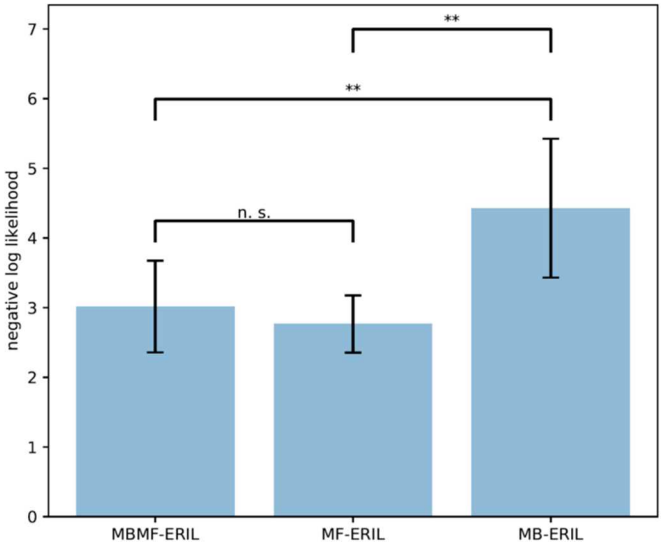


**図 3 Humanoid 環境で得られた最終方策の比較**

次に$\mathcal{D}^E$からのサンプル数を変えることで逆強化学習のデータ効率を評価した。Ho and Ermon (2016)に従い、一つの軌跡が50個の状態行動遷移対を含む、つまり1エピソードあたりのステップ数を50とする。図 4にMB-ERIL、MBMF-ERIL、MF-ERIL、DAC、BCを比較した結果を示す。すべての敵対的生成模倣学習はBCよりも少ないエキスパートデータ数で高い制御性能を示す方策を獲得している。一方でMB-ERIL、MF-ERILの間にはそれほど違いは見られず、モデルの学習による逆強化学習のサンプル効率の改善は得られなかった。MBMF-ERILのようにハイブリッドにしたことによる効果もなかった。

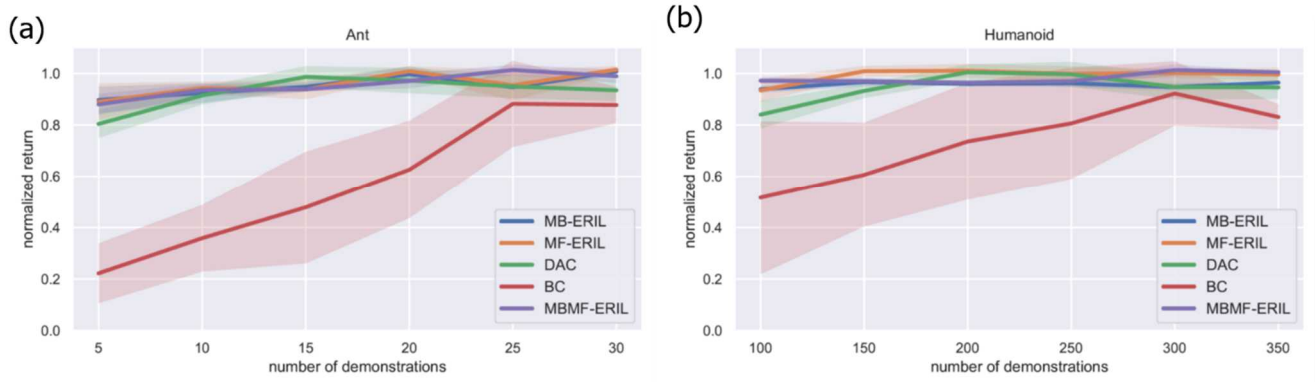


**図 4 逆強化学習の性能評価。(a) Ant 環境。(b) Humanoid 環境**

以上より開発手法はMB-ERIL、DAC、BCよりも学習効率を改善しつつ、通常のモデルベース法で問題となる漸近性能も改善することが確認された。またMF-ERILとMB-ERILを統合することで、漸近性能をさらに改善できることが確認できた。

### 3.3 (1-3) スパースネスの正則化を導入した離散状態・行動MDPを推定するアルゴリズムの開発

3.1、3.2節で開発したエントロピ正則化順・逆強化学習ではShannonエントロピを用いていたが、よりスパースな正則化が可能なTsallisエントロピが利用できるか検討した。

### 3.4 (1-4) モデルベース順・逆強化学習アルゴリズムの開発

#### 3.4.1 モデルフリーとモデルベース強化学習のための同期並列学習

環境の状態遷移を明示的に推定し、そこからのサンプルを利用して学習するモデルベース法はモデルフリー法よりも学習のサンプル効率は優れているが、モデル化誤差のために最終的な漸近性能はモデルベース法よりもモデルフリー法が優れている。両者は一長一短であるため、学習の

進捗状況に応じてモデルベース法とモデルフリー法を切り替えることが望ましい。そこで、我々がこれまでに開発した並列順強化学習法Cooperative and competitive Reinforcement And Imitation Learning (CRAIL) (Uchibe, 2018)を拡張し、モデルベースとモデルフリー強化学習アルゴリズムの動的な切り替えが学習効率の改善に寄与するか、また切り替えがどのようなタイミングで発生するかを調査した。具体的には、経験再生とモデルフリー法、ノンパラメトリックモデルとモデルフリー法、ノンパラメトリックモデルとモデルベース法、そしてパラメトリックモデルの勾配情報を利用したモデルベース法の4種類の性質の異なる強化学習モジュールをCRAILの形式で統合し、複数のモジュールを動的に切り替えることによって、単独の方法では得られないサンプル効率の改善や環境変化に対する適応性が得られるかを検証する。以降、開発した手法をModel-Based CRAIL (MB-CRAIL) (内部, 2020)と呼ぶ。

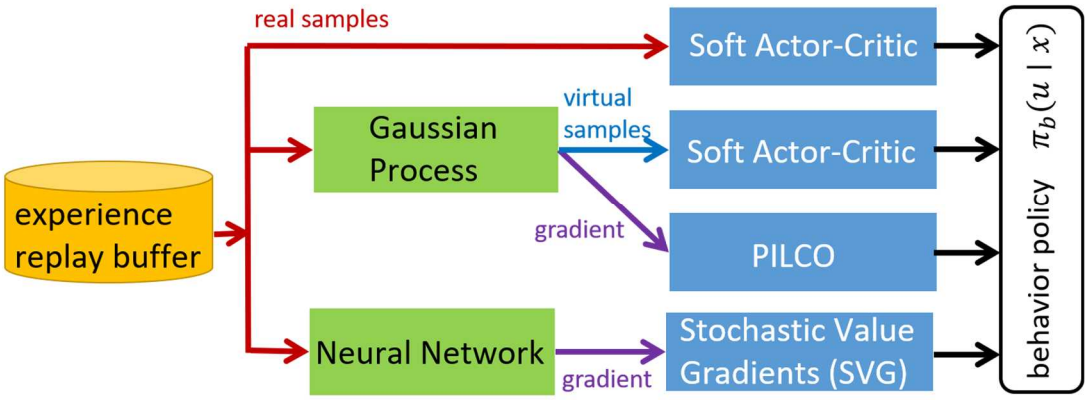


**図 5 同期型 MB-CRAIL のアーキテクチャの例**

図 5 に提案手法のアーキテクチャの例を示す。学習エージェントは$M$個の学習モジュールを持ち、各モジュールは方策$\pi_i(u \mid x)$、状態価値関数$V_i(x)$を持ち、それぞれ異なる強化学習アルゴリズムによって一つの報酬関数をもとに修正される。またエージェントは各モジュールの混合方策によって表現される挙動方策

$$\pi_b(u \mid x) = \sum_{i=1}^{M} \alpha(i \mid x)\pi_i(u \mid x)$$

を持つ。ここで$\alpha(i \mid x)$は状態$x$で学習モジュール$i$を選択する確率で

$$\alpha(i \mid x) \propto \exp\left(\beta V_i(x)\right)$$

と定義する。$\beta$は正のハイパーパラメータである。これは状態$x$でより大きな積算報酬の得られるモジュールを選びやすくする効果がある。

本研究で想定するモデルフリー強化学習は二通りある。一つは経験再生バッファ$\mathcal{D}$から直接生成されたサンプルから学習するモデルフリー強化学習 SAC で、もう一つは$\mathcal{D}$から状態遷移確率モデル$p_e$と報酬関数モデル$r$を明示的に推定し、推定した$p_e$から仮想的にサンプルを生成し、SAC を適用する方法である。モデル学習法としてはガウシアンプロセスを用いる。これを GP-SAC と呼ぶ。次に本実験で用いる二種類のモデルベース強化学習法を説明する。一つは PILCO (Deisenroth et al., 2015)と呼ばれる GP によって$p_e$を推定し、モデルの勾配を利用する方法である。もう一つはニューラルネットワークによって$p_e$を推定しつつ、その勾配情報を用いて価値関数を効率よく計算する方法 Stochastic Value Gradients (SVG) (Heess et al., 2015a)である。

提案手法の有効性を検証するために、OpenAI gym (Brockman et al., 2016)で提供されている Hopper、 Walker2d、 Swimmer、 Half-Cheetah、 Ant、FetchReach という 6 種類のロボット制御課題に適用する（**図 6** 参照）。FetchReach を除く 5 つのタスクの目的はできるだけ速く移動することである。FetchReach は 2 自由度のグリッパを持つ 7 自由度のマニピュレータの手先を目標位置に移動させるタスクで、手先が目標位置に到達した時に非零の報酬が得られるスパース課題、つまり状態空間のほとんどで報酬が 0 であるような課題である。提案手法 MB-CRAIL は先に述べた 4 種類の強化学習法を持つ。SAC で用いる価値関数や方策のネットワーク表現は我々の以前の研

究(Uchibe, 2018)を用いた。$p_e$をガウシアンプロセスで学習し、GP-SAC の強化学習部は SAC と同じものを用いた。SVG は 1 ステップだけ$p_e$を使って予測し、それ以降の系列に対する積算報酬値は価値関数で近似する SVG(1)を用いた。

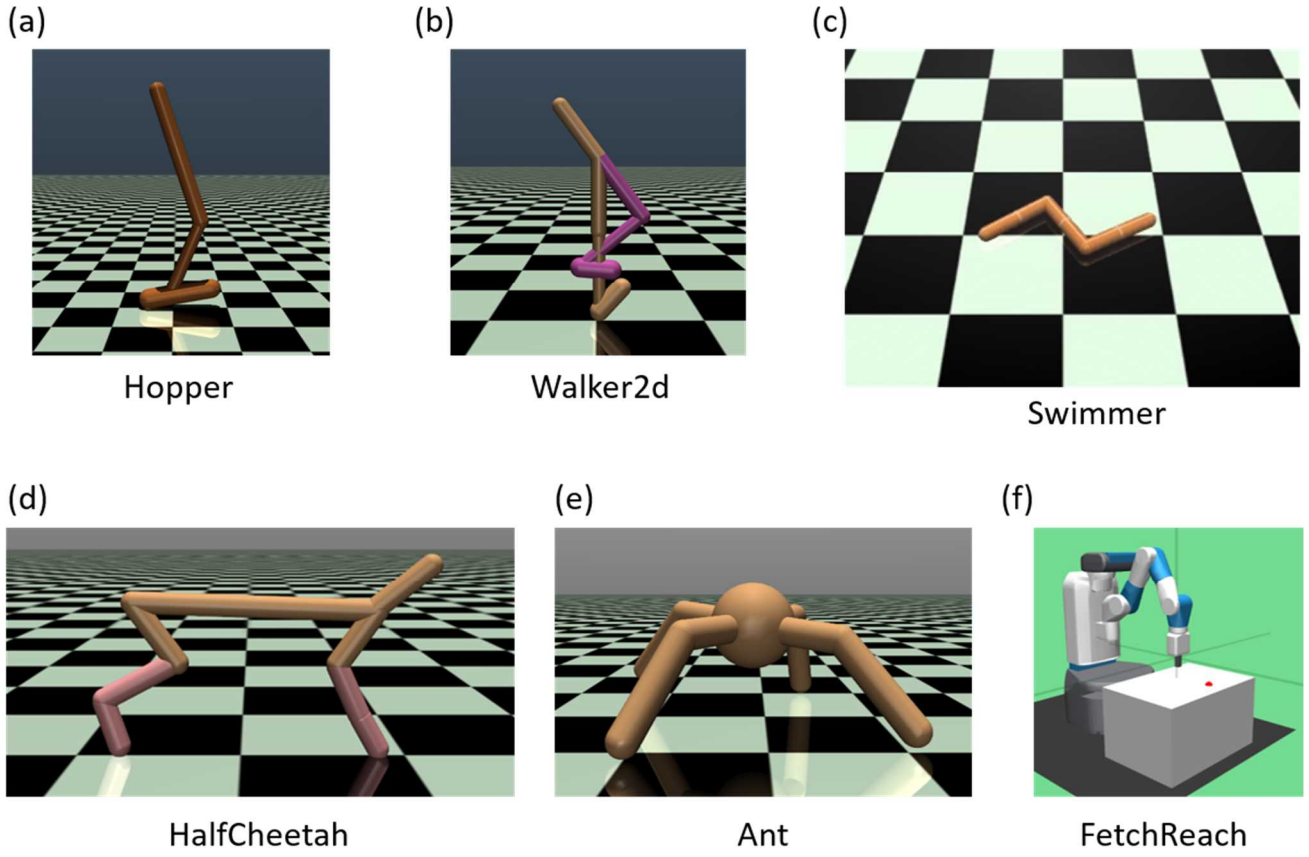


**図 6 OpenAI gym (Brockman et al., 2016)で提供されているベンチマーク課題。(a) Hopper、(b) Walker2d、(c) Swimmer、(d) HalfCheetah、(e) Ant、(f) FetchReach**

図 7 に MB-CRAIL と MB-CRAIL で使用している学習モジュールを単独で用いた場合の学習曲線を比較したものを示す。各手法は一つの課題につき 10 回実施した。見やすさのために図 7 は平均のみを表示している。Ant、 Hopper では MB-CRAIL は各学習モジュールを単独で用いるよりも学習の中期段階から効率よく学習している。一方、HalfCheetah、FetchReach では学習の後期段階では MB-CRAIL は他の手法と比べて性能の良い方策を獲得した。Walker2d では MB-CRAIL による性能改善はあまり得られなかった。モデルフリー法だけを用いたオリジナルの CRAIL と比較すると MB-CRAIL は学習の初期段階でのサンプル効率が改善されていないこともわかった。

図 8 は 1 回の実験でタスク Ant と FetchReach において、MB-CRAIL が学習途中と最終時点でどのようにモジュールを選択したかを示している。Ant では学習途中はモデルベース法である PILCO が選ばれる確率が高いのに対し、最終的にはモデルフリー法である SAC が選ばれる確率が高くなっている。理由として Ant では脚と床との接触がうまくモデル化できず、最終的にはサンプルだけから学習したほうが良かったためと考えられる。一方、FetchReach では PILCO から SVG へと使用されるモジュールは学習段階で変化しているがどちらもモデルベース法である。FetchReach は Ant と異なり手先を目標位置に動かすといった物理的な接触を伴わない課題であるためモデルを学習しやすく、モデルベースだけで学習が充分であったためと考えられる。

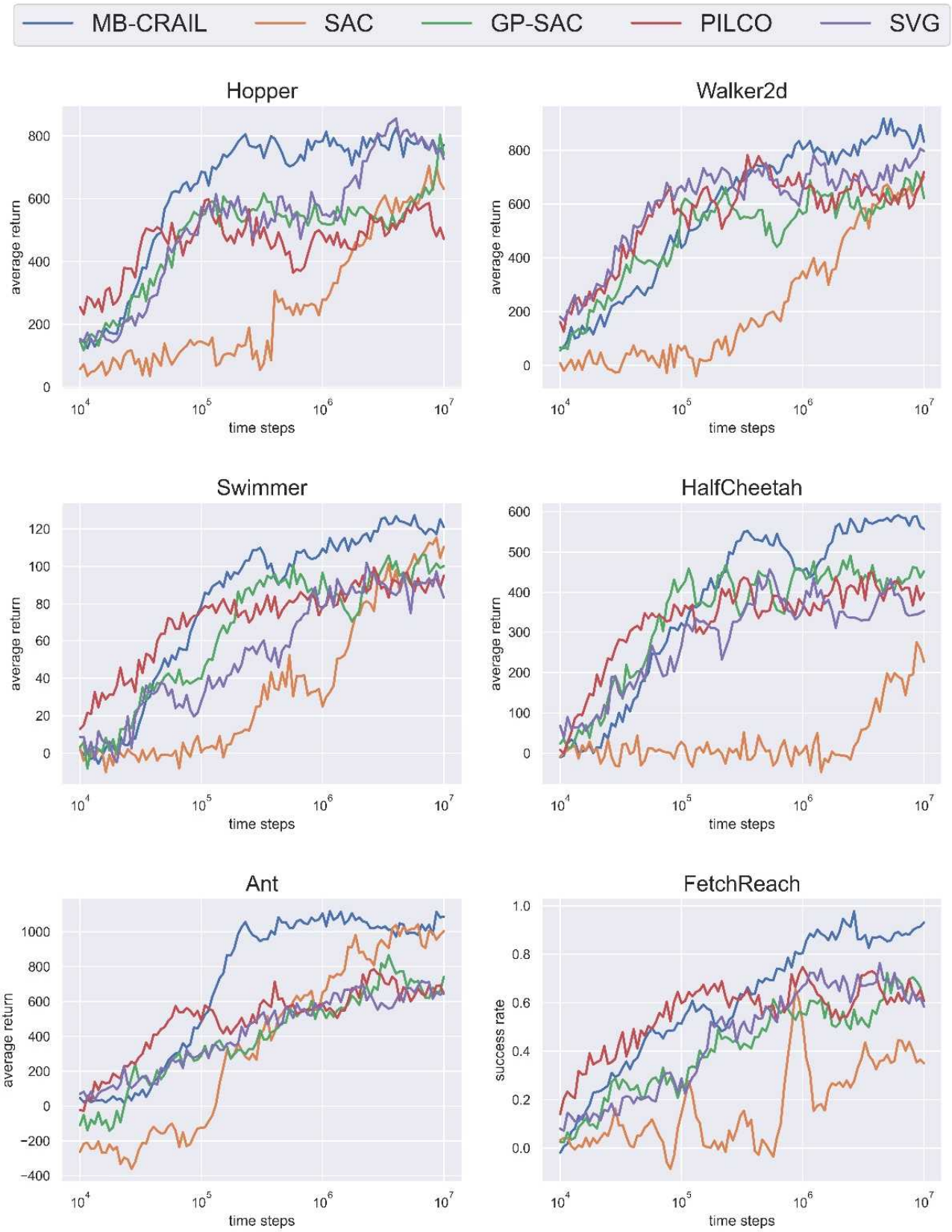


図 7 学習曲線の比較。各グラフにおいて MB-CRAIL は提案手法、SAC、 GP-SAC、 PILCO、 SVG は MB-CRAIL で使用している学習モジュールを単独で利用した場合の結果を示している。

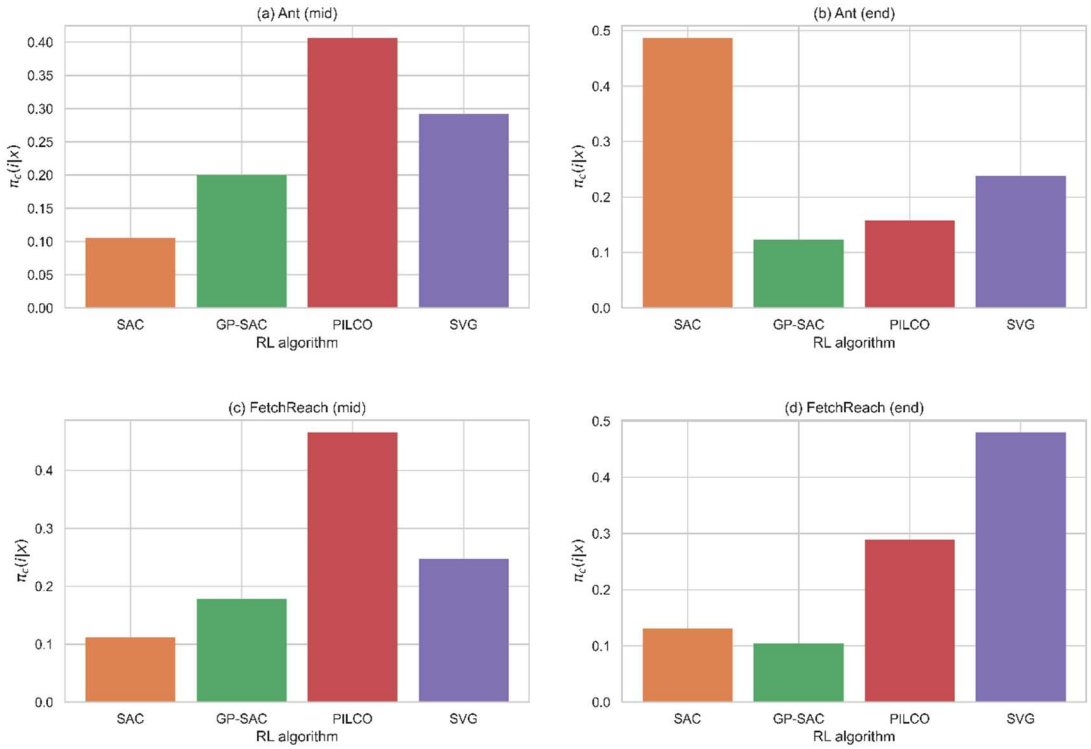


**図 8 学習途中と最終時点でのモジュール選択確率$\alpha(i \mid x)$。(a) Ant の学習途中、(b) Ant の最終時点、(c) FetchReach の学習途中、(d) FetchReach の最終時点。**

今回の実験では経験再生バッファから学習する SAC とガウシアンプロセスでモデルを学習し、モデルから生成されたサンプルから学習する GP-SAC では、GP-SAC がすべての場合において SAC の性能を超えていたことがわかる。ただしモデルフリー法だけを用いたオリジナルの CRAIL の実験結果と異なり、MB-CRAIL は学習の初期段階でのサンプル効率はあまり改善されなかった。オリジナル CRAIL では模倣に関する損失関数が重要であったのに対し、MB-CRAIL では各モデルの学習は完全に独立しており、ほかの学習モジュールを使って収集されたサンプルをうまく利用できていない問題がある。また、SVG のオリジナル論文では方策オフのための重点サンプリングの効果はあまり見られなかったが、MB-CRAIL では重点サンプリングの補正項が重要である可能性がある。

### 3.4.2 モデルフリーとモデルベース強化学習のための非同期並列学習

3.4.1節で述べたMB-CRAILは各学習器が意思決定に要する計算時間を考慮していなかったため、制御周期の短い単純なモデルフリー強化学習器を用いる利点を十分に発揮できていなかった。そこで各学習器の制御周期の違いを考慮した非同期並列強化学習法Asynchronous CRAIL (内部 2021a)を開発した。主要な貢献は各学習器で収集した経験を保存するバッファの分離と、制御周期の違いを吸収する経験再生バッファの変換である。これにより制御周期の異なる方策で収集されたデータをマージさせて学習することが可能になる。

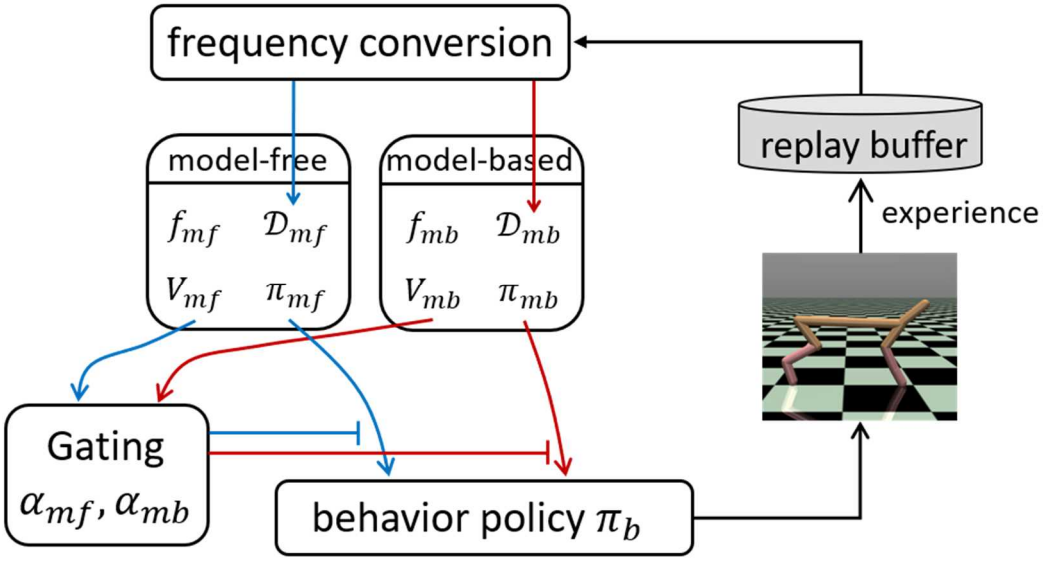


**図 9 モデルフリーとモデルベース強化学習モジュールを一つずつ持つ場合の非同期型 MB-CRAIL のアーキテクチャの例**

ここでも標準的なMDPを仮定する。図 9にモデルフリー(mf)強化学習モジュールとモデルベース(mb)強化学習を一つずつ持つ場合の提案手法のアーキテクチャを示す。たとえばモデルフリー強化学習モジュールは制御周波数$f_{mf}$で実行され、方策$\pi_{mf}(u \mid x)$、状態価値関数$V_{mf}(x)$を持つ。また方策や価値関数を更新するために用いられる経験（学習データ）を蓄える経験再生バッファ$\mathcal{D}_{mf}$を持つ。同様にモデルベース強化学習モジュールも方策$\pi_{mb}(u \mid x)$、状態価値関数$V_{mb}(x)$、経験再生バッファ$\mathcal{D}_{mb}$を持ち、制御周波数$f_{mb}$で実行される。

またエージェントは各モジュールの混合方策によって表現される挙動方策は3.4.1節と同様である。学習エージェントは挙動方策によって収集された経験$(x, u, x', r)$をいったんグローバルの経験再生バッファ$\mathcal{D}$に蓄える。ただし各モジュールの制御周期が異なるため、それぞれの方策で収集した経験を単純にマージすることはできない。そのため信号処理の分野で用いられるサンプリング周波数変換を用いる。図 10にサンプリング周波数変換を用いた経験の変換の例を示す。これはインターポレーションとデシメーションを組み合わせた標準的なもので、任意の周波数比のサンプリング周波数変換が可能である。ローパスフィルタ（LPF）はエリアジングひずみ発生を避けるためのもので、そのカットオフ周波数は入出力の低い方の周波数によって決定される。この枠組みで想定するモデルフリー強化学習は経験再生バッファから一様分布によって経験を直接サンプルし、方策オフ型モデルフリー強化学習を適用する方法である。ここでは代表的なアルゴリズムであるSACを用いる。モデルベース法は1ステップ予測モデルを組み合わせたものとマルチステップ予測モデルを用いたものを実装した。詳細は3.5.2節に示す。

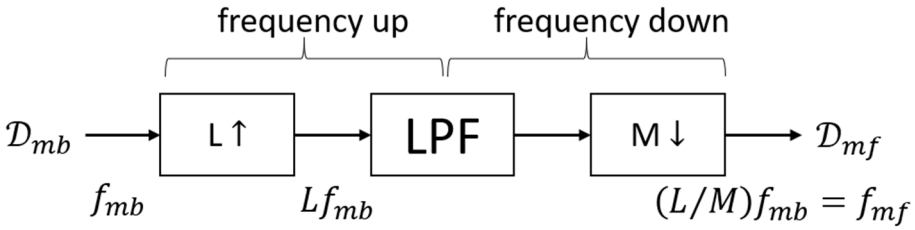


**図 10 ローパスフィルタ（LPF）を持つサンプリング周波数変換による経験の変換**

開発手法の有効性を検証するために、**図 6**に示すHalf-Cheetah、Ant、 FetchReachという3種類の

ロボット制御課題に適用する。提案手法Asynchronous CRAIL (Async-CRAIL)はSACと二種類のモデルベース強化学習の合計3種類のモジュールを持つ。予測するステップ数は$n = 5$とし、ロールアウト数は$K = 10$とした。SACの制御周波数を$f_{SAC}$としたとき、1ステップ予測モデルを用いるモデルベース強化学習の制御周波数は$f_{1-step} = f_{SAC}/2$、マルチステップ予測モデルを用いるモデルベース強化学習の制御周波数は$f_{multi-step} = f_{SAC}/5$とした。

開発手法の有効性を検証するために、制御周波数をマルチステップ予測モデルのものにすべて統一した同期並列強化学習法Synchronous CRAIL (Sync-CRAIL)とSAC単独で用いた場合の学習曲線を比較した。図 11に結果を示す。HalfCheetahとAntでは平均利得、FetchReachではタスク成功確率を制御性能として採用した。各手法は一つの課題につき10回実施した。3つの課題すべてにおいてAsync-CRAILはSync-CRAILやSACや1ステップ、nステップ予測モデルを単独で用いるよりも効率よく学習できたことが確認された。またAsync-CRAILにおいて周波数変換を用いない場合は学習性能が大幅に劣化することも確認できた。

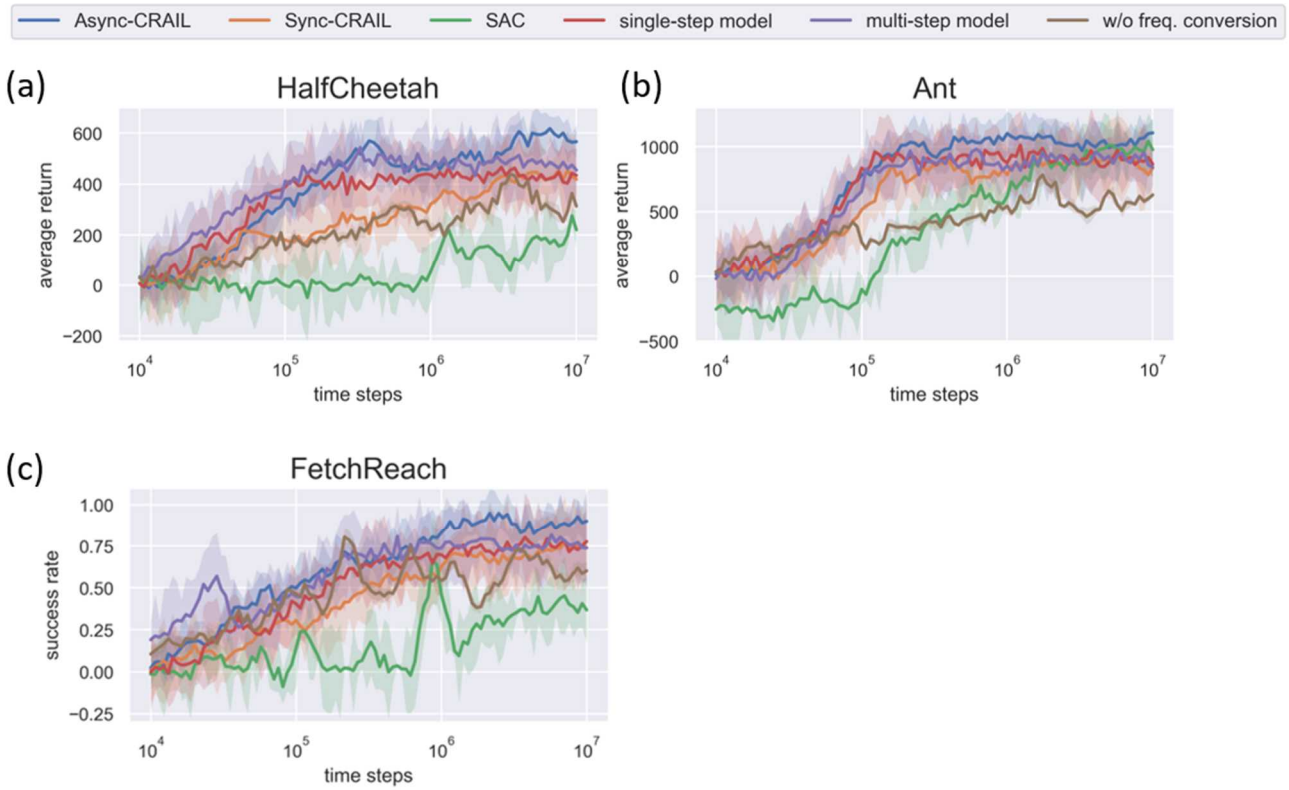


**図 11 学習曲線の比較。各グラフにおいて Async-CRAIL は 3.4.2 節、Sync-CRAIL は 3.4.1 節で述べた方法の結果である。SAC はモデルフリー強化学習 SAC を単独で利用した場合、single-step は 1 ステップ予測モデルを使った場合、multi-step は n ステップ予測モデルを使った場合のモデルベース強化学習の結果を示している。w/o freq. conversion は周波数変換を用いない Async-CRAIL の結果である。**

図 12はFetchReachにおいて、1回の実験でAsync-CRAILとSync-CRAILが学習初期(early)、途中(middle)と最終段階(end)でどのようにモジュールを選択したかを示している。学習の初期段階ではモデルベースとモデルフリーに性能差はないため、Async-CRAILは制御周期の早いSACが選ばれる。学習途中ではマルチステップ予測モデルを使ったモデルベース強化学習を使うことで学習効率を改善している。最終段階では再びSACが選ばれる結果となった。このことはモデル化誤差のため、モデルベースの漸近性能はモデルフリーに劣るという結果と合致する。一方Sync-CRAILでは学習初期でSACが利用されるのは同じであるが、学習途中では1ステップ予測モデルを用いたモデルベース強化学習、最終的にも同じモデルベース強化学習が選ばれる確率が高かった。これ

はSACの制御周期が最も遅いマルチステップ予測モデルを用いた強化学習に合わせているため、SACの性能を引き出せていないという従来の結果(内部2020)と一致する。

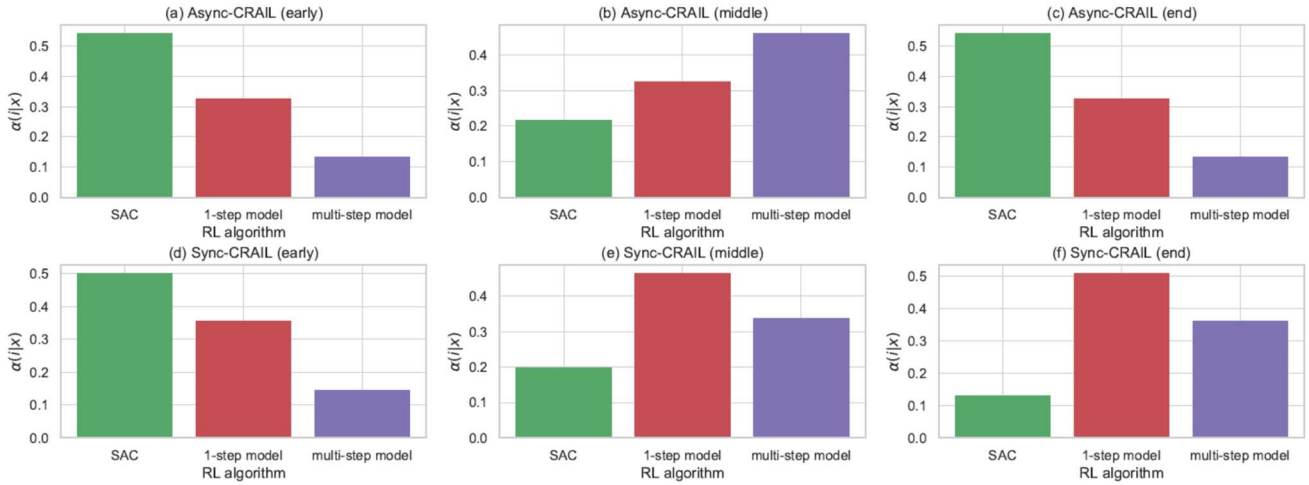


**図 12 FetchReach 課題における学習初期(early)、学習途中(middle)、最終段階(end)でのモジュール選択確率$\alpha(i \mid x)$の比較。(a)-(c)はそれぞれ Async-CRAIL、(d)-(f)は Sync-CRAIL の選択確率である**

3.4.3 学習器の切り替え条件に関する考察

3.4.1, 3.4.2節で述べたCRAILは学習器の性能を表すものとして状態価値関数を使用していた。しかし価値関数以外にもモデルから計算される状態予測誤差に基づく選択、価値関数から計算される報酬予測誤差に基づく選択，それらの重みづけによる方法など様々なものが考えられる。また切り替え条件を事前に規定する以外に、価値関数または方策を合成することで設定することも考えられる。さらにモデルベースとモデルフリーの切り替え条件そのものを学習することは検討していなかった。そこで本節では様々な切り替え条件を比較検討し、学習効率や環境の変化に対する追従能力，さらに神経科学の観点から考察する。

報酬予測誤差に基づく方法では、状態価値関数の代わりに価値関数の時間差分の符号を反転したTemporal Difference (TD)誤差$-\delta^2_{\mathrm{MF}}(x), -\delta^2_{\mathrm{MB}}(x)$を用いる。このとき報酬予測誤差の二乗が最も小さいものを学習が進んでいるとしてモデルフリーとモデルベースを選択していると解釈できる。以降、この方法をRPEと呼ぶことにする．状態予測誤差に基づく方法ではモデルベース法から計算される識別器の値を用いて

$$\alpha(\mathrm{MB} \mid x) = \frac{1}{1 + \exp\left(-\beta \epsilon^2_{\mathrm{MB}}(x)\right)}$$

と定義する。ここで$\epsilon_{\mathrm{MB}}(x)$は状態$x$でのモデルの誤差、$\beta$はハイパーパラメータである。これは識別器の出力が0.5から逸脱するほどモデルの予測精度が悪いと解釈して，モデルベースの選択確率を下げる方法である。他の基準と異なり、モデルフリーの情報は選択確率の計算に寄与しない。以降、この方法をSPEと呼ぶことにする。学習に基づく方法はモデルフリー、モデルベースそれぞれに対して選好度

$$f_{\mathrm{MF}}(x) = w_1 V_{MF}(x) + w_2 \delta^2_{\mathrm{MF}}(x)$$
$$f_{\mathrm{MB}}(x) = w_3 V_{\mathrm{MB}}(x) + w_4 \delta^2_{MB}(x) + w_5 \epsilon^2_{\mathrm{MB}}(x)$$

を計算し、それをもとに確率的学習器を選択する。重み$w_1, \ldots, w_5$はREINFORCEアルゴリズムで学習する。

提案手法の有効性をOpenAI gymで提供されている7自由度のマニピュレータFetchを用いたFetchReach、FetchSlide、FetchPickAndPlaceという3種類のベンチマーク制御課題を用いて検証する（図 13参照）。FetchReachはこの中で最も簡単な課題で、手先を目標位置に移動させることが目的である。手先が目標位置に到達した時に非零の報酬が得られる。FetchSlideはテーブルの上に置かれた箱を押して目標位置まで移動させることが目的である。ただしテーブルは滑りやすくなっ

ており、目標位置も手先の届かない範囲に設定される。そのためロボットは箱を適切にヒットさせる必要がある。FetchPickAndPlaceは箱を把持し、それを目標位置まで移動させる課題で、目標位置はテーブルの上だけでなく手先が到達可能な任意の位置にランダムに設定される。

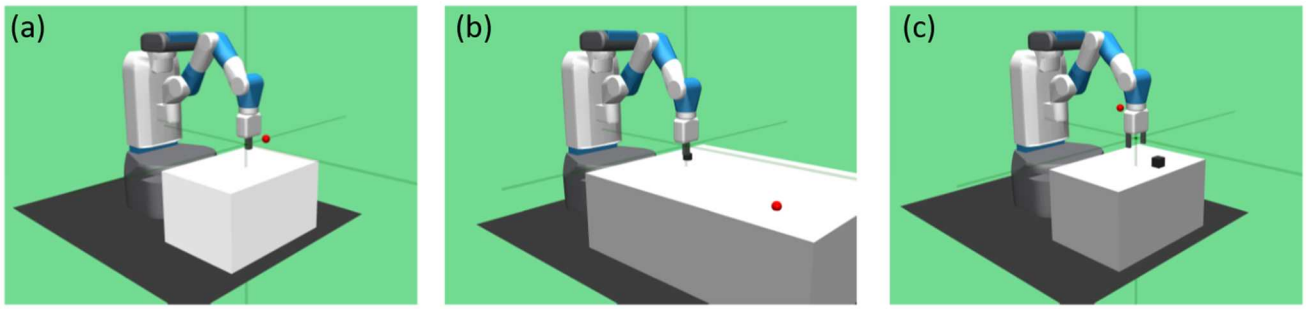


**図 13 ベンチマーク課題。(a) FetchReach、(b) FetchSlide、(c) FetchPickAndPlace**

図 14(a)は最も簡単なFetchReach課題でのモジュール選択確率の比較である。valueとlearningは学習が進むにつれてモデルフリーを選択する確率が増加するが、SPEは逆にモデルベースを選択する確率が増加する傾向にある。FetReachはDDPG単独でも容易に学習可能な課題であるため、制御周期の遅いモデルベースを使う必要がなくモデルフリーがvalueとlearningでは支配的になったが，SPEはモデルも素早く学習できるためモデルベースが支配的になったと考えられる。一方でRPEは学習が進んでもモジュール選択確率に変化はあまり見られなかった。

図 14(b)はFetchSlideの結果で、FetchReachよりも複雑な課題である。この課題でも学習が進むにつれvalueではモデルフリーの選択確率が増加する。一方でSPEの場合も同様の傾向があり、これはFetchReachとは異なる結果となっている。この原因として、箱が滑りながら移動する部分のダイナミクスの学習がFetchReachの手先位置だけのダイナミクスの学習よりも困難で、モデルベースの選択確率が大きくならなかったためである。RPE、 learningでは選択確率に関して顕著な傾向はみられなかった。

図 14(c)はFetchPickAndPlaceの結果である。このタスクではvalueは学習中期ではモデルベースを、学習後期ではモデルフリーを選ぶ傾向があり、これは従来結果（内部2021a）と一致する。RPE, SPEについてはFetchReachと同じ傾向である。learningは学習後期ではモデルフリーの選択確率が増加した。

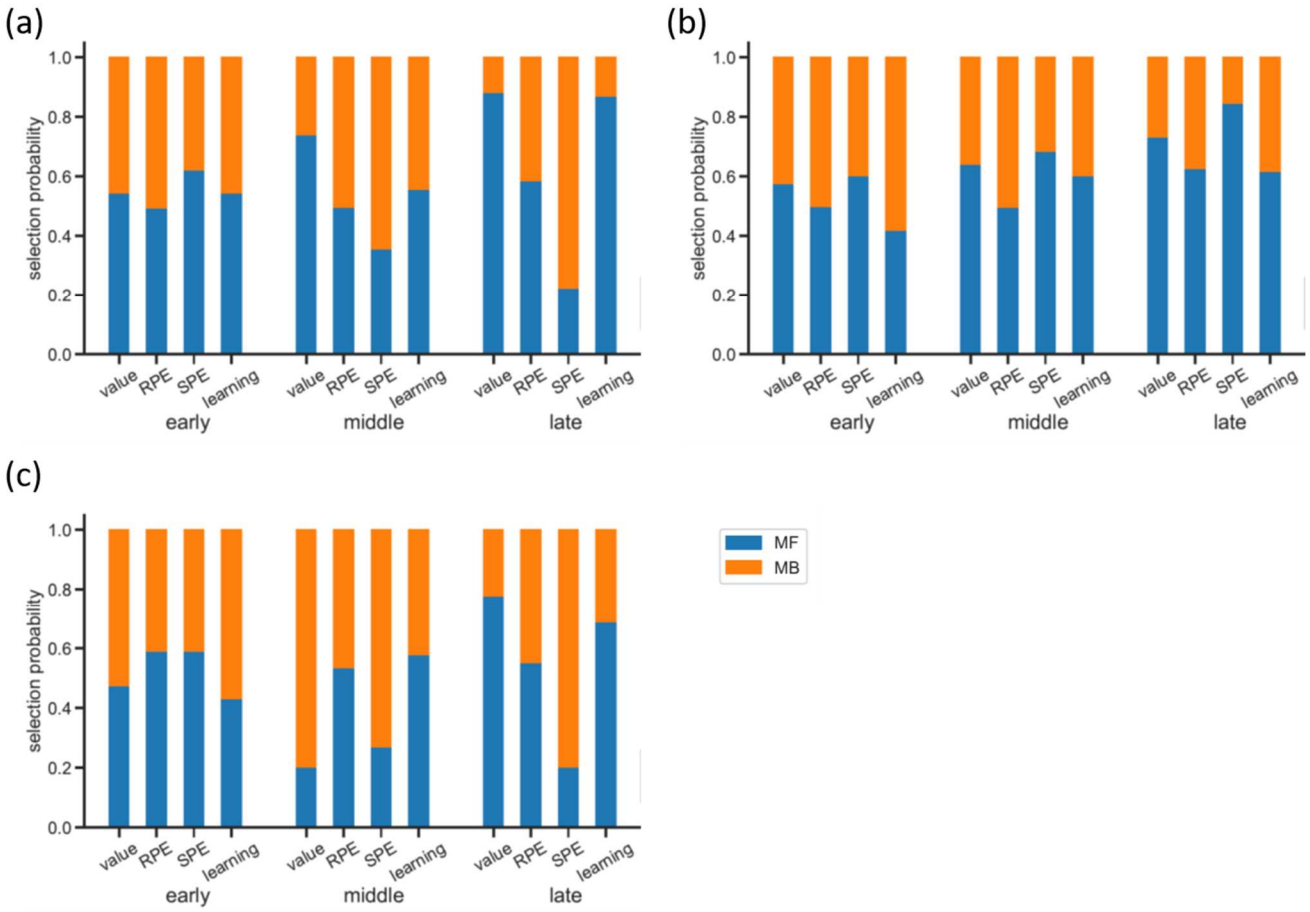


図 14 モデルフリー(MF)とモデルベース(MB)の選択確率の推移

3.4.4 アルゴリズム検証のための実ロボット実験環境の構築

移動ロボットTurtleBot3 Waffle PIを用いた実験環境を構築した。図 15にシミュレーション環境と実環境を示す。どちらもRobot Operating System (ROS)を用いて構築されており、同じソースコードを用いてシミュレーションと実ロボットの実験が可能である。このシステムを用いて、報酬の正負を分離して学習するMaxPainアーキテクチャを実装した(Wang et al., 2021)。

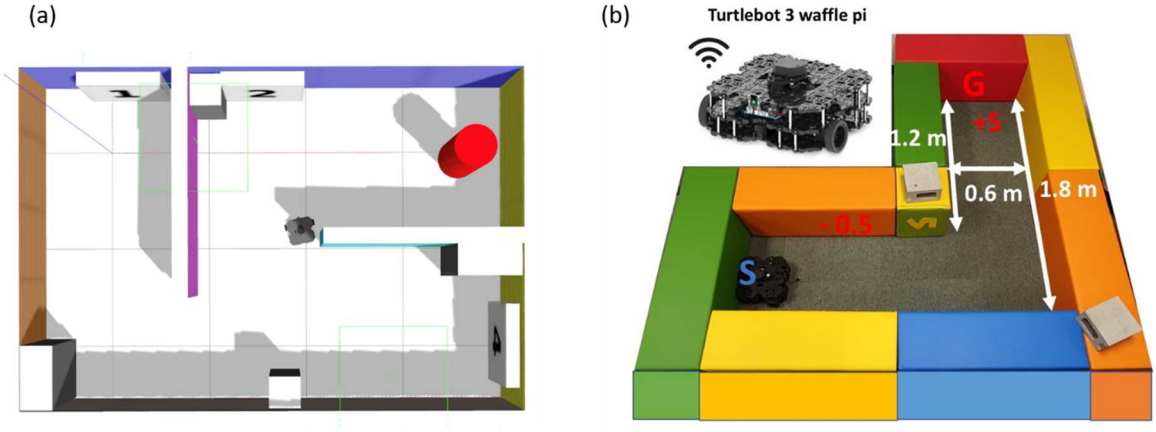


図 15 移動ロボット Turtle3 Waffle Pi の実験環境

また双腕ロボットを用いたマニピュレーション実験のために、カワダロボティクス社製Nextageを用いた実験環境を構築した（図 16）。ロボットには頭部に二つ、各手先に一つずつの合計4つのRGBカメラを持ち、それらを同時に入力して使用できる。

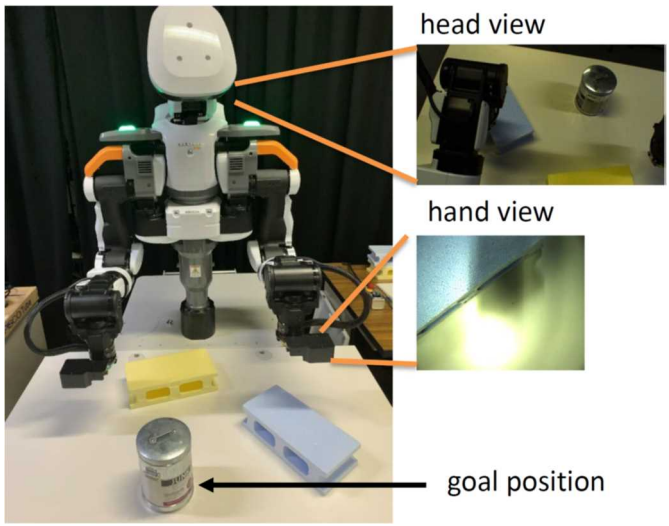


**図 16 双腕ロボット Nextage の実験環境**

### 3.5（1-5）モデル学習アルゴリズムの開発

環境のモデル学習法として、パラメトリック・ノンパラメトリックそれぞれのアルゴリズムを定式化した。

#### 3.5.1 パラメトリックなモデル学習法

世界モデル、すなわち状態遷移確率$q(x' \mid x,u)$を推定する最も簡単な方法は最尤推定などを用いて

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,u,x')\sim\mathcal{D}}[\ln q(x' \mid x,u;\theta)]$$

を最小にするように学習することである。$\theta$は分布$q$のパラメータである。ただしモデルの学習と順・逆強化学習を独立に学習するとモデルバイアスの影響により、順強化学習の漸近性能がモデルフリーの場合よりも劣ってしまう問題がある。

そこで順強化学習に適したモデル学習法を開発した。3.2節で述べた重点サンプリングを用いた補正を行った場合の順強化学習の目的関数は

$$\max_{\pi^L} \mathbb{E}_{(x,u,x')\sim q}\left[ISW\left(D^{(3)}\right)\tilde{r}(x,u,x')\right]$$

と表される。$\tilde{r}$は逆強化学習で得られた報酬と状態価値関数から計算されるTD誤差とする。上式は不偏推定量であるが、分散を最小にする最適なモデル$q$を以下のように構築できる。

$$q^*(x' \mid x,u) \propto |\tilde{r}(x,u,x')| p_e(x' \mid x,u)$$

しかし未知である真の状態遷移悪率$p_e$に依存しているため、最適なモデルは実際には計算不可能である。そこで両者のKLダイバージェンスを最小にするように$q$を推定する。KLダイバージェンスの勾配は

$$\nabla \mathrm{KL}(q^* \parallel q) \propto -\mathbb{E}_{p_e}[\tilde{r}(x' \mid x,u)\nabla \ln q(x' \mid x,u;\theta)]$$

と与えられる。これは通常の最尤推定法と比べると、推定された逆強化学習の結果で重みづけされたものになっており、TD誤差の大きい部分を重点的に学習するものとなっている。

開発したモデル学習法は3.2節で述べた方法と統合して用いた。モデルの実装は多変量正規分布$\mathcal{N}(\mu,\Sigma)$に従うと仮定し、平均$\mu$と共分散行列$\Sigma$の対角成分を明示的にニューラルネットワークで表現した。

#### 3.5.2 マルチステップ予測のためのノンパラメトリックなモデル学習法

モデルベース強化学習法で重要なのはロールアウトの生成方法であり、それに応じて学習すべきモデルの構造が決定される。**図 17**(a)に示す2種類のロールアウト方法を検討する。上の図は1ステップモデルを用いた方法で、通常の予測モデル$q(x' \mid x,u)$を$n$回用いて$n$ステップ先の予測を

$$\hat{x}_1 \sim q(\cdot \mid x_0,u_0), u_1 \sim \pi(\cdot \mid \hat{x}_1), \hat{x}_2 \sim q(\cdot \mid \hat{x}_1,u_1), \ldots, \hat{x}_n \sim q(\cdot \mid \hat{x}_{n-1},u_{n-1})$$

と計算する。これはモデルベース強化学習を用いたプランニングでは標準的に使われる方法である。モデル学習法としてGPを用いる。これは状態行動対$(x,u)$を入力、状態遷移の差分$\Delta x = x' - x$を出力とする確率モデルである。一方、報酬関数は決定論的関数として推定する。この方法の問題は予測を多段に繰り返すことであり、小さな予測誤差も複合的に発生し、長期予測を不正確なものにしてしまう。

**図 17**(b)はステップ数に応じて異なるモデルを用いるマルチステップ予測モデル(Asadi et al., 2018)を用いる方法である。これは$n$ステップ先の状態の予測は固有の予測モデルを用いて

$$\hat{x}_n \sim q^n(\cdot \mid x_0, u_0, u_1, \ldots, u_{n-1})$$

と計算する。ここで$q^1 = q$である。マルチステップ予測モデルは1ステップ予測モデルよりも複雑で、一般に学習に必要なデータ数はステップ数に応じて指数的に増大する。しかし、1ステップ予測モデルを再帰的に用いるよりも長期予測が正確であることが指摘されている(Asadi et al., 2018)。

どちらの方法であってもhorizon time、つまり先読みするステップ数$n$とロールアウトする回数$K$によって計算コストは変化し、$n$が大きいほど、$K$が大きいほど意思決定に要する時間は長くなる。また実装上、マルチステップ予測の方が学習に要する計算コストは高く、必要となるデータ数も多い。

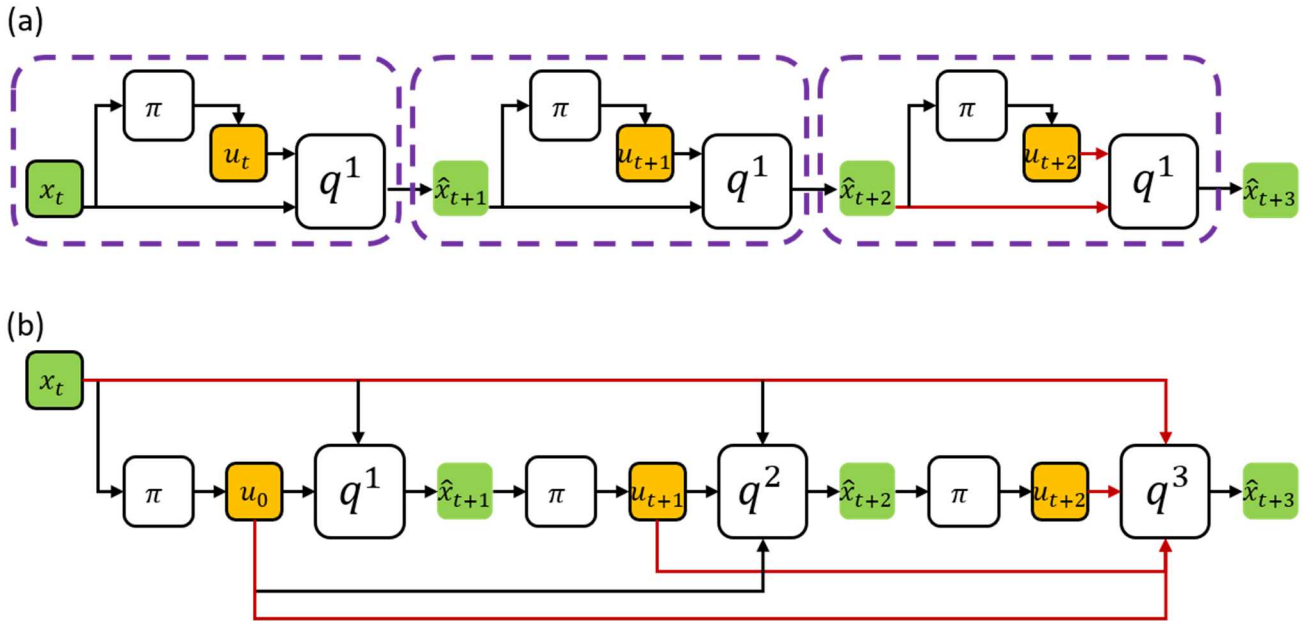


**図 17 環境モデルを用いたロールアウト。(a) 1ステップ予測モデルを用いた3ステップロールアウト。(b) マルチステップ予測モデルを用いた3ステップロールアウト。図はAsadi et al. (2018)から改変。**

図 18にAnt課題において、1ステップ法単独で用いた場合、3ステップ法単独で用いた場合、および3.4.2節で述べた非同期学習法と組み合わせて用いた場合のモデルの性能をRMSE(Root Mean Squared Error)によって評価した。提案手法は予測長が増大しても予測性能が他と比べて劣化しないことが確認できた。

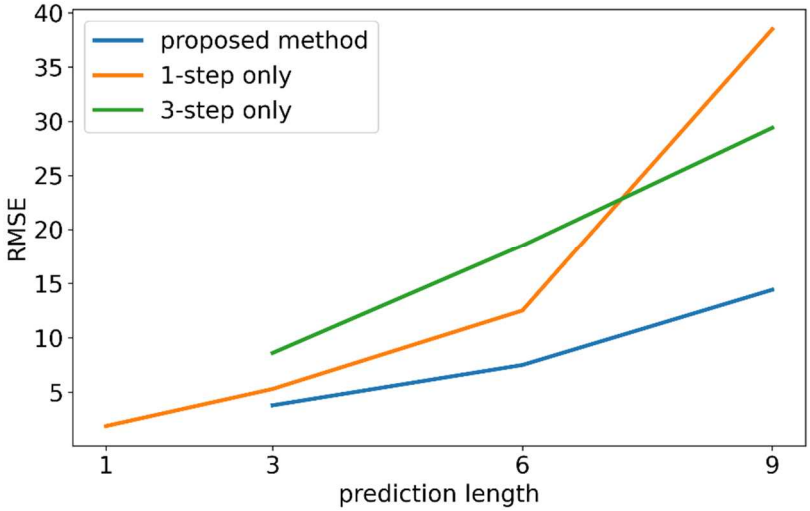


**図 18 Ant 課題におけるモデル比較の結果。**

### 3.6 (2-1) ニューロフィードバックで誘導の目標とする脳活動の定義

脳活動の空間パターンと疾患状態の因果的関係の解析を行うために、まずは過去のPTSD患者の脳活動データを整理するとともに、新規データ取得継続の準備を実施した。次にPTSD患者の脳活動の収集・解析を実施し、ニューロフィードバックで誘導の目標とする脳活動の定義について検討を進めた。今年度までに、64例のPTSD患者の安静時脳活動を収集に成功した。また、解析方法の開発の一環として機械学習アルゴリズムの実装に着手した。今年度は予備的患者40例と健常者108例の脳活動を用い、脳活動から患者・健常者を84%の精度で判別することに成功した。判別には、左前頭葉、右淡蒼球、右海馬といった、PTSDの中心的病態である情動制御と関連する脳部位の活動が関与することが分かった（図 19参照）。

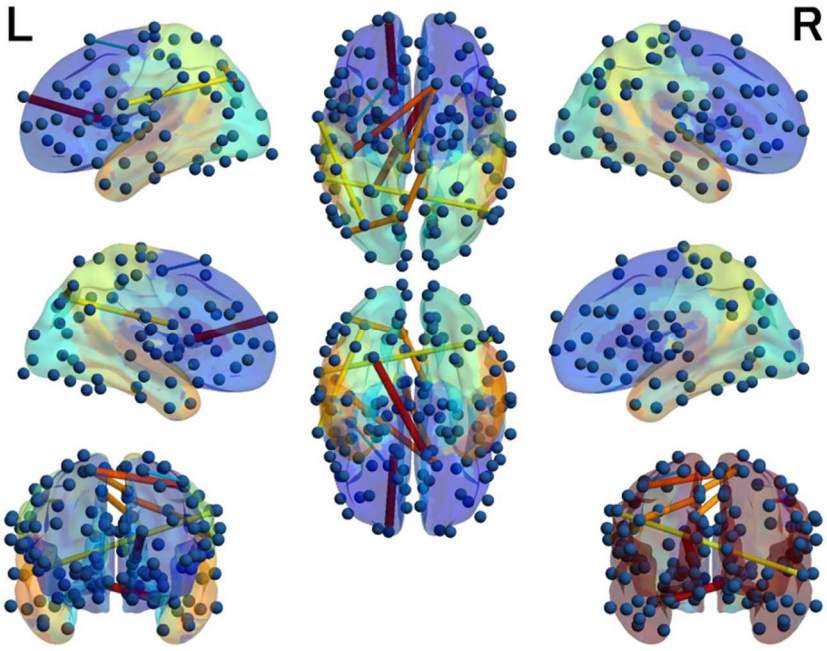


**図 19 PTSD を特徴づける脳活動：健常者との判別に寄与する脳領域**
**各点は解析に用いられた関心領域の座標を示し、各点間の、棒は群間差を認めたことを意味する。棒が太いほど群間差は大きい。**

また、ストレス体験が脳活動に及ぼす影響の解明のため、自衛隊のレンジャー訓練開始前に56

例から脳活動を収集した。このうち、訓練を完遂した37例から訓練直後に脳活動を収集した。まず、ストレス曝露により変化する脳活動パターンの同定のため、訓練前と訓練後を判別する判別器を作成した。その結果、71%の精度を誇る判別器の作成に成功した。判別に寄与する脳領域は、腹内側前頭前野や背外側前頭前野といった、情動制御ネットワークに含まれる領域及び、運動野・補足運動野といった運動機能に関連する領域が中心であった（図 20参照）。

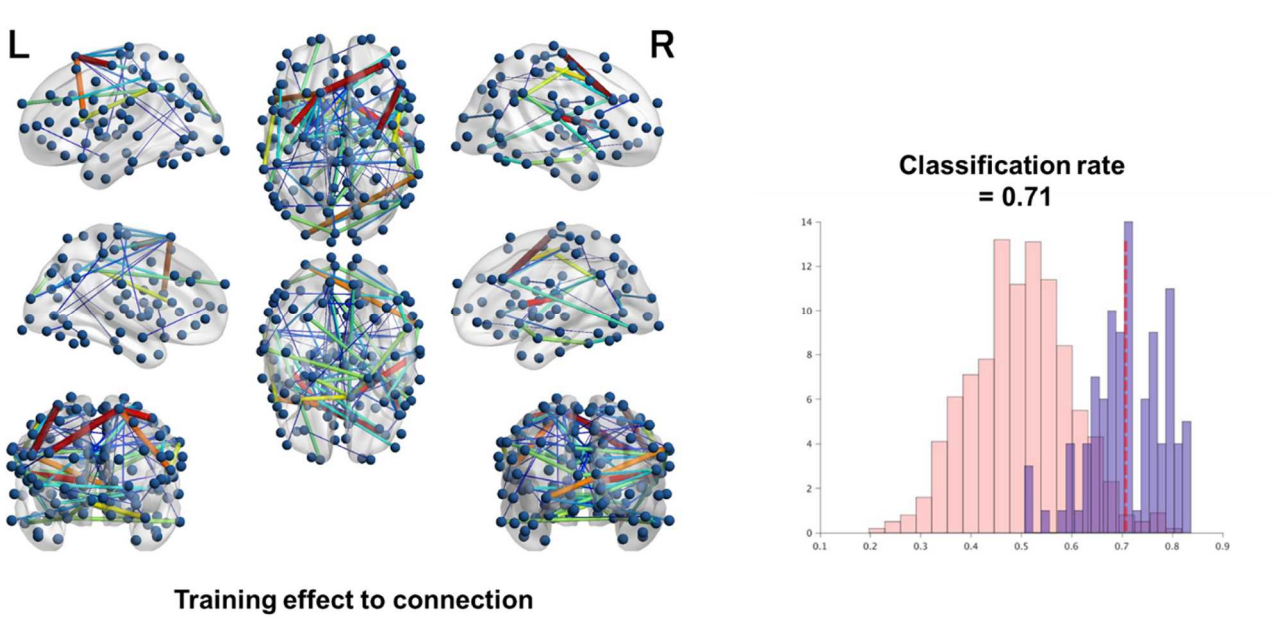


**図 20 ストレスによる影響を特徴づける脳活動：訓練前後の判別に寄与する脳領域**

次に、ストレス耐性と関連する脳領域を同定するため、訓練完遂群と脱落群を訓練前の脳活動から予測する判別器を作成した。その結果、61%の精度の判別器の作成に成功した。判別に寄与する脳領域は、背側前帯状皮質や眼窩前頭前野、背外側前頭前野といった情動制御ネットワークに含まれる領域が中心であった（図 21参照）。

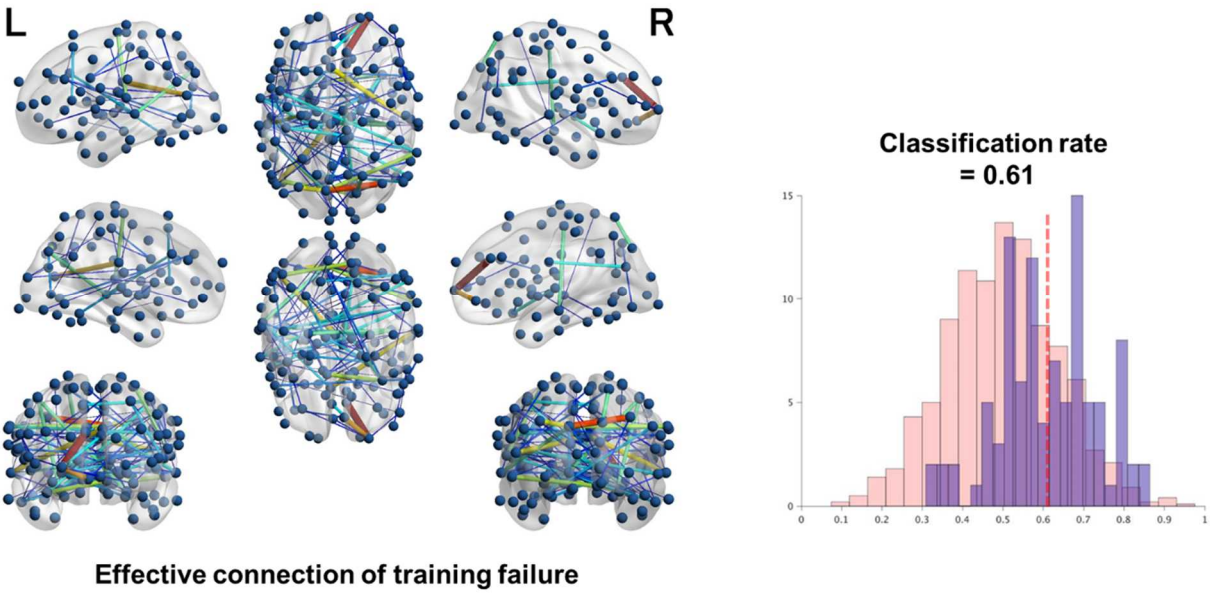


**図 21 ストレス抵抗性を特徴づける脳活動：訓練脱落の判別に寄与する脳領域**

以上、PTSD、ストレス曝露の影響、ストレス耐性と関連する脳活動の特徴を元とした判別器では、一貫して情動制御と関連する脳領域が選択されることがわかり、得られた結果の妥当性を示唆するものであるといえる。今後は、PTSD についてはデータの拡充を、レンジャーデータにつ

いては独立データを取得し判別器の性能を検証する。得られた知見は縦断データにおける PTSD 患者の脳活動ダイナミクスとの比較に供される。さらに、PTSD の DecNef の目標脳活動の決定、効果判定に用いられる。

### 3.7（2-2）ニューロフィードバックの実施

うつ症状軽減のための機能結合ニューロフィードバックによる治療効果

うつ度軽減ニューロフィードバックのプロトコル開発の達成度は 80%である。現在までに 3 種類のプロトコル（先行研究の手法、異なる時間窓、異なる教示方法）についてニューロフィードバックを実施した。図 22 左が実験前後での、うつの程度の自己評価尺度である BDI (Beck Depression Inventory)の変化で、緑色が以前に行った実験、藍色が今回行った実験、灰色が両者を合わせたもので、うつ症状が減少していることが確認できた。また、安静時の脳機能変化との相関も確認した。図 22 右が時間経過後のグラフで、赤色が 1 か月後、黒色が 2 か月後の結果で、ニューロフィードバックの長期効果も確認できた。今回は左背外側前頭前野と左楔前部の結合に関わる症状のみの改善を認めた。

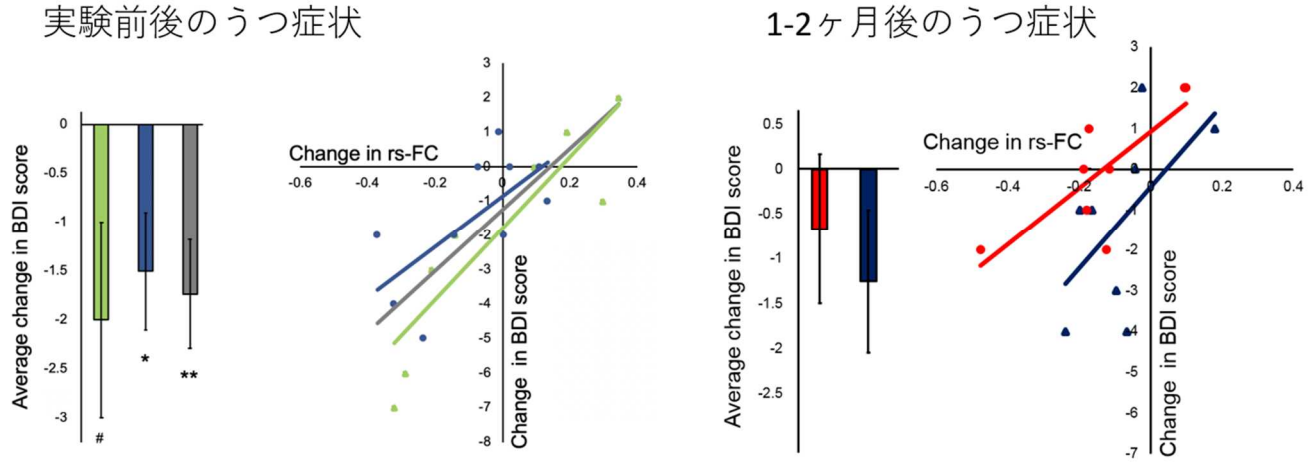


**図 22 うつ症状軽減のための機能結合ニューロフィードバックによる治療効果**

PTSD 患者を対象としたニューロフィードバックによる治療効果

先行研究のパラメータで実施したニューロフィードバックにおいて、効果、安全性、再現性を確認した。並行して、間欠実験の実施を進めた。PTSD患者6例を対象に、ニューロフィードバックを実施し、介入前後で顕著な症状改善を認めることを確認した。患者の現在の脳活動が怒り顔を提示したときと普通の顔を提示したときのどちらに近いかに応じて、目的関数に相当する報酬を患者にフィードバックする。そのとき患者は報酬を最大化するように脳活動を変容させる。この脳内処理が順強化学習に対応する。怒り顔を提示する実験条件6例では、全ての症状において介入によってPTSD重症度を軽減できることを確認した。さらに対照条件を4例に対し実施し、実験条件で得られた効果が単なるプラセボでない可能性が高いことを確認した。この効果は訓練実施後2か月後まで持続することを確認した。

### 3.8（3-1）脳活動データに潜在脳推定アルゴリズムを適用するための予備解析

縦断的なデータ解析によるPTSDの症状の長期的な変動の評価

縦断的なデータを解析し、PTSDの症状の周期性を同定した。PTSD患者12人からクリニック受診時の質問紙項目を取得し、以前我々が定義した（Chiba et al., Mol. Psy 2020）、情動制御状態の指標である症状インバランスの時系列データを抽出した。これにより、患者の情動制御状態は約2か月と約1年という複数のモードで周期的な変動を確認した（図 23）。

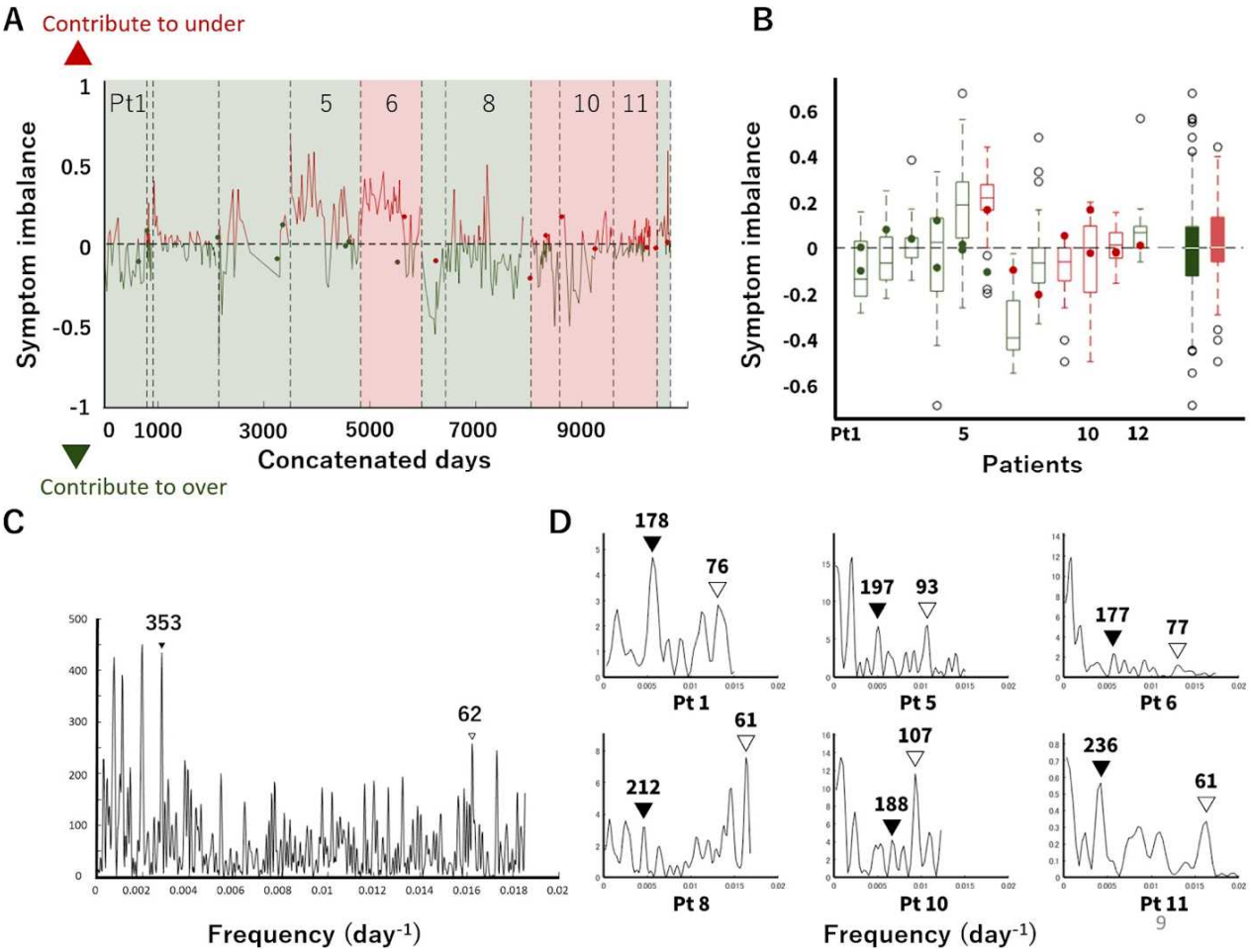


**図 23 患者の症状ダイナミクスに関する周波数解析**

大規模なオンライン調査を実施して PTSD の症状とマルチディメンジョナルな精神疾患症状・各種行動との関係性を明らかにした(Oka et al., 2021)。まず、厚労省の公表している本邦の自殺者数を interrupted time-series analyses により解析し、コロナ禍において自殺が顕著に増加傾向にあることを明らかにした（図 24 参照）。更に、この増加と関係する精神状態を解明するために、オンラインデータと比較解析した。その結果、年代・性で層別化した自殺の増加率は該当する群の PTSD 症状により、よく説明できることがわかった（図 25 参照）。PTSD 症状の自殺増加の説明力はうつ病や不安障害のスコアよりも大きく、PTSD 症状に着目することで自殺の高リスク群を抽出可能であることが判明した(Chiba et al., 2020)。

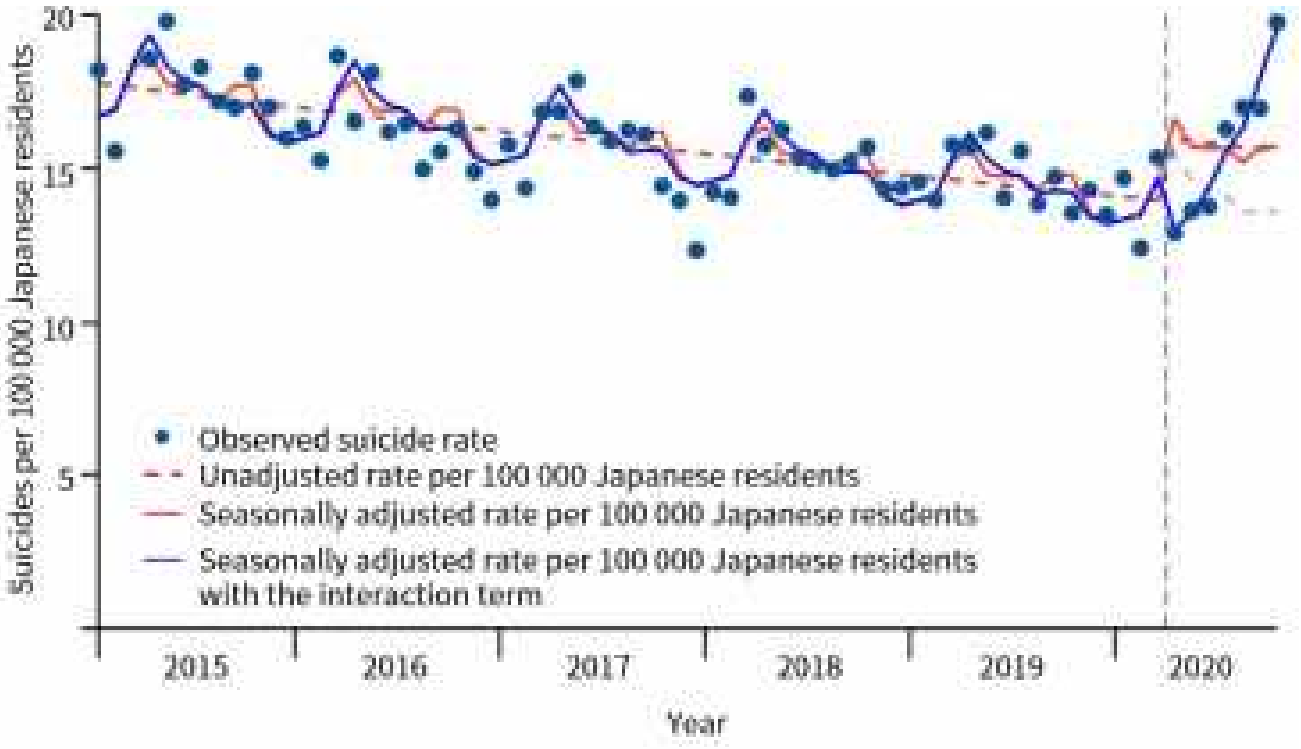


図 24 本邦におけるコロナ禍の自殺数増加

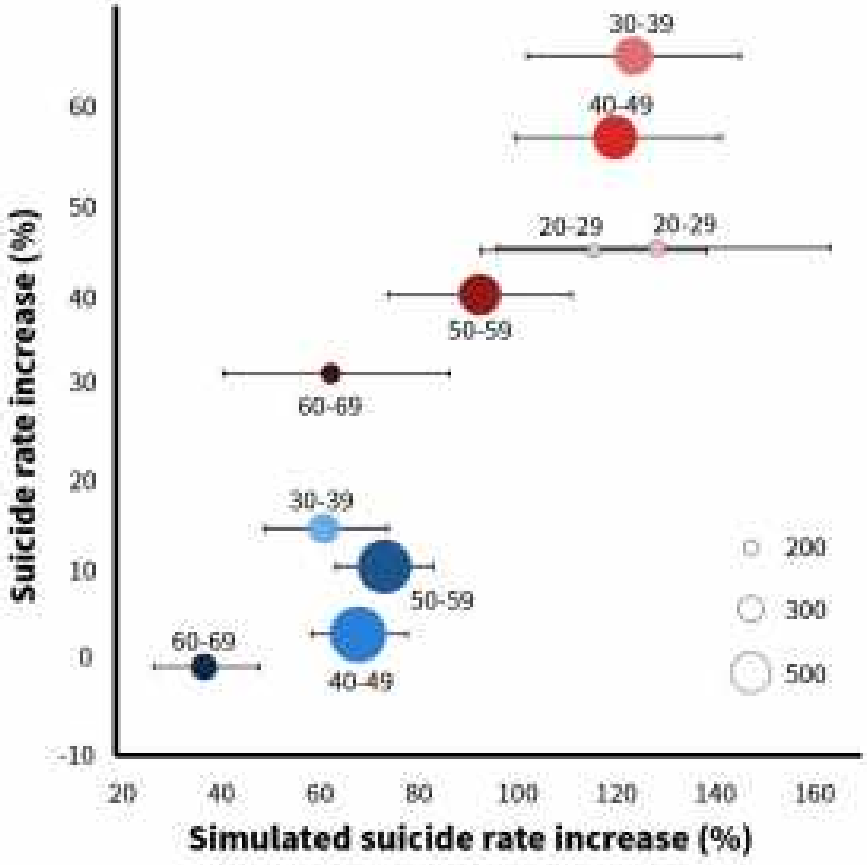


図 25 コロナ禍における自殺増加率と PTSD スコアとの関連

さらに、コロナ禍におけるストレスが精神症状に及ぼす影響のダイナミクスを評価するため、同集団から定期的にオンラインデータを取得した。その結果、種々の精神症状は様々な時系列推移のパターンを示すことを確認した（図 26 参照）。ただし、共変動する要素を主成分分析によ

り抽出すると、全般的な精神負荷を示す要素（PC1）、社交不安を示す要素（PC2）、アルコール関連の問題を示す要素（PC3）、うつと不安を示す要素（PC4）の要素により分散の 60%を説明可能であることが示された。PC1,2,4 はコロナ禍の比較的初期から急激な増加をきたし、PC1 はその状態を維持、することが分かった（図 27 参照）。また、PC４が観察機関の初期に、PC2 は後期にピークを迎えることがわかった。よって、初期には PC4 を構成する鬱や不安に、後期には社交不安に焦点をあてて予防策をとると効果的である可能性が示された。

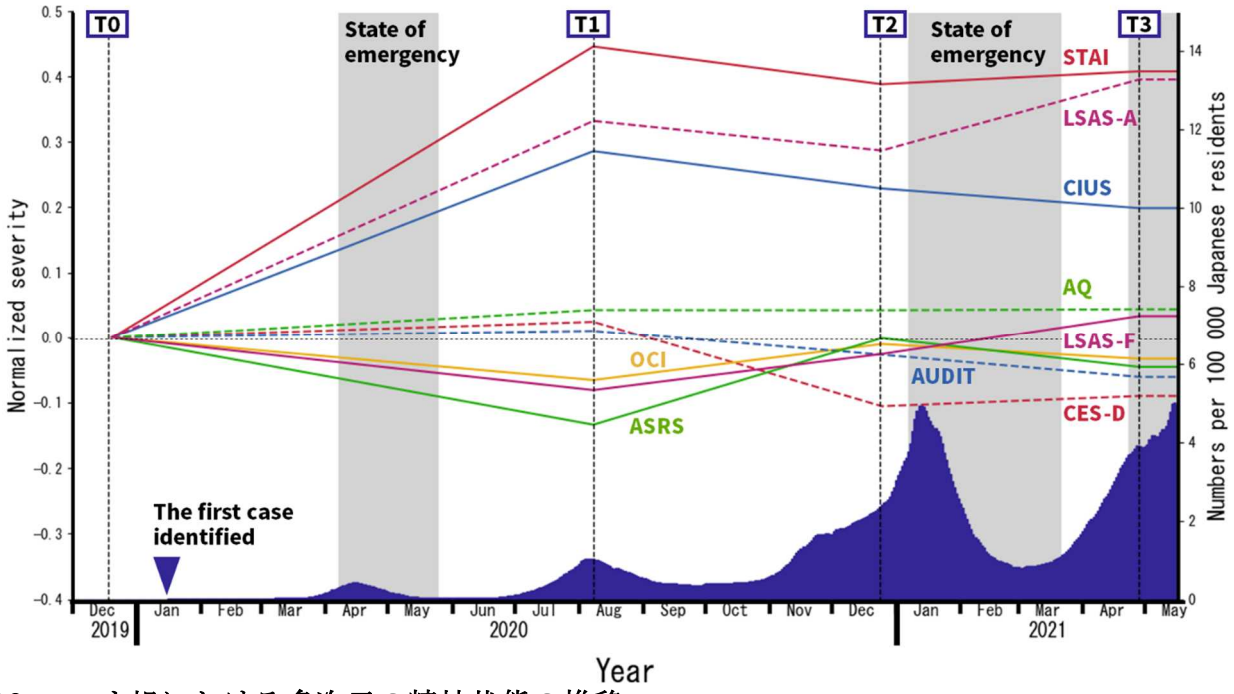


図 26 コロナ禍における多次元の精神状態の推移

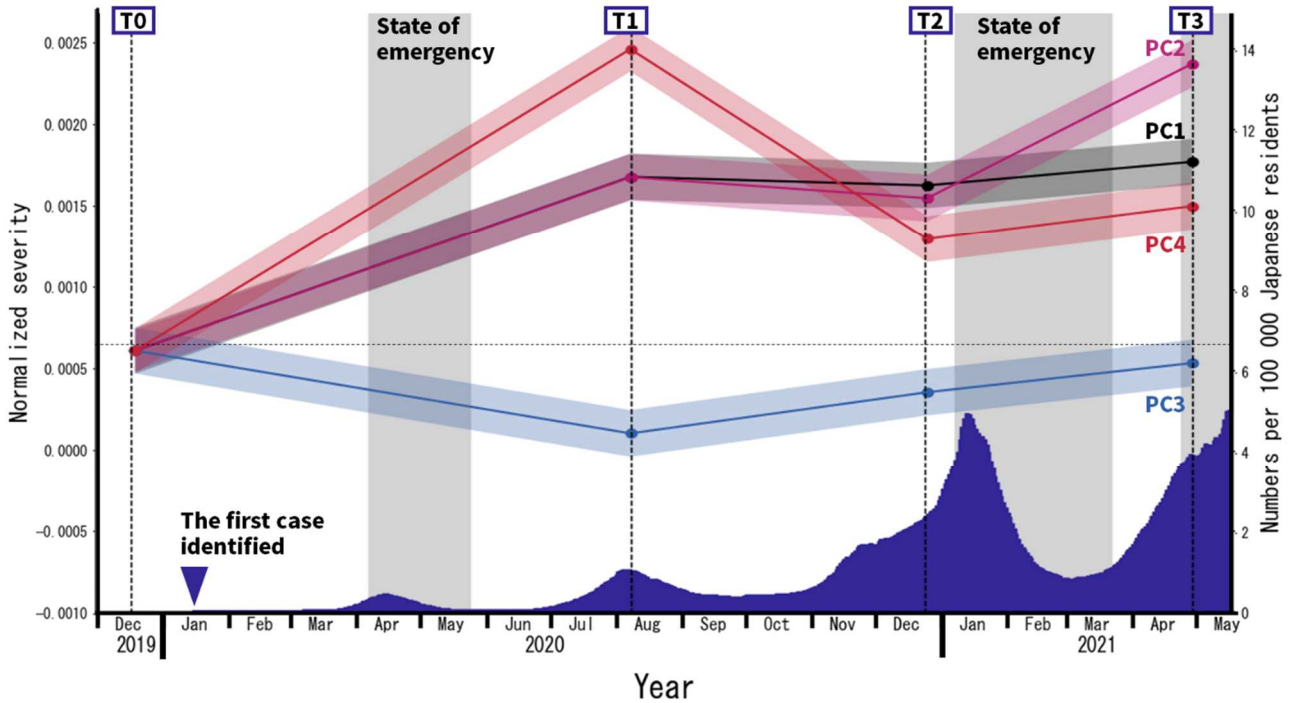


図 27 コロナ禍における共変動する多次元の精神状態の要素の推移

### 3.9 (3-2) 潜在脳ダイナミクスを考慮した模倣学習のための予備解析

期間中には実施していない。

### 3.10 (4) プロジェクトの総合的推進

プロジェクト全体の進捗状況を確認するために、毎週リモート会議ツールを使ったミーティングを実施した。また研究員ごとに進捗を確認するために、必要に応じて数時間のミーティングを実施した。

### 参考文献


- Asadi, K. et al. (2018). Towards a simple approach to multi-step model-based reinforcement learning. CoRR.
- Chiba, T. et al. (2020). A reciprocal inhibition model of alternations between under-/overemotional modulatory states in patients with PTSD. Molecular Psychiatry.
- Chiba, T. et al. (2020). PTSD symptoms related to COVID-19 as a high risk factor for suicide - Key to prevention. medRxiv preprint.
- Fujimoto, S. et al. (2018). Addressing Function Approximation Error in Actor-Critic Methods. In Proc. of the 35th International Conference on Machine Learning.
- Haarnoja, T. et al. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Proc. of the 35th International Conference on Machine Learning.
- Oka, T. et al. (2021a). Prevalence and risk factors of internet gaming disorder and problematic internet use before and during the COVID-19 pandemic: A large online survey of Japanese adults. Journal of Psychiatric Research 142, 218–225.
- Oka, T. et al. (2021b). Multiple time measurements of multidimensional psychiatric states from immediately before the COVID-19 pandemic to one year later: a longitudinal online survey of the Japanese population. Translational Psychiatry 11(1).
- Tayor et al. (2022). Depressive symptoms reduce when dorsolateral prefrontal cortex-precuneus connectivity normalizes after functional connectivity neurofeedback. Scientific Report, 12(1).
- Uchibe, E. (2018). Model-free deep inverse reinforcement learning by logistic regression. Neural Processing Letters, 47(3): 891-905.
- 内部英治。(2020)。モデルフリーとモデルベースの協同による並列深層強化学習。第34回人工知能学会全国大会予稿集。
- 内部英治。(2021a)。モデルフリーとモデルベース強化学習のための非同期並列学習。第35回人工知能学会全国大会予稿集。
- 内部英治、松原崇充、森本淳。(2020)。形態の異なるロボット間での敵対的生成模倣学習。第38回日本ロボット学会学術講演会予稿集。
- 内部英治。(2021b)。方策と環境モデルを生成モデルとして学習する敵対的生成模倣学習。第39回日本ロボット学会学術講演会予稿集。
- Uchibe, E., and K. Doya. (2021). Forward and inverse reinforcement learning sharing network weights and hyperparameters. Neural Networks 144: 138-153.
- Wang, J., S. Elfwing, and E. Uchibe. (2021). Modular deep reinforcement learning from reward and punishment for robot navigation," Neural Networks, 135: 115-126.
- Zhu, J.-Y. et al. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proc. of the IEEE International Conference on Computer Vision.


## ４．委託業務全体の成果

### ４．１　計画時に想定していなかった成果（副次的成果）や、目標を超える成果

4.1.a 形態の異なるロボット間でのモデルベース順・逆強化学習

3.2節で述べたモデルベース順・逆強化学習はエキスパートと学習者が同じ状態空間・行動空間を持ち、状態遷移確率も同一であることを仮定していた。そのためモーションキャプチャで取得した人の行動データをロボットに転移することができなかった。我々がすでに開発した構造化された識別器(Uchibe, 2018; Uchibe and Doya, 2021)にモデル情報とCycle Consistencyと呼ばれる正則化を導入することで、この問題を解決する。

エキスパートはMDPに従い、${}^{E}\mathcal{M} = \langle {}^{E}\mathcal{X}, {}^{E}\mathcal{U}, {}^{E}p_T, {}^{E}r, {}^{E}\gamma \rangle$で与えられると仮定する。ここで${}^{E}\mathcal{X}$は状態空間、${}^{E}\mathcal{U}$は行動空間、${}^{E}p_T({}^{E}x' \mid {}^{E}x, {}^{E}u)$, ${}^{E}x, {}^{E}x' \in {}^{E}\mathcal{X}$, ${}^{E}u \in {}^{E}\mathcal{U}$は状態遷移確率、${}^{E}r \in \mathbb{R}$は報酬、${}^{E}\gamma \in [0,1]$は将来の報酬に対する割引率である．エキスパートは期待積算報酬の最大化によって最適方策${}^{E}\pi({}^{E}u \mid {}^{E}x)$を学習し、最適方策に従って状態遷移の集合

$$ {}^{E}\mathcal{D} = \{({}^{E}x_i, {}^{E}x_i')\}_{i=1}^{{}^{E}N} $$

が得られていると仮定する。${}^{E}N$は状態遷移対の個数である。エキスパートは人や動物からの行動データを想定しているので、エキスパートの行動は直接観測できず、新規にデータを収集することもできない。学習者はMDPから報酬を取り除いた${}^{L}\mathcal{M} = \langle {}^{L}\mathcal{X}, {}^{L}\mathcal{U}, {}^{L}p_T, {}^{L}\gamma \rangle$を想定する。報酬の代わりにエキスパートと学習者の状態の対応関係

$$ {}^{L}\mathcal{D} = \{({}^{L}x_i, {}^{L}x_i')\}_{i=1}^{{}^{L}N} $$

が与えられている。${}^{L}N$は状態遷移対の個数である。学習者の目的はエキスパートとの状態の対応関係を満足しつつ、エキスパートと同様の行動を模倣することである。学習者が利用できるのは${}^{E}\mathcal{D}$, ${}^{L}\mathcal{D}$のみで、状態遷移確率や報酬は未知である。また割引率は簡単のため${}^{E}\gamma = {}^{L}\gamma$と仮定し、以降$\gamma$と記述する。

エキスパートと学習者の同時分布${}^{E}p$, ${}^{L}p$が異なるため、拡大した識別器を

$$ D({}^{E}x', {}^{L}x' \mid {}^{E}x, {}^{L}x) \triangleq \frac{{}^{E}p({}^{E}x' \mid {}^{E}x)}{{}^{E}p({}^{E}x' \mid {}^{E}x) + {}^{L}p({}^{L}x' \mid {}^{L}x)} $$

と定義する（内部ら2020）。図 28に識別器の構造を示す。エキスパートの状態から学習者の状態に変換するネットワークを${}^{L}F_E$、その逆変換ネットワークを${}^{E}F_L$とする。目的関数を

$$ D({}^{L}p, {}^{L}F_E, {}^{E}F_L) \triangleq \mathbb{E}_{{}^{E}p}\left[\ln D\left({}^{E}x', {}^{L}F_E({}^{E}x') \mid {}^{E}x, {}^{L}F_E({}^{E}x)\right)\right] $$
$$ + \mathbb{E}_{{}^{L}p}\left[\ln\left(1 - D\left({}^{E}F_L({}^{L}x'), {}^{L}x' \mid {}^{L}x, {}^{E}F_L({}^{L}x)\right)\right)\right] $$

と設定する。$\mathbb{E}_{{}^{E}p}$はエキスパートから生成された状態遷移に関する期待値で、$\mathbb{E}_{{}^{L}p}$も同様である。

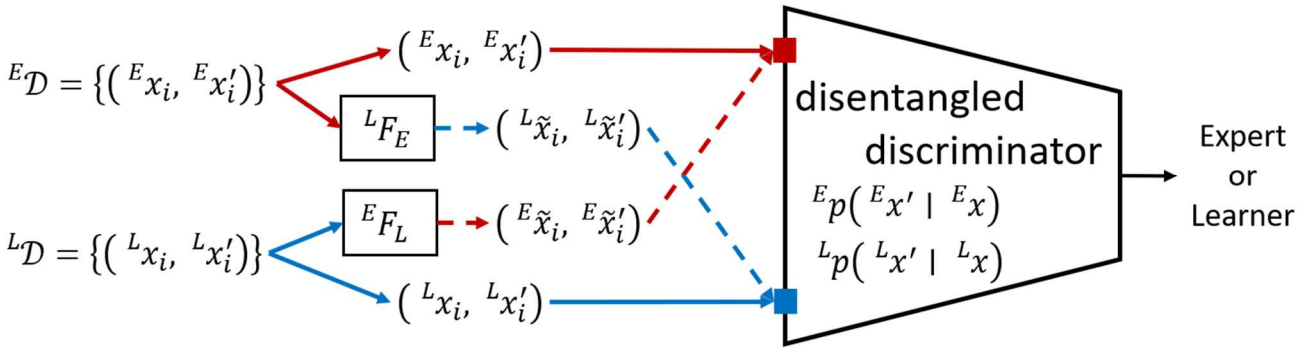


**図 28 形態の異なるロボット間でのモデルベース順・逆強化学習のための拡大した識別器**

さらに${}^{E}F_L$、${}^{L}F_E$の学習のために以下の二種類の損失関数を追加する。一つは事前に与えた状態の対応関係に関するpairwise損失

$$ V({}^{L}p, {}^{L}F_E, {}^{E}F_L) \triangleq \frac{1}{{}^{P}N}\sum_{i=1}^{{}^{P}N} {}^{L}d({}^{L}F_E({}^{E}x_i), {}^{L}x_i) + {}^{E}d\left({}^{E}x_i, {}^{E}F_L({}^{L}x_i)\right) $$

である。${}^{E}d$、${}^{L}d$はそれぞれエキスパートと学習者の状態空間上での対応関係を測る距離関数である。もう一つの損失関数はCycleGAN (Zhu et al., 2017)と同様のcycle consistency

$$L_{cycle}\left({}^{L}F_{E},\ {}^{E}F_{L}\right)=\frac{1}{{}^{E}N}\sum_{i=1}^{{}^{E}N}\left({}^{E}x_{i}'-{}^{E}F_{L}\left({}^{L}F_{E}\left({}^{L}x_{i}\right)\right)\right)^{2}+\frac{1}{{}^{E}N}\sum_{i=1}^{{}^{E}N}\left({}^{L}x_{i}'-{}^{L}F_{E}\left({}^{E}F_{L}\left({}^{E}x_{i}\right)\right)\right)^{2}$$

である。以上の三種類の損失関数をまとめた

$$\min_{{}^{L}F_{E},\ {}^{E}F_{L}}\left[-V+\eta_{cycle}L_{cycle}+\eta_{pair}L_{pair}\right]$$

を状態変換ネットワークの目的関数とする。$\eta_{cycle}$、$\eta_{pair}$は正のハイパーパラメータである。

生成器の学習はGANと同様に${}^{E}F_{L}$、${}^{L}F_{E}$を固定しつつ$V\left({}^{L}p,\ {}^{L}F_{E},\ {}^{E}F_{L}\right)$を最小化するように${}^{L}p$を更新する。ただし通常のGANとは異なり、識別器が${}^{L}p$に依存するため右辺第1項を省略できないことに注意されたい。また

$${}^{L}p\left({}^{L}x'\mid{}^{L}x\right)=\int{}^{L}p\left({}^{L}x'\mid{}^{L}x,\ {}^{L}u\right){}^{L}\pi\left({}^{L}u\mid{}^{L}x\right)d\,{}^{L}u$$

であり、今回は${}^{L}p$、${}^{L}\pi$ともに多変量正規分布と仮定し、解析的に${}^{L}p\left({}^{L}x'\mid{}^{L}x\right)$も多変量正規分布で表現する。ただし平均と分散はニューラルネットワークで表す。一方、エキスパートの状態遷移確率${}^{E}p$は${}^{E}\mathcal{D}$から最尤推定などを用いて

$${}^{E}\theta^{*}=\arg\min_{{}^{E}\theta}\ln{}^{E}p\left({}^{E}x'\mid{}^{E}x\right)$$

を最小にするように事前に学習する。${}^{E}\theta$は分布${}^{E}p$のパラメータである。ここではエキスパート分布が多変量正規分布$\mathcal{N}\left({}^{E}\mu,\ {}^{E}\Sigma\right)$に従うと仮定し、平均${}^{E}\mu$と共分散行列${}^{E}\Sigma$の対角成分を明示的にニューラルネットワークで表現する。

提案手法を倒立振子安定化のための行動を自転車の安定化の学習に転移する課題を用いて検証した。図 29にそれぞれのモデルを示す。倒立振子は台車の位置と速度、振子の角度と角速度を状態、台車に与える力を行動とする。なお、すべての状態は区間[-1, 1]で正規化されている。一方、自転車は地面の法線方向と自転車のなす角度と角速度、ハンドルバーの角度と角速度を状態、ハンドルバーに与えるトルクを行動とする。自転車は限界速度よりも遅い速度(2.778 [m/s])で前進することで自転車を安定化しにくい設定としている。両課題は状態の次元は同じで制御の目的も安定化と共通しているがダイナミクスは異なる。

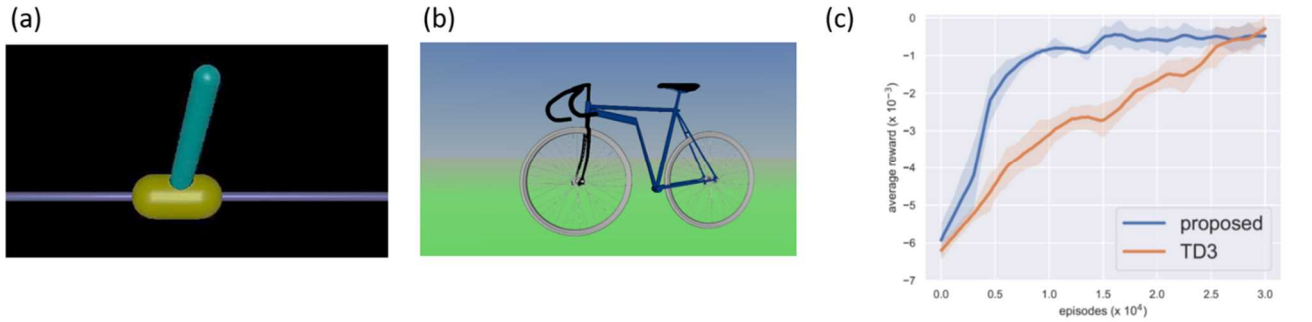


**図 29 倒立振子の安定化課題と自転車の安定化課題。(a) 倒立振子モデル。(b) 自転車の安定化。(c) 学習曲線の比較。**

倒立振子のエキスパートデータはTwin Delayed Deep Deterministic policy gradient (TD3)アルゴリズム (Fujimoto et al., 2018) を用いて訓練した方策を用いて${}^{E}N=100$個の状態遷移の組を生成し

た。倒立振子と自転車の状態間の距離関数は不安定平衡点周りの角度をマッチングするために振子の角度と、自転車と地面の法線方向のなす角度を対応させるよう定義した。原点を含む$^{P}N =$ 100個のランダムに生成した状態の組を$^{P}\mathcal{D}$として用いた。比較のために自転車の安定化課題を直接TD3によって学習した。その際の報酬は転倒した場合に$-1$、それ以外は0と設定した。学習曲線を比較した結果を図 29(c)に示す。TD3の学習曲線は通常のTD3の学習中に得られるエピソードごとの平均報酬を示している。一方、提案手法の学習曲線は比較のために学習中に得られた報酬の平均を表示しているが、提案手法の学習には報酬は使用していないことに注意されたい。提案手法はTD3を用いて直接学習するよりも約1/3程度の学習データ数で学習できていることがわかる。

次に図 30 (a)、(b)に示す、状態空間の次元の異なるReacher間での転移課題を用いて提案手法の有効性を検証する。目的は二つの障害物との衝突を避けつつ、手先を緑色の目標位置に移動させることである。目標位置と障害物はエピソードごとにランダムに配置される。Reacherの状態は各リンクの角度と角速度、目標地点の位置、障害物の位置で構成される。すなわちエキスパートが学習する2リンクモデルでは10次元の状態ベクトル、4リンクモデルでは14次元の状態ベクトルを用いる。それぞれの行動は関節角変位とする。両課題は制御の目的は同じであるが、状態空間の次元が異なるため従来研究はそのままでは適用できない。

2リンクReacher課題のエキスパートデータとして、TD3を用いて訓練した方策を用いて$^{E}N =$ 100個の状態遷移の組を生成した。二つのReacherの状態間の距離関数は、関節角度のみの二乗距離を用いて定義した。また目標位置や障害物の位置は共有できるため、関節角度と角速度のみを変換するようネットワークを構成した。比較のために4リンクReacher課題を直接TD3で学習した。報酬は手先と目標位置との距離に応じて設定した。シミュレーション結果を図 30(c)に示す。提案手法とTD3の学習性能を比較するためにエピソード終了状態での報酬値の平均を示している。この実験においても提案手法はTD3を用いて直接学習するよりも素早く学習できていることがわかる。すなわち、開発した手法はTD3の半分程度の学習データ数で最適方策を学習している。

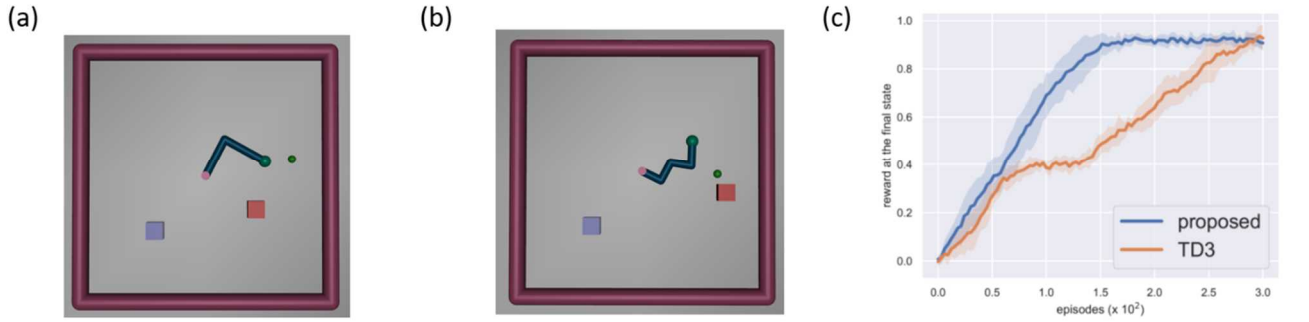


**図 30 Reacher 課題。(a) 2 リンクモデル。(b) 4 リンクモデル。(c) 学習曲線の比較。**

### ４．２　研究課題の発展性（間接的成果を含む）

開発した順・逆強化学習をベースとして、拠点である国際電気通信基礎技術研究所で様々な研究がスタートしている。一つは人間参加型の順・逆強化学習である。学習システムとの関係の観点からは、本研究ではエキスパートは目標となるデータを事前に与える静的なものであったのに対し、新しいプロジェクトではどのような情報をエキスパートに提供すれば学習システムを含めた全体が最適化できるか、といった共進化を対象としている。また3. 4節で述べたモデルベース法とモデルフリー法の非同期協調の研究で着目した機械学習アルゴリズムのサンプリング周期の問題は、機械学習アルゴリズムを実際の問題に適用する際に今後重要になると考えられる。

一方で、ニューロフィードバック研究においては、コロナウイルス感染拡大の影響もあり、定期的に被験者・患者を研究拠点に迎えることが困難となり、実験計画に大幅な変更があった。また実験の性質上、被験者・患者の訴えにより実験を中断するなどの問題も発生した。そのため、途中で中断したデータも利用可能な統計処理、機械学習アルゴリズムの開発が必要であることも判明した。

### ４．３　研究成果の発表・発信に関する活動

内部英治が研究「モデルフリーとモデルベース強化学習のための非同期並列学習」について、2021年度人工知能学会全国大会優秀賞を受賞した。これは3.4.2節の内容に対応する。

## ５．プロジェクトの総合的推進

### ５．１　研究実施体制とマネジメント

円滑に領域横断的な研究を実施するために、メンバー全員参加のミーティングを毎週実施し、進捗状況を確認した。ただし新型コロナウイルス感染拡大防止の観点から、Zoom を用いたネット会議を用いた。また必要に応じて少数メンバーに限った対面ミーティングを実施した。対面ミーティングの回数は減少したが、国内外の研究者とのミーティングが可能となり、より高度かつ専門的な意見交換が実施できた。

購入した GPU 計算機サーバ群は Kubernetes を用いて仮想化することで、使用者の要求に柔軟に対応できる開発・シミュレーション環境を構築した。また実ロボットを用いた評価を前倒しで実施したことにより、アルゴリズムの実装に必要なソフトウェアを計画より早く整備できた。

### ５．２　経費の効率的執行

コロナ禍の状況においてデータ収集に遅れがあったが、おおむね計画通りに実施した。

## ６．まとめ、今後の予定

本研究は残り2年を残して終了となったため、最終目標として計画していた項目に取り組む予定である。(1) アルゴリズムの開発と工学応用については、モーションキャプチャ等で計測したエキスパートの柔軟物の操りなどの行動をロボットに移植する作業を進める。このとき、4.1節で述べた異構造間での模倣学習を発展させる。(2)ダイナミクスを記述する特徴量・学習データの抽出と医療応用においては、うつ度軽減機能結合ニューロフィードバックおよびPTSD患者を対象としたニューロフィードバックによる治療を継続し、治療の長期効果を確認するとともに、マインドフルネス介入の併用効果を検証する。(3) アルゴリズムの医療応用では、エントロピ正則化強化学習を人の順強化学習のモデルとして解析に用いるとともに、ニューロフィードバックのデコーダと逆強化学習の識別器の統合を進める。さらに人データに適用して潜在脳ダイナミクスの同定を進める計画である。

## ７．研究発表、知的財産権等の状況

### （１）研究発表等の状況

| 種別 | 件数 |
|---|---|
| 学術論文 | 7件 |
| 学会発表 | 9件 |
| 展示・講演 | 17件 |
| 雑誌・図書 | 9件 |
| プレス | 該当なし |
| その他 | 5件 |

### （２）知的財産権等の状況

該当なし

### （３）その他特記事項

該当なし