

# 博士学位論文

氏名 (本籍)	高橋 春輝 (埼玉)
学位の種類	博士 (情報学)
学位記番号	第 189 号
学位授与年月日	2024 年 9 月 30 日
学位授与の要件	学位規則第 4 条第 1 項
学位論文題目	冗長な観測に対する状態削減問題と 目的指向な強化学習環境の推定

論文審査委員	主査 大和淳司 教授
	副査 真鍋義文 教授
	// 田中久弥 教授
	// 竹川高志 教授
	// 鮫島和行 教授 (玉川大学)

工学院大学大学院

# 目次

1	はじめに	1
2	強化学習環境に対する状態削減問題としての定式化	6
2.1	既存の強化学習エージェント	6
2.2	状態削減とコア	10
2.3	観測に対するノイズ情報の分類	11
3	次元削減によるアプローチ	15
3.1	線型変換による定式化	15
3.2	判別分析	16
3.3	ラプラシアン固有マップ法	17
3.4	判別的ラプラシアン固有マップ法	18
3.5	類似度の低次元化処理	21
3.6	平均化と低次元処理の効果検証	22
4	環境推定によるアプローチ	24
4.1	冗長観測性と部分観測性によるアプローチ	24
4.2	変分ベイズによる環境推定	27
4.3	棒折過程によるノンパラメトリック推定	28
4.4	完全な環境推論	29
4.5	目的指向な環境推論	33
4.6	忘却によるオンライン処理	37
4.7	行動価値トンプソンサンプリングによる行動制御	38
5	数値実験	39
5.1	コアとノイズによるテスト環境	39
5.2	非定常なテスト環境	41

5.3	状態削減の評価指標 . . . . .	42
5.4	コア推定による戦略学習の効果検証 . . . . .	43
6	考察とまとめ	49
	謝辞	53
	参考文献	54

# 1 はじめに

人間や動物は、感覚情報を通じて世界を認識し、試行錯誤を行いながら学習を行い、段階的に適切な行動を行うことができるようになる。特に人間は視覚を中心に膨大な入力信号を受け取りながら、その中から余計なノイズ情報を排除し、目的に必要な情報を効果的に抽出しながら、長期的な手順の計画を必要とするタスクの解決方法を見つけ出すことができる [1, 2]。例えば、自動車の運転時には、膨大な視覚情報の中から、信号や標識などの静的な情報や他の車や歩行者の動きなどの動的な情報を抜き出している。そして、接触事故などの致命的な失敗を避けながら、短い訓練期間で車庫入れやクランク走行などの技術を獲得している。このような高度な知的能力はどのように実現されているのだろうか。人間と同等かそれ以上の能力をもつエージェントが実現できれば、危険の伴うタスクや高精度で安定した作業が求められるタスクなどを代替することができようになる。

このようなエージェントには、なるべく少ない試行錯誤でタスクに適応できるようなサンプル効率が高い条件を満たすことが望ましい。また、実世界におけるタスクに応用した際には、なぜエージェントがその行動を選択したのかを開発者やユーザーが解釈できなければ、エージェントの正当性や安全性を評価することができない。よって、実世界のタスクに応用できるエージェントを実現するには、有益な情報を抽出する情報処理と目的を達成するための行動制御の2つの技術的課題に加えて、エージェントの行動原理に対する説明可能性 [3, 4] が課題になる。計算論的神経科学や人工知能の分野では、これらの課題を解決するために有用な手法が研究されてきた。

計算論的神経科学の分野では、神経細胞が結合したネットワーク（神経回路網）をモデル化し、教師信号による学習により入力信号に対して望ましい出力を実現する情報処理が実現できることを示した。ネットワークを構成するニューロンの発火パターンは、ニューロン同士の結合力パラメータを学習することで複雑に変化する。複数の層上にニューロンを配置した多層ニューラルネットワークモデルに対して、ネットワークが出力する予測誤差を小さくするように結合力パラメータを効率的に学習する方法であるバックプロパゲーション [5, 6] が提案されることで、複雑なパターン認識タスクを解くことができるようになった。これら



の技術は人工知能分野にも応用され、各層のニューロンが2次元上に配置された画像処理に特化した Convolutional Neural Network [7, 8] や、中間層のニューロンの出力が自己結合する時系列処理に特化した Recurrent Neural Network [9] が提案された。近年では、膨大なデータを用いてパラメータを学習することで、物体領域 [10] や癌部位 [11] の特定などの画像認識や、文章生成や翻訳 [12, 13] などの自然言語処理で、複雑な情報から有益な情報を抽出する情報処理を自動化することが可能となった。

また、人工知能の分野におけるベイズ推論では、情報の因果構造をモデル化することで、有益な情報を抽出する試みが行われている。Gaussian Mixture Model (GMM) [14, 15] は、有益な情報に付加されたノイズを除去するモデルであり、データ分析におけるクラスタリングや異常検知のタスクに応用されている。また、Complete Environment Inference [16] はノイズを除去した有益な情報を表現している状態（クラスタ）が行動によって変化する時間発展を予測するモデルである。ベイズ推論に基づくエージェントは計算論的神経科学の分野においても研究が行われている。モデルによる情報の予測誤差を最小化する自由エネルギー原理に従う Active Inference [17, 18] によって、脳の仕組みの解明に取り組まれている。

強化学習と呼ばれる人工知能の分野では、エージェントが制御する行動によって変化する情報の因果構造を環境として定義し、その中で行動制御の自動化を図っている。エージェントの目的は、食糧や貨幣などの報酬をなるべく多く獲得することとして解釈される。エージェントが選択した行動は、すぐに得られる報酬だけでなく、将来の報酬にも影響するため、長期的な計画を考慮して行動を制御しなければならない。Q-Learning [19] は離散的な情報をもとに、行動の結果として得られる実際の報酬と予測した報酬の誤差を小さくするように学習することで、将来の報酬を考慮した行動制御が可能であることを示した。人間が行動を学習する過程においても、脳内のドーパミン細胞の活動が報酬の予測誤差を示すことがわかっており、強化学習は人間の意思決定にも深く関わっていることが示唆されている [20]。強化学習は、スマートグリッドにおけるエネルギー管理システム [21] や患者の特性に基づいた治療戦略の個別化 [22] など時系列イベントにおける行動制御の自動化に応用されてきた。しかし、入力信号が離散的であるという制限を置いた手法では、視覚情報などの複雑な入力信号に対して行動制御を自動化することは困難である。

近年では、Deep Learning と強化学習の技術が発展したことで、それらを融合した深層強化学習によって、複雑な情報処理に基づいた行動制御の自動化の試みが行われている。深層強化学習は、ネットワークが出力する報酬の予測誤差を小さくするように学習することで、膨大な試行錯誤を通じて目的を達成する行動を実現する。Deep Learning は環境に対して特定の因果構造を指定することなく情報処理を可能にするため、因果構造が不透明なタスクに対しても広く適用できる。しかし、ネットワーク内部で因果構造自体の学習からしなくてはならないことが原因で、サンプル効率は低い学習となっている。Deep Q-Network(DQN) [23] は Atari ゲームなどの画像で表現できる複雑な視覚情報に基づいた行動制御を可能にした。深層強化学習による取り組みは、ロボット制御 [24, 25] や自動運転 [26] などにも広がっており、実世界でのエージェントの開発の機運が高まってきている。

深層強化学習のアプローチによるエージェントの実現は、情報処理と行動制御の課題を解決できる。しかし、学習におけるサンプル効率の低さは多くの試行錯誤が必要なことを意味するため、学習にかかる物理的な時間の制約がこのアプローチによるエージェントの実現を困難にする。実際に DQN が Atari ゲームの学習に使用したデータ量は 2 億フレームにのぼり、物理的な時間に換算すれば 900 時間以上のデータ量となる。この課題を解決するために、まずシミュレーション上で高速にコストをかけずに大量のデータでエージェントを訓練し、その後に実世界で学習させる転移学習という方法が考えられる [27, 24]。しかし、シミュレーションは実世界を完全に再現できないため、シミュレーションと実世界の学習条件のギャップが生じることで、転移学習はうまく働かないことがある [28]。特に、センサーから得られる実世界の情報にはノイズによる不確実性が伴うことや、また照明の変化や人の動きなどのエージェント置かれた環境の動的な変化など、完全な再現が難しい要素が多くある。

これに対して、ベイズ推論と強化学習を組み合わせたアプローチは、情報の因果構造を直接モデル化するため、学習時におけるサンプル効率の課題を解決できる可能性がある。また、状態の時間発展を記述する遷移則も学習するため、選択した行動が未来の報酬に与える影響を予測できるようになり、行動原理を人が解釈しやすい利点もある。しかし、全ての情報に対する因果構造をモデル化するため、情報量の多い観測をもつ環境では、状態数の 2 乗に

比例する記憶容量が必要な遷移則の学習に、メモリ容量が不足する実装上の課題がある。

実環境におけるエージェントのセンサーから得られる情報には、有益な情報だけでなく、無駄な情報も多く含まれる。そのため、全ての情報に対してモデル化する必要はなく、有益な情報に限定した簡潔な因果構造をモデル化することができれば、メモリ要求は軽減する。さらに、因果構造を簡潔化することはエージェントの利用時における説明可能性の向上につながる。よって、強化学習環境において有益な情報に限定した因果構造をモデル化するアプローチが、実世界でのエージェントの開発に有効であると考えられる。

強化学習環境におけるエージェントの目的はなるべく多くの報酬を獲得することである。よって、報酬の予測に寄与する情報が有益な情報である。また、有益な情報は現在の時刻で観測できる情報だけでなく、過去の時刻にも含まれている場合がある。このことを考慮し、過去の時系列に含まれる全ての情報から、報酬に関連する情報を状態として抽出し、その状態の時間発展を推定する問題として状態削減問題を定義する。報酬に関連する情報のみを表現したコア状態をエージェントが獲得できれば、最もサンプル効率の高い学習が可能であり、同時に最も簡潔な因果構造でエージェントの行動原理を説明することが可能となる。

本論文は、有益な情報に限定した簡潔な因果構造のモデル化を目的とした

- [29] Takahashi, Kazuki, and Takashi Takekawa. "Discriminant laplacian eigenmaps by the approximation of discriminant analysis using similarity." *Nonlinear Theory and Its Applications, IEICE 13.2* (2022): 300-305.
- [30] Takahashi, Kazuki, et al. "Goal-oriented inference of environment from redundant observations." *Neural Networks 174* (2024): 106246.

の成果をまとめたものである。2章では、膨大な情報から有益な情報を抽出する過程を強化学習環境の学習に組み込んだ状態削減問題を定義する。3章は [29] の内容であり、情報の構造を利用して状態削減問題にアプローチする次元削減による手法を提案する。4章と5章は [30] の内容であり、構造を利用できない状態削減問題に対して、観測の生成過程を仮定したベイズ推定によるアプローチを提案する。6章はこれらの結果に対する考察とまとめを述べる。環境を概念化した簡潔な因果構造に基づいたベイズ推論は、サンプル効率の課題を解決

し説明可能性の高い実世界でのエージェントとしての応用に加えて，人間の高度な知的能力を解明する足がかりとなることが期待される．

## 2 強化学習環境に対する状態削減問題としての定式化

### 2.1 既存の強化学習エージェント

実世界の多くのタスクでは、様々な摂動や非定常変化を伴う未知の環境ダイナミクスから出現する観測履歴に基づいて、将来の報酬を最大化するための適切な行動を決定する必要がある [31, 32, 33]. 適応制御や強化学習の分野では、このような複雑な環境において適切な行動を学習するための数学的枠組みを構築する試みがなされてきた [34, 35, 36, 37].

強化学習は、エージェントが未知の環境ダイナミクスで試行錯誤を通じて、与えられた状況に適した最適な行動を生成する戦略を学習する数学な枠組みである (図 1). エージェントの行動  $a_t$  の結果として、環境は観測  $o_{t+1}$  と報酬  $r_{t+1}$  をエージェントに返す. 自動車の運転の場合では、信号や車間距離・速度などの情報から、道路の中央からの距離である負の報酬を最大化するようにハンドルやアクセルなどの行動を制御するタスクとして記述される. この時、エージェントの目的は、未来の報酬の累積を最大化する戦略  $\pi(a_t | o_t)$  を学習することである. 次の瞬間に車体が道路中央に位置するように急なハンドル操作を行えば、さらに次の瞬間には車体は道路中央を越え蛇行運転することになる. そのため、短期的に報酬が最大になる行動が、長期的に最適な行動になるとは限らない. 強化学習では未来の報

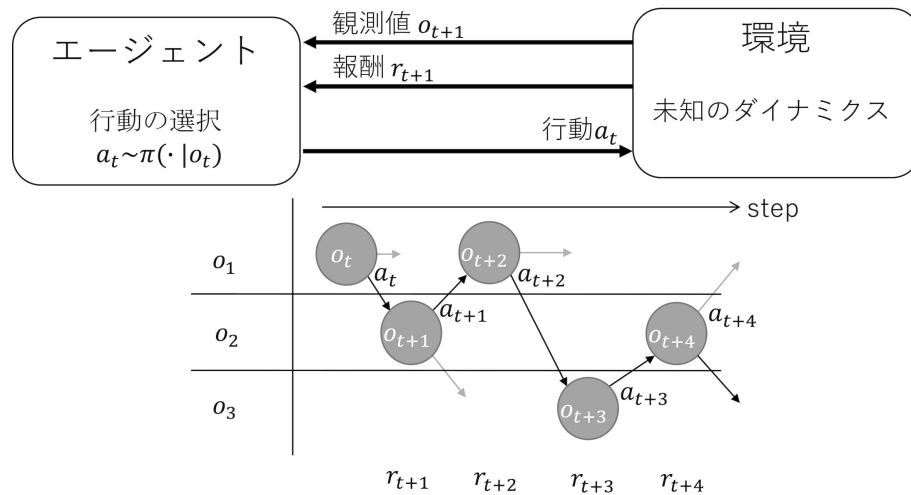


図 1: エージェントと環境の相互作用による強化学習環境の定式化. エージェントは歩道 ( $o_1$ ) や反対車線 ( $o_3$ ) へ外れずに、走行車線 ( $o_2$ ) に収まるような行動制御を学習する.

報酬累積が発散することを防ぐため、その代わりに未来の報酬量を割引率  $0 < \gamma \leq 1$  で低減した累積である長期報酬  $\sum_{t=0}^{\infty} \gamma^t r_t$  を最大化する。割引率は未来の報酬を考慮するためなるべく 1 に近づけるように設定する。この長期報酬を最大にする最適戦略は、行動価値  $Q(o, a) \equiv E[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | o_t=o, a_t=a]$  を計算することで達成される [19]。行動価値は、ある観測と行動のもとでの期待長期報酬を意味しているため、与えられた観測のもとでこの行動価値を最大にする行動を選択することで、その行動は最適戦略となる。この行動価値を計算するには、未知の環境ダイナミクスに何かしらの仮定を置く必要がある。

Q-Learning は環境ダイナミクスがマルコフ決定過程 (Markov Decision Process: MDP) [38] に従うと仮定できる場合に、行動価値を近似的に計算する手法であり、環境ダイナミクス自体は推定しないためモデルフリー手法に分類される。MDP では、次の観測と報酬は、現在の観測と行動のみに依存して決まり、

$$p(r_{t+1}, o_{t+1} | a_t, o_t, r_t, a_{t-1}, o_{t-1}, r_{t-1}, \dots) = p(r_{t+1}, o_{t+1} | a_t, o_t), \quad (1)$$

観測はマルコフ性を満たすとも言う。このとき、観測の遷移則  $p(o_{t+1} | o_t, a_t)$  と報酬則  $p(r_{t+1} | o_{t+1})$  を用いて、行動価値の最大化は最適ベルマン方程式、

$$Q(o, a) = \sum_{o_{t+1}} \left\{ \sum_{r_{t+1}} r p(r_{t+1} | o_{t+1}) + \gamma \max_{a_{t+1}} Q(o_{t+1}, a_{t+1}) \right\} p(o_{t+1} | o_t, a), \quad (2)$$

の解として与えられる (図 2)。この行動価値を最大化する行動をそれぞれの観測で選択することで、長期報酬を最大化する最適な戦略が得られる [39]。Q-Learning は遷移則や報酬則が未知のまま、環境のこの性質を利用して、環境を探索することで行動価値を局所的に近似していく。環境を探索するほど行動価値の近似精度は向上する一方で、過度な探索は報酬獲得の機会損失につながる。そのため、最適戦略を実行可能なほどに行動価値を近似した後には、その行動価値を活用して長期報酬を最大化するように、探索と活用のトレードオフを制御する必要がある。Q-Learning では、温度パラメータ  $\beta$  とソフトマックス関数を用いて、 $\pi(a_t | o_t) = \text{softmax}(\beta Q(o, a))$  で、近似している行動価値をその大きさに比例する確率に変換することで、このトレードオフを制御する。あるいは、あらかじめ決めた割合に応じてランダムに探索と行動価値を最大化する活用を行う  $\epsilon$ -greedy 法が使われることが一般的

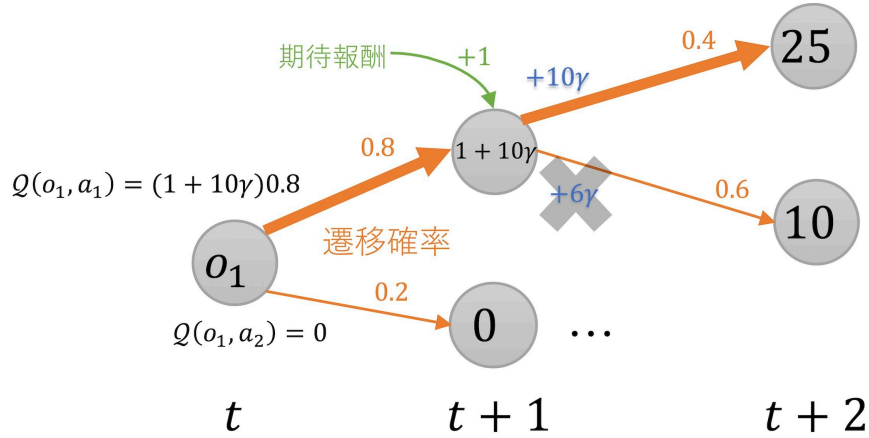


図 2: 遷移則と報酬則による行動価値の再帰的な分解

である.

一般的なタスクにおいては、マルコフ性に従う状態が隠れており、観測からは状態の部分的な情報しか得られないものがある. 特に、自動運転のタスクでは、エージェントが走行中の車線 (状態) は直接観測することはできず、カメラやレーダーによる情報から推定しなければならない. このような状況は部分観測マルコフ決定過程 (Partially Observable MDP: POMDP) [40] で数学的に表現される. POMDP ではエージェントから隠れた状態  $s_t$  がマルコフ性を満たし、その状態から観測と報酬が生成される. したがって、環境が POMDP に従う場合には、Q-Learning は隠れた状態を区別できないため、観測に対する行動価値を近似しても最適戦略を取ることができない. そこで、環境が POMDP に従うと仮定し、観測や報酬そして行動の履歴から、隠れた状態の遷移則と報酬則を推定するモデルベース手法が提案されている. Complete Environment Inference (CEI) は POMDP に従うように履歴を確率的にモデル化する. 遷移則と報酬則のパラメータ  $[\Theta^M, \Theta^N]$  をベイズの定理により更新することで、遷移則と報酬則を確率分布として推定し、その確率分布から履歴を生成する尤もらしい状態を推定する. このことにより、行動価値の分布  $\hat{Q}(s, a)$  が理論的には計算できる (図 3A).

Q-Learning では行動価値を直接近似することで、長期報酬の期待値のみ推定していたため、期待値を確率に変換する必要であった. 一方で、行動価値の分布が計算できる場合には、この分布から直接、環境推定の精度に相関した探索と活用のトレードオフを制御するこ

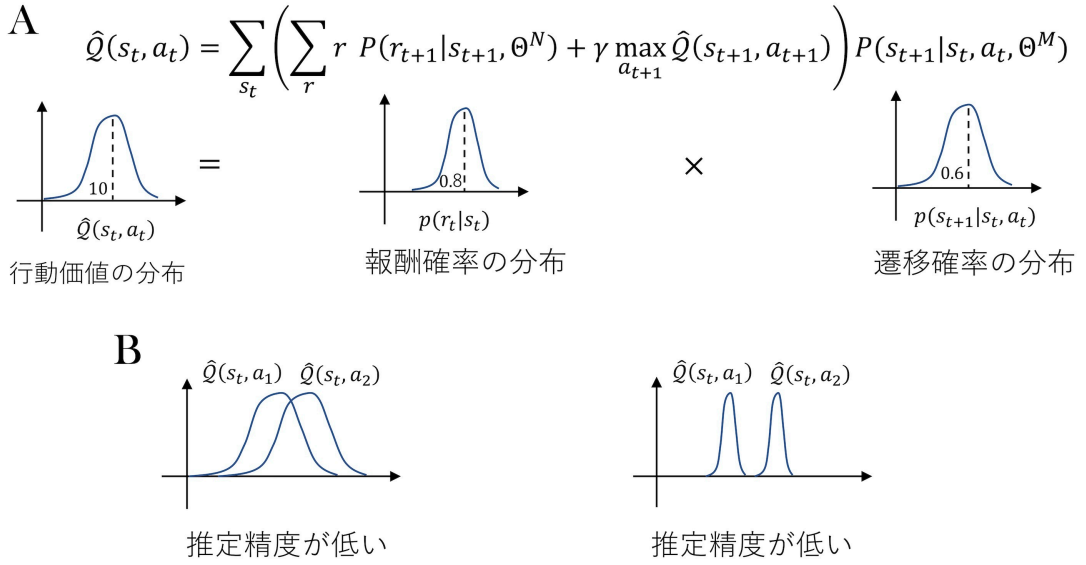


図 3: 行動価値分布による探索と活用のトレードオフの制御. それぞれのグラフは行動価値と報酬確率そして遷移確率を確率変数 (横軸) としたときの確率密度 (縦軸) を示している.

とができる (図 3B). 遷移則と報酬則の推定精度が低く, その確率分布の分散が大きい場合には, それに比例して行動価値分布の分散も大きくなる. そのような場合には, 環境を探索して遷移則と報酬則の推定精度上げる必要がある. 反対に, 行動価値分布の分散が小さい場合には, 遷移則と報酬則の推定精度は最適戦略を取るのに十分であるため, 行動価値の期待値を最大にとる行動を選択し, 報酬獲得を優先した行動を取ることができる.

近年では, これらモデルフリーとモデルベース手法に Deep Neural Network (DNN) を組み合わせることで, 将棋や Atari ゲームなど, 視覚的に複雑な観測をもとに最適戦略を求める手法が提案されている. DQN は, 複雑な観測を行動価値に変化する関数を DNN で近似したモデルフリー手法である [23, 41]. また, 環境ダイナミクスを DNN で近似したモデルベース手法である MuZero が提案されている [42]. MuZero は, 特徴量抽出モジュールと, その特徴量の遷移モジュール, 特徴量から長期報酬を推定するモジュールに, ネットワーク構造を分解することで, 視覚的に複雑な情報からの戦略学習において最先端の性能を示している.



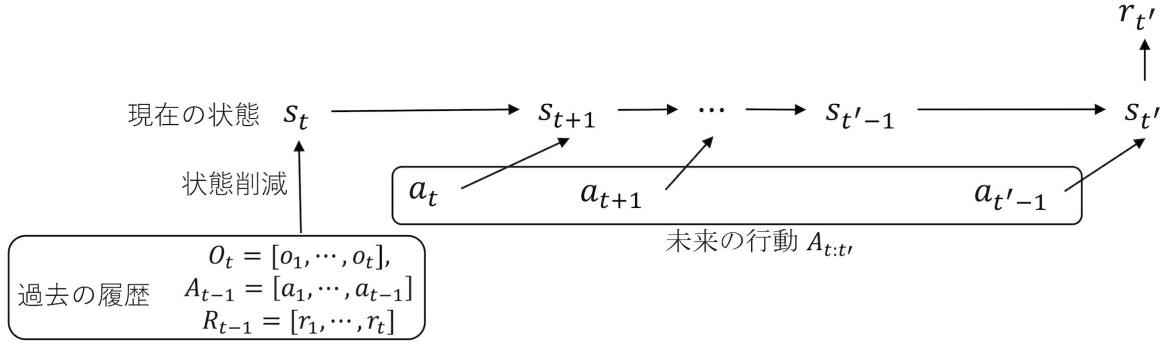


図 4: 状態削減の条件

## 2.2 状態削減とコア

DNN を用いた手法は、複雑な観測が持つ戦略学習に有益な情報を保持するように状態を推定できる。一方で、戦略学習と状態推定に加えて、削減できるノイズ情報の性質を特定する必要があるため、DNN による戦略学習には統計的に十分な大量のデータが必要になる。実世界における応用時には、大量のデータを取得する時間的・金銭的成本を無視することはできない。よって、学習データを減らすために、ノイズ情報の種類ごとに適切なアプローチで状態削減を行うことで、実世界での応用可能性が高まると期待できる。そこで、MuZero でブラックボックス化されているタスクに特化した状態が削減しているノイズ情報を定性的に分類するために、まずはエージェントが直面する状態削減問題を数学的に定義する。以下では、小文字の  $p(\cdot)$  は真の確率分布を意味し、大文字の  $P(\cdot)$  はエージェントがモデル化したものを意味する。

最適な行動を選択するために必要な現在の状態  $s_t$  は、一般に、最近の行動、報酬、観察を含む最近の文脈に依存する。しかし、実際の環境では、エージェントは利用可能な情報の履歴から何を現在の状態と見なすかを決める必要がある。従って、エージェントは履歴から状態  $s_t$  を推測する必要がある。行動  $a_t$  を選択する前の時刻  $t$  において利用可能な情報は、行動履歴  $\mathcal{A}_{t-1} \equiv [a_{t-1}, a_{t-2}, \dots]$ 、報酬履歴  $\mathcal{R}_t \equiv [r_t, r_{t-1}, \dots]$ 、そして報酬と行動以外のエージェントの観測履歴  $\mathcal{O}_t \equiv [o_t, o_{t-1}, \dots]$  である。

極端な例として、現在の行動を決定する前の利用可能なすべての情報を使って状態  $s_t$  を

定義することができる：

$$s_t \equiv \{\mathcal{A}_{t-1}, \mathcal{R}_t, \mathcal{O}_t\} = [r_t, o_t, a_{t-1}, r_{t-1}, o_{t-1}, a_{t-2}, \dots].$$

この定義による状態集合は報酬の有意な情報を失わないため、この過程は定義により MDP の仮定 (1) を満たす。しかし、このような状態集合に対する MDP は状態数が無限となり、現実的には解くことができない。従って、エージェントは、結果として得られる過程が MDP の仮定 (1) を満たし、現在の状態  $s_t$  が、エージェントがどのような行動系列を生成しようとも、全ての履歴から推測可能な将来の報酬に関する情報を失わないような有限の小さな状態集合を見つけるという問題に直面する。エージェントの目的が将来の報酬を最大化することであれば、エージェントは利用可能な全ての情報を用いて以下の報酬確率を推論しなければならない：

$$\forall t' > t, \forall \mathcal{A}_{t:t'},$$

$$p(r_{t'} | \mathcal{A}_{t:t'}, \mathcal{A}_{t-1}, \mathcal{R}_t, \mathcal{O}_t) = \sum_{s_t, \dots, s_{t'} \in S} p(r_{t'} | s_{t'}) \prod_{\tau=t}^{t'-1} p(s_{\tau+1} | a_\tau, s_\tau) p(s_t | \mathcal{A}_{t-1}, \mathcal{R}_t, \mathcal{O}_t). \quad (3)$$

ただし、 $\mathcal{A}_{t:t'} \equiv [a_t, a_{t+1}, \dots, a_{t'-1}]$  は時刻  $t$  から  $t'$  での行動系列である。

エージェントは、条件 (3) を満たす状態削減モデル  $p(s_t | \mathcal{A}_{t-1}, \mathcal{R}_t, \mathcal{O}_t)$ 、状態遷移モデル  $p(s_{t+1} | a_t, s_t)$ 、報酬モデル  $p(r_t | s_t)$  を同時に推論しなければならない。状態削減モデルは、遷移モデルと報酬モデルが適用される状態の集合  $S$  への削減を定義するので特に重要である。一般に、異なる状態集合に対して、条件 (3) を満たす状態削減は無限に存在する。本論文では、最小のサイズ  $|S|$  を持つ状態集合を「コア」と定義する。そして、状態削減問題は与えられた環境のコアへの履歴の削減を求める問題と定義する。正確なコアを見つけることは困難であるため、エージェントは実質的に  $S$  をできるだけ簡略にする削減方法を探す。

## 2.3 観測に対するノイズ情報の分類

本論文では、コア状態の推定に寄与しないノイズ情報の種類を大きく 2 つに分類し (図 5)、それぞれに対する状態削減問題へのアプローチを提案する。一つ目は、観測の構造に起因す

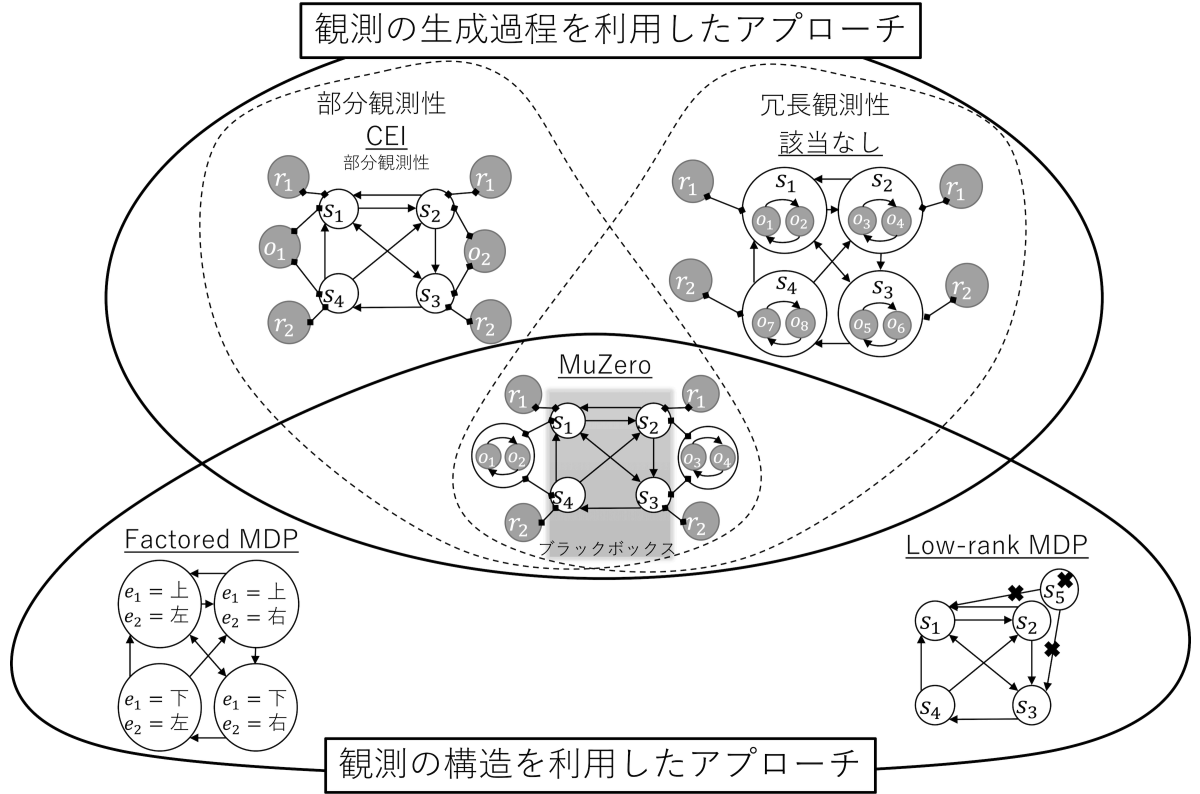


図 5: タスクに特化した状態削減の分類

表 1: 強化学習手法の比較

	モデルフリー		モデルベース		
	Q-Learning	Abstract 強化学習	MuZero	POMDPs Factored MDP Low-rank MDP	提案手法
非定常性	△	△	×	△	○
説明可能性	×	×	△	○	○
状態削減	△	○	○	△	○

るものであり、次元削減などの前処理により解決できるノイズ情報である。二つ目は、観測の生成過程に起因するものであり、前処理では解決できないノイズ情報である。生成過程に起因するノイズ情報に対しては、観測の生成過程に特定の仮定を与え、エージェントモデルによって環境ダイナミクスを推定するアプローチが有効である。モデルベース強化学習でしばしば用いられる部分観測性の仮定は、生成過程に起因するノイズ情報の一つの例である。

強化学習環境において遷移則や報酬則が非定常に変化するタスクは Lifelong 強化学習と呼ばれる [32, 37]. Lifelong 強化学習の場合、ノイズ情報の状態削減問題は困難に直面する. 実世界の環境の真の状態空間は一般に、比較的少数のタスクに関連する状態と、他の多くのタスクに関連しない状態からなり、前者だけが報酬に関連する. コアと呼ぶべき報酬に関連する状態の集合が完全に未知である場合、エージェントが環境の変化に追従することは困難である. 強化学習の実問題への適用性を向上させるためには、コア状態以外の情報を含んだ冗長な観測から報酬に関連する状態空間を自己組織化し、動的に変化する環境を学習するためのサンプル効率の良い戦略を提供する理論的枠組みが必要である [43].

観測から直接行動価値を推定するモデルフリー手法である Q-Learning は、環境の非定常な変化を想定していない. ただし、行動価値の学習率と探索と活用を制御する温度パラメータを調整することで、サンプル効率は低い非定常な変化に追従して戦略を学習できる. しかし、環境ダイナミクス自体を推定しないため、冗長な観測を報酬に関連する状態への削減も、その状態による環境ダイナミクスを利用した行動の説明も行うことができない.

最近のいくつかの研究では、エージェントの過去の観測の完全な集合から構成される状態空間の「コア」を求めることが試みられている. 環境推論に基づく手法は、MDP に従う環境の構造を推定する. 例えば、POMDP は、隠れ状態の履歴から隠れ状態の遷移則を推論し、高い説明力を得ることができる. また、状態空間推定に基づく状態削減法も提案されている [16, 17, 44, 45]. 例えば、Factored MDP [46] は、完全な観測が複数の要素から構成されると仮定して、完全な観測空間を部分的な遷移則の組み合わせとして表現するものである. また、Low-rank MDP [47] は完全観測 MDP の遷移則を低次元の行列に分解する. しかし、これらの方法は、報酬に関連する観測と関連しない観測の両方に基づいて状態空間を削減することに注意すべきである. したがって、冗長な観測に対する困難は部分的にしか解決されておらず、推定された状態空間は必ずしも最適戦略に必要な最小の状態空間ではない.

他の方法として、環境変化の規則を推論しようとせずに、行動価値に従って観測をクラスタリングするものが提案されている. 例えば、Abstract 強化学習 [48] は、行動価値のような指標を用いて、MDP に従う完全な観測空間を、コア状態空間に似た目的指向の状態空間

に削減する [49, 50, 51]. DQN は、複雑な観察から報酬を予測するのに必要な特徴を表す、タスクに特化した状態を推定することができる [23]. しかし、これらの方法は環境ダイナミクスをブラックボックスにしたままであるため、なぜ特定の行動がより大きな報酬につながるのかを説明することができない [42]. 説明可能性を向上させるために、MuZero は特徴抽出モジュールを環境のダイナミクスを推定するモジュールから明示的に分離し、特徴抽出モジュールを報酬推定に使用することで、タスクに特化した状態空間を使った戦略学習において最先端の性能を示している [52, 53].

それにもかかわらず、Lifelong 強化学習の問題設定では、壊滅的忘却 [33] により、DNN は過去の環境を再学習しなければならず [54], DNN による関数近似は非効率的である. 環境推論に基づく手法の中には、環境の変化に応じて複数のモジュールを切り替え、以前に学習した知識を忘れることなく新しい環境に適応するなど、サンプル効率の良い戦略学習を用いてこの困難を解決しようとするものがある [36, 37]. つまり、コア状態とそのダイナミクスを推定しながら、非定常性に適応する有効な手法は知られていない. 環境推論に基づく状態空間の削減手法は前者の問題を解決できず、DNN に基づく手法は後者の問題を解決することが困難である (表.1). Lifelong 強化学習環境における状態空間の最適推定を達成するためには、目的指向の環境推論手法の新しい枠組みが必要である.

エージェントは原理的に、状態空間がすべての可能な観測にまたがる場合と、コア状態のみにまたがる場合のどちらでも、長期報酬を予測することができる. しかし、コア状態は状態削減問題を満たす最小の状態空間であるため、エージェントはコア状態とその遷移則を特定できれば、長期報酬を最大化する最適な戦略をより効率的に学習することができる. このことはまた、エージェントの説明可能性を高めるためには、コアとなる状態を特定することが重要であることを示唆している. さらに、非定常環境における継続的な強化学習と状態空間の動的推定を組み合わせることで、戦略の学習速度を向上させ、環境の急激な変化に追従できる可能性がある. 一般に、状態空間の冗長性が増加すると、状態と行動の組み合わせの数が増加し、最適な戦略を学習するための試行錯誤の回数も増加する. 動的推定は状態空間の無駄な拡大を抑制し、学習速度を向上させると期待される.

### 3 次元削減によるアプローチ

本章では、次元削減により状態削減問題へアプローチする手法を提案する。次元削減は主にクラスタリングの前処理として使われる手法であり、高次元データ  $X$  を低次元の特徴データ  $Y$  に変換する、 $Y = f(X)$ 。この低次元化によって、高次元データのクラスタリングにおける課題である次元の呪い [55] を解消することができる。次元の呪いは、データが高次元になるほど超球面上にデータが分布する現象であり、データ間の距離が均一になることでクラスタリングを困難にする要因である。

高次元データを観測空間  $O$  として低次元データをコア状態空間  $S$  とみなせば、次元削減手法は制約付き状態削減問題を解いていると解釈することができる。このアプローチによる状態削減は、観測値が持つ情報の構造を低次元で表現できる場合に有効である。具体的な例として、画像のピクセル配列から環境を構成する部品に分割するセグメンテーションが挙げられる。ブロック崩しのタスクでは、画像のピクセル配列から環境ダイナミクスを構成するブロックやボールそしてボールを跳ね返す台などの部品の領域を特定するような状態削減が該当する。次元削減は、構造によって複雑化した観測値を状態削減するアプローチとして有効な方針である。

本章の残りでは、データ分布とクラスラベルを用いた教師あり次元削減手法である判別分析と、データ間の類似度を用いた教師なし次元削減手法であるラプラシアン固有マップ法を整理する。そして、両手法を比較することで判別分析を教師なし次元削減に拡張した判別的ラプラシアン固有マップ法を提案し、その性能を検証する。

#### 3.1 線型変換による定式化

以下では、 $n$  個のサンプルデータ  $X_{d \times n} = [x_1, \dots, x_n]$  の線形次元削減を射影ベクトル  $a_{d \times 1}$  による  $y_{n \times 1}$  への線形変換と定義する： $y = X^\top a$ 。また、 $X$  の非線形変換で定義された特徴データ  $\Phi_{d' \times n} = [\phi(x_1), \dots, \phi(x_n)]$  を用いる場合、非線形次元削減を  $y = \Phi^\top a_{d' \times 1}$  と定義する。

非線形次元削減は次元  $d'$  に比例した計算時間を必要とする。ここで、射影ベクトル  $a$  を

特徴データの線形和を意味する  $\Phi b_{n \times 1}$  と  $\delta \perp \phi_i (i = [1, \dots, n])$  を満たす  $\delta_{n \times 1}$  に分解することを考える．このとき、非線形次元削減は

$$y = \Phi^\top (\Phi b + \delta) = \Phi^\top \Phi b, \quad (4)$$

に書き換えることができる．よって、カーネルを  $K_{n \times n} = \Phi^\top \Phi$  と定義することで、非線形次元削減を標本サイズ  $n$  に比例した計算時間で実行でき、理論的には無限次元  $d'$  の特徴量からの次元削減を可能にする．

$r$  次元へ削減する場合には、 $Y_{n \times r} = [y_1, \dots, y_r]$ 、線形と非線形次元削減は射影行列  $A_{d \times r}$  を用いてそれぞれ、 $Y = X^\top A_{d \times r}$  と  $Y = K^\top A_{n \times r}$  で表せられる．クラス数  $k$  があらかじめ分かる場合には、 $r = k - 1$  次元へ削減すると各クラスを  $r$  次元の超平面で分割できる．

### 3.2 判別分析

線形またはカーネル判別分析 (Discriminant Analysis: DA) は、ラベルの行列  $Z_{n \times k} = [Z_1, \dots, Z_k]$  を用いて、 $Y$  空間上の  $k$  個のクラス間での分離度を最大にするように最適な射影ベクトル  $a^*$  を計算する．ラベル行列はサンプルデータ  $x_i$  が  $j$  番目のクラスに属する場合に、 $z_{ij} = 1$  となり、そうでない場合には  $z_{ij} = 0$  となる．このとき分離度は、クラス間分散  $S_b$  とクラス内分散  $S_w$  の比で定義され、これらはそれぞれ、

$$S_b = \sum_{j=1}^k c_j \left( \frac{1}{c_j} \sum_{i=1}^n (x_i z_{ij}) - \frac{1}{n} \sum_{\ell=1}^n x_\ell \right) \left( \frac{1}{c_j} \sum_{i=1}^n (x_i z_{ij}) - \frac{1}{n} \sum_{\ell=1}^n x_\ell \right)^\top \quad (5)$$

$$= X M Z C^{-1} Z^\top M X^\top, \quad (6)$$

$$S_w = \sum_{j=1}^k \left( \sum_{i=1}^n z_{ij} (x_i - \frac{1}{c_j} \sum_{\ell=1}^n x_\ell z_{\ell j}) (x_i - \frac{1}{c_j} \sum_{\ell=1}^n x_\ell z_{\ell j})^\top \right) \quad (7)$$

$$= X M (I - Z C^{-1} Z^\top) M X^\top, \quad (8)$$

で定義される [56]．ただし、 $I$  は単位行列、 $M = I - \frac{1}{n} \mathbf{1}_{n \times n}$  は平均化行列、 $c_j = \sum_{i=1}^n z_{ij}$  は  $j$  番目のクラスサイズを意味しており、 $C_{k \times k} = \text{diag}([c_1, \dots, c_k])$  と定義した．

このときクラス間分散とクラス内分散の和として総分散  $S_t = \sum_i (x_i -$

$\frac{1}{n} \sum_{\ell=1}^n x_\ell)(x_i - \frac{1}{n} \sum_{\ell=1}^n x_\ell)^\top = S_w + S_b$  を定義すると、分離度を最大にする最適な射影ベクトル  $a^*$  は、

$$a^* = \arg \max_a \frac{a^\top S_b a}{a^\top S_w a} = \arg \min_a \frac{a^\top S_w a}{a^\top S_t a} = \arg \min_a \frac{a^\top X M (I - Z C^{-1} Z^\top) M X^\top a}{a^\top X M I M X^\top a}, \quad (9)$$

によって与えられる。この最適化は一般固有値問題

$$X M (I - Z C^{-1} Z^\top) M X^\top a = \lambda X M M X^\top a, \quad (10)$$

を解くことで、 $a^*$  は最小固有値に対応する固有ベクトルとして求めることができる。また、分離度  $\sigma$  は  $\frac{1}{\lambda} - 1$  で計算できる。

式 (10) の固有値が 1 未満の固有ベクトルは  $\text{rank}(S_b) = k - 1$  個存在する。これらに対応する固有ベクトルの行列  $A^*$  を用いると、分離度が最大となる  $r = k - 1$  次元へサンプルデータを削減することができる [56]。

### 3.3 ラプラシアン固有マップ法

ラプラシアン固有マップ法 (Laplacian Eigenmaps: LE) は、重みづけ総分散を固定した上で類似度がデータ同士を近づけるように次元削減をする [57]。類似度行列  $W \in [0, 1]^{n \times n}$  は、データ  $x_i$  と  $x_j$  が似ているほど  $w_{ij}$  が 1 に近づき、似ていないほど 0 に近づく。また、データ同士の類似度には対称性を仮定する、 $w_{ij} = w_{ji}$ 。このとき、次元削減データは重みづけ総分散が一定である条件、 $y^\top D y = 1$ 、で

$$\arg \min_y \frac{1}{2} \sum_{i,j=1}^n w_{ij} \|y_i - y_j\|_2^2 = \arg \min_y y^\top (D - W) y, \quad (11)$$

を満たすように与えられる。ただし、 $d_i = \sum_{\ell=1}^n w_{i\ell}$  であり、 $D = \text{diag}(W 1_n)$  と定義した。サンプルデータ  $x_i$  の周辺にデータが密集するほど  $d_i$  は高くなるため、 $d_i$  は局所的な密度として解釈できる [58]。

LE は射影ベクトル  $a^*$  を計算する代わりに、次元削減後のデータ  $y^* = X^\top a^*$  を、目的



関数

$$y^* = \arg \min_y \frac{y^\top (D - W)y}{y^\top D y}, \quad (12)$$

を最小化することで直接計算する．DA と同様に， $y^*$  は一般固有値問題

$$(D - W)y = \lambda D y, \quad (13)$$

の最小固有値に対応する固有ベクトルとなる．式 (13) は自明な解として， $y = 1_n$  のとき最小固有値  $\lambda = 0$  となるが，この次元削減に意味はない．そのため，LE では一般的に，2 番目から  $k$  番目の最小固有値に対応する固有ベクトルを次元削減データとして使用する [57]．

類似度に基づいて直接次元削減データを求めるのではなく，データ分布  $X$  や  $K$  を用いて次元削減する手法は，Locally Preserving Projection(LPP) と呼ばれる [41]．LPP も DA や LE と同様に，目的関数

$$\frac{a^\top X(D - W)X^\top a}{a^\top XDX^\top a}, \quad (14)$$

を最小化するように，一般固有値問題の固有ベクトルが射影ベクトルとして与えられる．

### 3.4 判別的ラプラシアン固有マップ法

類似度をクラスラベルの代わりに用いることで，近似したクラス内分散とクラス間分散を，

$$\tilde{S}_b = XWX^\top = \sum_{j=1}^{k'} \omega_j \left( \frac{1}{\omega_j} \sum_{i=1}^n x_i v_{ij} \right) \left( \frac{1}{\omega_j} \sum_{i=1}^n x_i v_{ij} \right)^\top, \quad (15)$$

$$\tilde{S}_t = XDX^\top = \sum_{i=1}^n d_i x_i x_i^\top, \quad (16)$$

$$\tilde{S}_w = X(D - W)X^\top = \tilde{S}_t - \tilde{S}_b, \quad (17)$$

と定義することで，一見すると LE と LPP は近似した分離度を最大化しているように見える．ただし， $v_{ij}$  と  $\omega_j$  はそれぞれ，類似度  $W$  の固有ベクトルの要素と固有値の逆数であり，類似度  $W$  のランク  $k'$  はクラス数と解釈できる．しかしながら，LE と LPP はサンプル

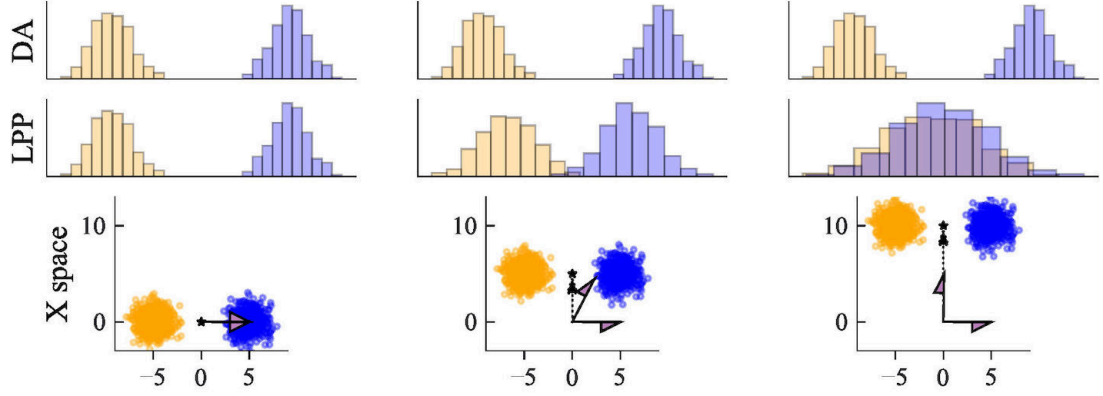


図 6: 上段の 2 つの図は、理想的な類似度  $ZC^{-1}Z^T$  を使用した DA と LPP の最適な次元削減データ (横軸) のヒストグラムを示している。下段の図は、青と黄色のクラスターで構成される 2 次元のサンプルデータの例を散布図で示している。右の図にいくほどサンプルデータは平均が、縦方向に大きくなる。破線の矢印は全データの平均ベクトルを表し、左半分と右半分に分かれた矢印はそれぞれ LPP と DA の最適な射影ベクトルを表している。

データ  $X$  の平均が  $0_n$  でない場合には分離度を最大化するように次元を削減することはできない。図.6 は、DA と LPP の最適な次元削減  $y^*$  の結果を示している。LPP で使用する類似度は  $W = ZC^{-1}Z^T$  としており、LPP と DA の差はサンプルデータ  $X$  を平均化するか否かにある。サンプルデータの平均ベクトル  $\bar{x}$  が 0 からずれるほど、LPP による最適射影ベクトル  $a^*$  は  $\bar{x}$  の方向を向くように、分離度を最大化する方向からずれていく。LE の場合では、式 (13) が  $n$  個の直交した解を持つことから、 $\bar{x}$  は常に  $0_n$  でない。そのため、LE による次元削減は分離度を最大化する射影方向からずれた結果となる。仮に、 $\bar{x} = 0_n$  が成り立つ場合には、LE の次元削減データの次元は  $n - 1$  個になる。

そこで、サンプルデータ  $X$  を平均化するように近似したクラスター間分散とクラスター内を

それぞれ,

$$\hat{S}_b = \sum_{j=1}^{k+1} \omega_j \left( \frac{1}{\omega_j} \sum_{i=1}^n x_i v_{ij} - \frac{1}{n} \sum_{\ell=1}^n x_\ell \right) \left( \frac{1}{\omega_j} \sum_{i=1}^n x_i v_{ij} - \frac{1}{n} \sum_{\ell=1}^n x_\ell \right)^\top \quad (18)$$

$$= XMWMX^\top, \quad (19)$$

$$\hat{S}_w = \sum_{j=1}^n \left( \sum_{i=1}^n v_{ij} \left( x_i - \frac{1}{\omega_j} \sum_{\ell=1}^n x_\ell v_{\ell j} \right) \left( x_i - \frac{1}{\omega_j} \sum_{\ell=1}^n x_\ell v_{\ell j} \right)^\top \right) \quad (20)$$

$$= XM(D - W)MX^\top, \quad (21)$$

と定義しなおすことで, 類似度で近似した分離度を最大化する判別的ラプラシアン固有マップ法 (Discriminant Laplacian Eigenmaps:DLE) を提案する. 重みつき総分散  $\hat{S}_t$  とこの近似したクラス内分散とクラス間分散の関係は, DA と同様に,

$$\hat{S}_t = \hat{S}_b + \hat{S}_w \quad (22)$$

を満たす. そのため, DLE による近似した分離度を最大化する  $y^*$  は,

$$y^* = \arg \min_y \frac{y^\top M(D - W)My}{y^\top MDMy}, \quad (23)$$

によって与えられ, 一般固有値問題

$$M(D - W)My = \lambda MDMy. \quad (24)$$

の最小固有値に対応する固有ベクトルとして与えられる. 平均化行列のランクは  $n - 1$  であるため, 式 (24) の右辺  $MDM$  は特異行列となり, 直接固有値ベクトルを求めることはできない. そこで, 平均化行列を  $M = FI_{(n-1) \times (n-1)}F^\top$  と固有値分解し,  $s = F^\top y$  を一般固有値問題

$$F^\top (D - W)Fs = \lambda F^\top DFs, \quad (25)$$

の最小固有値に対応する固有ベクトルとして  $s$  を求める. そして, 平均化した次元削減データ  $\bar{y} = My = Fs$  で計算する. DLE は LE と異なり, 式 (25) は自明な解を持たないため, 1 番目から  $k - 1$  番目の最小固有値に対応する固有ベクトルが次元削減データとなる.

### 3.5 類似度の低次元化処理

DLE は近似分離度を最大化するため、 $W$  のランクはクラスタ数を意味する。  $W$  のランクが実際のクラスタ数より大きくなると、類似度から擬似クラスタが生成され、DLE の弱点となる。 擬似クラスタには 2 種類あり、1 つ目は実際のクラスタの部分的なクラスタある。 そのようなクラスタ間分散を最大化すると、実際のクラスタ内分散を最大化することになる。 もう 1 つの擬似クラスタは、異なる実際のクラスタの複合型であり、そのクラスタ内分散を最小化することで、実際のクラスタ間分散を最小化することになる。 結果として、DLE は擬似クラスタ間の誤った分離度を最適化する可能性がある。 この問題を解決する安直な案として  $W$  の固有値分解や特異値分解によって  $W$  のランクを下げる方法が考えられる。 しかしながら、 $W$  の固有値が大きい方から  $k$  個の成分を取り出すような再構成された類似度  $H = V_{n \times k} \Omega_k V^\top$  は、再構成前の局所密度  $D = \text{diag}(W 1_n) \neq \text{diag}(H 1_n)$  を維持できない。

そこで、局所密度を保持したまま、数値的に類似度のランクを下げる方法を提案する。  $\Lambda$  を式 (13) の固有値  $\lambda$  の対角行列とすると、式 (13) は制約条件  $YDY^\top = I$  のもとでの  $WY = DY(I - \Lambda)$  と書き直せ、 $YY^\top = D^{-1}$  と  $YWY^\top = (I - \Lambda)$  が成り立つ。 したがって、類似度行列は  $W = DY(I - \Lambda)Y^\top D$  と書き直せる。  $(I - \Lambda)$  の対角成分は  $[0, 1]$  の範囲に収まるため、

$$DY(I - \Lambda)^2 Y^\top D = WD^{-1}W \quad (26)$$

で対角成分を累乗することで、数値誤差として類似度  $W$  のランクを下げるができる。 この操作では、局所密度  $\text{diag}(WD^{-1}W 1_n) = D$  は元の類似度  $W$  と理論的に一致する。 この操作は再帰的に実行でき、 $H_0 = W$  とすることで、

$$H_{r+1} = H_r D^{-1} H_r, \quad (27)$$

により局所密度  $\text{diag}(H_r 1_n) = D$  を保持しながら、類似度のランクを下げられる。

式 (27) の導出から明らかなように、LE の固有問題はどのような  $H_r$  に対しても同時対角化可能である。 つまり、式 (27) の類似度の低次元化によって、LE の固有ベクトルが変化することはないため、この再帰的な方法では LE の次元削減の結果は変化しない。 一方、DLE

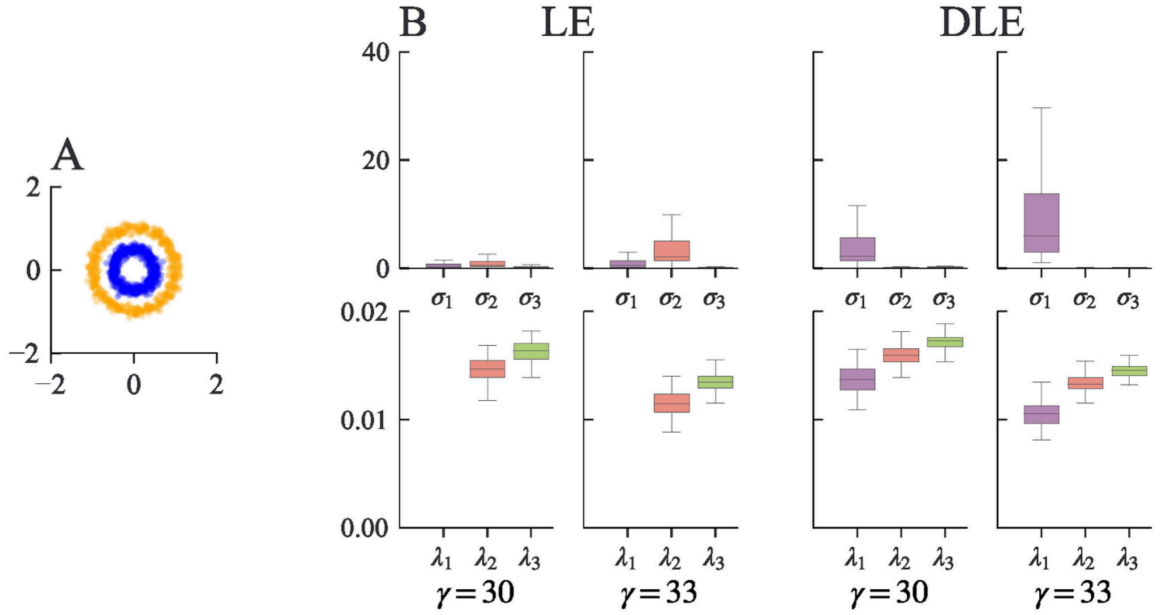


図 7: (A) 実験に使用したサンプルデータを散布図で描いており、100 回再サンプリングして LE と DLE による次元削減データの分離度を評価した．(B) 上段の箱ひげ図は、目的関数の最小固有値に対応する第 1 から第 3 固有ベクトルに対する LE および DLE の実際の分離度を、昇順に示している．下段はそれらに対応する固有値を示している．

は平均化により同時対角化が不可能な場合があるため、この再帰的手法により次元削減の結果が変化することが予想される．

### 3.6 平均化と低次元処理の効果検証

LE と DLE を比較することで、サンプルデータの平均化と類似度の再帰的处理の効果を検証する数値実験を行った．実験に用いたサンプルデータは、ガウシアンノイズを持つ 2 つの円の入れ子からなる人工的なものであり（図 7A）、100 回リサンプリングした．類似度  $w_{ij}$  として RBF 関数  $\exp\{-\gamma\|x_i - x_j\|_2^2\}$  を  $\gamma = 30, 33$  で用い、1 番目から 3 番目までの最小固有値 ( $\lambda_1, \lambda_2, \lambda_3$ ) に対応する次元削減データの実際の分離度を評価した．

図 7B は、DLE が安定的に 1 次元に削減でき、分離度を最大化できることを示している．一方、LE は 2 番目に固有値が小さい固有ベクトルで分離度を上げるように次元削減ができている．一方、理論的に自明な解である  $\lambda_1 = 0$  で  $y = 1_n$  となる理論的に自明な解に反して、 $\sigma_1 \neq 0$  となった．この理論と実験の不整合は、LE の解の曖昧性にある．簡単な例とし

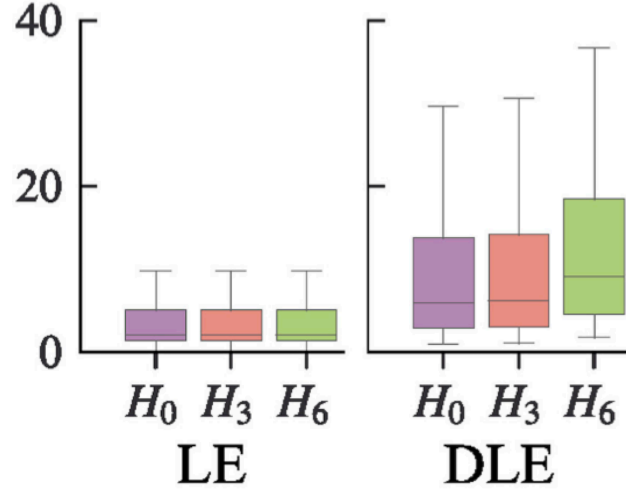


図 8: DLE の第 1 解と LE の第 2 解に対して,  $\gamma = 33$  を用いた類似度の再帰的处理による実際の分離度を箱ひげ図で示している.

て,  $W = ZC^{-1}Z^{\top}$  という理想的な類似度の場合を想定する. この場合, 同じクラスタ内のデータ間の類似度のみが正で, それ以外の類似度は全て 0 である. このとき, LE の最適解は  $W$  の固有ベクトルと一致し, 可能な固有ベクトルのうち 1 つは  $ZC^{-\frac{1}{2}}$  で, そのとき固有値は  $I_k$  である. また, この解  $ZC^{-\frac{1}{2}}$  を回転行列  $R_k$  で回転した任意の行列  $Y = ZC^{-\frac{1}{2}}R$ , 回転行列の性質  $RR^{\top} = I_{k \times k}$  から,  $YI_{k \times k}Y^{\top} = ZC^{-\frac{1}{2}}RR^{\top}C^{-\frac{1}{2}}Z^{\top} = ZC^{-1}Z^{\top}$  となるため LE の最適解となる. 実験では, 理想的な類似度を用いていないが,  $\lambda_2$  と  $\lambda_3$  が 0 に近づくことで, LE の  $\lambda_1$  に対応する固有ベクトルが  $1_n$  からズレていると考えられる.

図 8 より, 類似度の低次元化の処理を繰り返すことで, DLE による最適な次元削減データの分離度が改善されることがわかる. 一方, LE では  $H_r$  に対する同時対角化可能性によって, 理論的にも数値上も変化がない. LE に平均化の制限を新たに加えた DLE は分離度を最大にすように次元削減ができていることに加え, LE の目的関数の次元削減に役に立たない自明な解を取り除くのに有効である.

## 4 環境推定によるアプローチ

本章では、観測データがフルランクになるような、観測が待つ情報を低次元で表現できない次元削減によるアプローチは適切ではない状態削減問題に取り組む。このような状態削減問題は観測の生成過程によって生じる。特に、一般の強化学習手法が仮定する部分観測性ではなく、観測の生成過程に冗長観測性を仮定した状態削減問題を考える。ブロック崩しの例では、ボールを直接ブロックに当てるのも、壁に跳ね返してブロックに当てるのも、ブロックを崩すという目的に対してボールの軌道は区別する必要がない。つまり、コア状態にはボールの軌道の情報は含まれない。他の例として、カードの合計値を 21 に近づけるようにその大きさを競う BlackJack では、観測はカードの組み合わせであるが、最適な戦略を取るには合計値をコア状態とみなすことができる。報酬の冗長性による複雑化に対しては、環境ダイナミクスに対する最適戦略の観点から状態を削減する必要がある。

報酬に無関係な観測が多数存在する環境における強化学習のために、冗長観測マルコフ決定過程 (Redundantly Observable MDP: ROMDP) を定式化する。このような環境では、観測は一般に報酬予測に必要なコア状態を定義する報酬に関連する観測と報酬に無関係な観測に分類される。非定常 ROMDP における継続的な強化学習タスクに対処するため、コア空間を推定する確率モデルとして目的指向な環境推論 (Goal-Oriented Environment Inference: GOEI) を提案し、GOEI により推定された状態空間構造に基づく行動制御に行動価値 Thompson Sampling (Action-value Thompson Sampling: ATS) を用いる。GOEI と ATS の組み合わせにより、状態空間を推定すると同時に、学習した知識を保持しながら最適な戦略を学習できることを示す。

### 4.1 冗長観測性と部分観測性によるアプローチ

状態削減問題の条件 (3) は、現在の状態に基づく行動依存なマルコフ連鎖によって将来の報酬が予測できるように、行動と報酬そして状態の関係に制約を与える。このような制約は、各状態で最適な行動を選択するために必要である。しかし、式 (3) では、状態と観測の関係は規定されていない。全ての可能性の中から適切な関係を推測することは困難であ

るため、エージェントは状態削減問題 (3) を実際に解くために、状態と観測の間に特定の関係を仮定する必要がある。言い換えれば、エージェントは状態と観測に関する特定の仮定に基づいた近似アプローチをとる。

状態と観測に関する 2 つの単純な仮定を考える。最初のアプローチは本論文の焦点であり、次のように記述される、

[冗長観測性]

$$p(s_t | \mathcal{A}_{t-1}, \mathcal{R}_t, \mathcal{O}_t) \approx P(s_t | o_t), \quad (28)$$

は、冗長な観測結果から報酬予測に関連する情報のみを抽出する状態削減を目的とする。これは式 (3) の状態削減モデル  $p(s_t | \mathcal{A}_{t-1}, \mathcal{R}_t, \mathcal{O}_t)$  をより単純な依存関係  $P(s_t | o_t)$  に置き換えることを仮定する。この仮定は、現在の状態  $s_t$  が現在の観測  $o_t$  を通してのみ過去に依存することを意味する。言い換えれば、 $o_t$  は  $s_t$  を決定するのに十分な情報を持っていることを意味し、 $o_t$  が  $s_t$  に無関係な情報を含むことを許容する。したがって、このアプローチにおけるエージェントは、冗長な観測から報酬の予測に不必要な情報を除外するような状態削減を目指す。

現代の実世界のタスクでは、観測値は多くのセンサーによって収集されることが多く、それゆえ多くの不要な情報を含むことがある。観測値が冗長な環境では、全ての観測値を表現する状態集合を見つけることはメモリコストがかかり経済的ではない。最初のアプローチは、現在の観測値  $o_t$  から報酬に無関係な情報を除外することで最小の状態集合を見つけようとするもので、実世界のタスクに対して経済的でメモリ効率の良い解を与えられられる。

一方、最初のアプローチ (28) は、 $s_t$  を決定するには現在の観測で十分であると仮定することで、過去の履歴  $\mathcal{A}_{t-1}, \mathcal{R}_t, \mathcal{O}_{t-1}$  から有用な情報を抽出するための状態削減を放棄している。しかし、この仮定が不適切で、状態の過去系列への依存性を探索することが状態削減のために必要な状況もある。典型的な例は、 $o_t$  が  $s_t$  を決定するのに十分な情報を持っていない状況である。このような状況に対処するために、状態削減における  $s_t$  と  $o_t$  の役割を交換する逆関係を導入することができる：



[部分観測性]

$$p(o_t | \mathcal{A}_{t-1}, \mathcal{R}_t, \mathcal{O}_{t-1}, s_t) \approx P(o_t | s_t). \quad (29)$$

この仮定は、現在の観測値  $o_t$  は  $s_t$  を通してのみ過去に依存することを意味する．言い換えれば、エージェントは、現在の状態に関する部分的な情報しか観測できないと仮定して、 $o_t$  を決定するのに十分な情報を持つように状態  $s_t$  を決定する．

式 (3) の状態削減モデル  $p(s_t | \mathcal{A}_{t-1}, \mathcal{R}_t, \mathcal{O}_t)$  は部分観測性の仮定 (29) の式から直接導出することはできないが、ベイズ推論によって経験  $\mathcal{A}_{t-1}, \mathcal{R}_t, \mathcal{O}_t$  の履歴からその近似を導出することができる．まずは、部分観測性の仮定 (29)、式 (3) の確率分布を全てベイズ推論するための生成モデルを構築する．そして、式 (3) の確率分布と仮定 (29) を用いて、行動系列  $\mathcal{A}_{t-1}$  による  $\mathcal{S}_t, \mathcal{R}_t, \mathcal{O}_t$  の生成モデルは、

$$p(\mathcal{S}_t, \mathcal{R}_t, \mathcal{O}_t | \mathcal{A}_{t-1}) = \prod_{\tau=1}^t P(o_\tau | s_\tau) P(r_\tau | s_\tau) P(s_\tau | a_{\tau-1}, s_{\tau-1}) P(s_0), \quad (30)$$

となる．ただし、 $\mathcal{S}_t \equiv [s_t, s_{t-1}, \dots]$  である．この生成モデルにベイズの定理を適用することで、状態系列  $\mathcal{S}_t$  の確率分布とこれらの状態間の遷移モデル  $P(s_\tau | a_{\tau-1}, s_{\tau-1})$ 、報酬モデル  $P(r_\tau | s_\tau)$ 、そして観測モデル  $P(o_\tau | s_\tau)$  を現在の既知な経験の履歴から同時に推定する．状態削減モデルは推論された分布の一部であり、

$$p(s_t | \mathcal{A}_{t-1}, \mathcal{R}_t, \mathcal{O}_t) = \frac{\sum_{\mathcal{S}_{t-1}} p(\mathcal{S}_t, \mathcal{R}_t, \mathcal{O}_t | \mathcal{A}_{t-1})}{\sum_{\mathcal{S}_t} p(\mathcal{S}_t, \mathcal{R}_t, \mathcal{O}_t | \mathcal{A}_{t-1})}, \quad (31)$$

で表される．このように、2つ目のアプローチでは、現在の既知な経験  $\mathcal{A}_{t-1}, \mathcal{R}_t, \mathcal{O}_t$  から抽出された重要な情報を、有限な状態集合における現在の状態  $s_t$  に削減する．しかし、仮定 (29) は、推論された状態に観測の予測に役立つあらゆる情報を含めるようエージェントを促すため、このアプローチでは報酬予測に無関係な冗長な情報を除外することはできない．

この生成モデル (30) は従来の POMDP と同一であり、第二のアプローチで行われる状態削減は POMDP のそれと等価である．POMDP はもともと非マルコフ環境として導入されたが、本論文では POMDP を状態削減問題を解くアプローチとして解釈する．この視点において、生成モデル (30) はエージェントがモデル化した仮定を表すものであり、従来の視点のような環境設定を表すものではない．このような違いを考慮して、第二のアプローチを

「POMDP アプローチ」と呼ぶことにし、一つ目のアプローチを「ROMDP アプローチ」と呼ぶ。

まとめると、POMDP と ROMDP は状態削減問題 (3) に対する異なる近似アプローチを記述している。状態削減問題 (3) を直接解くことは非常に困難であるため、様々な実用的な仮定が必要である。ROMDP と POMDP のアプローチは、それぞれ冗長な観測を持つ環境と部分的に観測可能な環境に対して有効である。部分的に観測可能な状態と冗長な状態の両方を含む、より一般的な環境では、POMDP と ROMDP のアプローチを組み合わせることができる。本論文では、一般的な環境での状態削減に取り組むための足掛かりとして、ROMDP アプローチに焦点を当てる。

以降の節では、これらのアプローチの方法論的な詳細に焦点を当てる。前述のように、POMDP アプローチは報酬予測に不要な情報が含まれていても、すべての観測を表現できるように状態を削減する。そのため、POMDP アプローチのエージェントを「完全な環境推論 (Complete Environment Inference: CEI)」と呼ぶことにする。対照的に、ROMDP アプローチは報酬予測に重要な情報のみを抽出するように機能し、ROMDP アプローチのエージェントを「目的指向な環境推論 (GOEI)」と呼ぶ。

## 4.2 変分ベイズによる環境推定

エージェントは環境から得られた情報の因果関係をモデル化し、環境ダイナミクスを推論するしなければならない。ベイズ推論は、情報が与えられるたびに事前分布を真の環境動態に近づけるために事後分布を更新する枠組みを提供する。変分ベイズ推論は、複雑なモデルの事後分布を近似するための方法である。

モデルの予測変数を  $D$ 、与えられた条件を  $\bar{D}$ 、パラメータを  $\Omega$  とする。それぞれの定義について、以降の章で定義するエージェントモデルと事前分布をそれぞれ  $P(s, D \mid \Omega, \bar{D})$ 、 $P(\Omega)$  と表す。変分ベイズ (VB) は、隠れ状態  $s$  とパラメータ  $\Omega$  の真の事後分布が独立していると仮定する。次に、自由エネルギー (FE) は環境推論モデル  $P(s, \Omega, D \mid \bar{D}) =$

$P(s, D \mid \Omega, \bar{D})$ ,  $P(\Omega)$  および近似事後分布  $Q(s)$ ,  $Q(\Omega)$  に対して定義される.

$$\text{FE} = \sum Q(s) Q(\Omega) \ln \frac{Q(s) Q(\Omega)}{P(s, \Omega, D \mid \bar{D})} = \text{KL} \left[ Q(s) Q(\Omega) \parallel P(s, \Omega \mid D, \bar{D}) \right] - \ln P(D \mid \bar{D}). \quad (32)$$

ただし,  $\text{KL}[\cdot \parallel \cdot]$  は Kalback-Leibler (KL) ダイバージェンスを意味する.

FE を最小化することは, 近似事後分布と真の事後分布の KL ダイバージェンスを最小化することと同等である. 変分定理により,  $Q(s)$  が与えられたときに FE を最小化する E ステップ (式 (33)) と,  $Q(\Omega)$  が与えられたときに FE を最小化する M ステップ (式 (34)) を定義できる. 変分ベイズでは, FE を連続的に減少させるために, E ステップと M ステップを交互に繰り返す.

$$Q^{(i)}(s) \propto \exp \left[ \mathbb{E} \left[ \ln P(s, D \mid \Omega, \bar{D}) \right]_{Q^{(i)}(\Omega)} \right], \quad (33)$$

$$Q^{(i+1)}(\Omega) \propto \exp \left[ \mathbb{E} \left[ \ln P(s, D \mid \Omega, \bar{D}) \right]_{Q^{(i)}(s)} \right] P(\Omega). \quad (34)$$

本稿では, 前回の学習で得られた事後分布を VB の初期条件として使用した:  $Q_{\tau}^{(0)}(\Omega) = P_{\tau}(\Omega) = Q_{\tau-T}^{(\infty)}(\Omega)$ .

### 4.3 棒折過程によるノンパラメトリック推定

エージェントモデルを定義するためには, 非定常環境動態の数やそれを構成する状態の数などのコンポーネント数を事前に知っておく必要がある. ディリクレ過程は理論上, 無限のコンポーネントを扱うことができるため, コンポーネントの数に関する事前知識に依存しないエージェントモデルの定義が可能となる. 棒折過程 (Stick Breaking Process: SBP) は, 変分ベイズ法を用いてディリクレ過程を計算する方法であり, 無限のコンポーネントを複数のベータ分布で表現する.

サンプル分布  $\bar{F} = [F^1, \dots]$  とし, 基底分布  $B_F$  のディリクレ過程を以下のように SBP を

用いて定義する：

$$P(B_F) = \prod_w^\infty P(F^w; \Phi^{F^w}) \text{SBP}(W; \Phi^W), \quad (35)$$

$$P(v_w) = \text{Beta}\left(v_w \left| 1 + \Phi_w^W, \alpha_w + \sum_{w'=w+1}^\infty \Phi_{w'}^W \right.\right), \quad (36)$$

$$\text{SBP}(W = w; \alpha_w, \Phi^W) = P(v_w) \prod_{k=1}^{w-1} (1 - P(v_k)). \quad (37)$$

ここで,  $\alpha_w > 0$  は SBP のハイパーパラメータであり,  $W$  は割り当て指標  $w$  の確率である. 一般に SBP では交換可能性が成り立たない問題への対策として,  $w$  はハイパーパラメータ  $\Phi^W$  の降順にソートされ, その後に式 (36) および式 (37) が適用される.

M ステップでは, SBP 事後分布のハイパーパラメータ  $\Theta^W$  が次の式を用いて更新される：

$$\begin{aligned} \sum_{t=\tau-T}^\tau Q(F_t) \ln \text{SBP}(W = w; \Phi^W) = \\ \sum_{t=\tau-T}^\tau Q(F_t = F^w) \ln P(v_w) + \sum_{w'=w+1}^\infty \sum_{t=\tau-T}^\tau Q(F_t = F^{w'}) \ln(1 - P(v_{w'})). \end{aligned} \quad (38)$$

さらに, E ステップでは隠れ変数の期待値を次の式を用いて計算する：

$$\mathbb{E}[\ln \text{SBP}(W = w)]_{Q(B_F)} = [\psi(\Theta_w^V) - \psi(\Theta_w^V + \bar{\Theta}_w^V)] \sum_{k=1}^{w-1} [\psi(\bar{\Theta}_k^V) - \psi(\Theta_k^V + \bar{\Theta}_k^V)]. \quad (39)$$

ただし, 記法を簡略化するために,  $1 + \Theta_w^W = \Theta_w^V$ ,  $\alpha_F + \sum_{w'=w+1}^\infty \Theta_{w'}^W = \bar{\Theta}_w^V$  としている.

#### 4.4 完全な環境推論

まず, POMDP アプローチを達成するための標準アルゴリズムの詳細を説明する [17, 18]. 環境の変化の可能性を考慮して, 報酬モデルおよび遷移モデルのパラメータのマルチモジュールバージョンを導入した. 具体的な手順を説明するために, 履歴  $\mathcal{S}$ ,  $\mathcal{O}$ ,  $\mathcal{R}$  および  $\mathcal{A}$  を固定された区間  $T$  で再定義した.

現在時刻  $\tau$  を考慮し, 過去の状態の系列を  $\mathcal{S} = [s_{\tau-T}, \dots, s_\tau]$ , 観測を  $\mathcal{O} = [o_{\tau-T}, \dots, o_\tau]$ , 報酬を  $\mathcal{R} = [r_{\tau-T}, \dots, r_\tau]$ , 行動を  $\mathcal{A} = [a_{\tau-T-1}, \dots, a_{\tau-1}]$ , 遷移モデ

ルモジュールのパラメータを  $\tilde{M} = [M_{\tau-T}, \dots, M_\tau]$ , および報酬モデルモジュールのパラメータを  $\tilde{N} = [N_{\tau-T}, \dots, N_\tau]$  とする. これらの系列に対して, エージェントモデルは,

$$P(\mathcal{O}, \mathcal{R}, \mathcal{S} | \mathcal{A}, s_{\tau-T-1}, L, \tilde{M}, \tilde{N}) = \prod_{t=\tau-T}^{\tau} P(o_t | s_t, L) P(s_t | s_{t-1}, a_{t-1}, M_t) P(r_t | s_t, N_t), \quad (40)$$

で定式化される. ここで,  $L, M_t$  および  $N_t$  は, それぞれ観測モデル  $P(o_t | s_t)$ , 遷移モデル  $P(s_t | s_{t-1}, a_{t-1})$ , および報酬モデル  $P(r_t | s_t)$  のパラメータである. さらに,  $M_t \in \bar{M} = \{M^1, \dots\}$  および  $N_t \in \bar{N} = \{N^1, \dots\}$  は, それぞれハイパーパラメータ ( $\alpha_M > 0$  および  $\alpha_N > 0$ ) と基底分布 ( $B_M$  および  $B_N$ ) を持つディリクレ過程 [59] に従う.

ベイズ推論の目的は, エージェントがすでに取得した過去の系列  $[\mathcal{O}, \mathcal{R}, \mathcal{A}]$  から, 事後分布  $P(\mathcal{S}, \tilde{M}, \tilde{N}, L, B_M, B_N | \mathcal{O}, \mathcal{R}, \mathcal{A})$  を得ることである. しかし, 事後分布を解析的に得るのは困難であるため, 変分ベイズ法 [60, 61] を用いて,

$$Q(\mathcal{S}, \tilde{M}, \tilde{N}) Q(L, B_M, B_N) \simeq P(\mathcal{S}, \tilde{M}, \tilde{N}, L, B_M, B_N | \mathcal{O}, \mathcal{R}, \mathcal{A})$$

で近似事後分布を計算する.

行動, 観測, および報酬が離散的である場合 (すなわち, one-hot ベクトル  $a, o, r, s$  として表される場合), パラメータ  $L, M_t$  および  $N_t$  はそれぞれカテゴリカル分布に従う:

$$P(o_t | s_t, L) = \text{Cat}(o_t; L[s_t]), \quad (41)$$

$$P(s_t | s_{t-1}, a_{t-1}, M_t) = \text{Cat}(s_t; M_t[s_{t-1}, a_{t-1}]), \quad (42)$$

$$P(r_t | s_t, N_t) = \text{Cat}(r_t; N_t[s_t]). \quad (43)$$

この場合, これらのパラメータの事前分布は, ハイパーパラメータ  $\Phi^L, \Phi^{M^x}, \Phi^{N^z}, \alpha_x, \Phi^X, \alpha_z$  および  $\Phi^Z$  をそれぞれ持つ独立分布である:

$$\begin{aligned} P(L, B_M, B_N) &= P(L) P(\bar{M}, X) P(\bar{N}, Z) \\ &= \text{Dir}(L; \Phi^L) \prod_x^{\infty} \text{Dir}(M^x; \Phi^{M^x}) \text{SBP}(X; \alpha_x, \Phi^X) \prod_z^{\infty} \text{Dir}(N^z; \Phi^{N^z}) \text{SBP}(Z; \alpha_z, \Phi^Z). \end{aligned} \quad (44)$$

ここで,  $X$  および  $Z$  は, SBP に従う割り当て指標  $x \sim \text{Cat}(X)$  および  $z \sim \text{Cat}(Z)$  の確

率であり、基底分布  $(B_M, B_N)$  はサンプル分布  $(\bar{M}, \bar{N})$  と割り当て指標  $(X, Z)$  に分解される。このとき、パラメータの近似事後分布もまた、ハイパーパラメータ  $\Theta^L, \Theta^{M^x}, \Theta^{N^z}, \Theta^x$  および  $\Theta^z$  を持つ独立分布の積として与えられる。

$$\begin{aligned}
Q(\Omega_{\text{CEI}}) &\equiv Q(L, B_M, B_N) = Q(L, \bar{M}, \bar{N}, X, Z) \\
&= Q(L) \prod_x^\infty Q(M^x) \prod_z^\infty Q(N^z) Q(X) Q(Z) \\
&= \text{Dir}(L; \Theta^L) \prod_x^\infty \text{Dir}(M^x; \Theta^{M^x}) \text{SBP}(X; \alpha_x, \Theta^X) \prod_z^\infty \text{Dir}(N^z; \Theta^{N^z}) \text{SBP}(Z; \alpha_z, \Theta^Z).
\end{aligned} \tag{45}$$

CEI は報酬  $r$  と観測  $o$  を予測するが、行動  $a$  は予測しない。要約すると、CEI の自由エネルギーは、環境推論モデルに対して式 (40) および (44) を式 (32) に代入し、 $D = \{\tilde{r}, \tilde{o}\}$ ,  $\bar{D} = \{\tilde{a}\}$ , および  $\Omega = \{L, B_M, B_N\}$  と設定することによって得られる。

CEI の M ステップは、エージェントモデル (40) を自由エネルギー (FE) の式 (34) に代入することで導出される：

$$\begin{aligned}
Q(L, B_M, B_N) = \exp \left[ \sum_{t=\tau-T}^{\tau} Q(s_t, M_t, N_t) \left\{ \ln L_{o_t s_t} + \ln M_{s_t s_{t-1} a_{t-1}}^{x_t} \right. \right. \\
\left. \left. + \ln N_{r_t s_t}^{z_t} + \ln \text{SBP}(x_t) + \ln \text{SBP}(z_t) \right\} \right] P(L, B_M, B_N). \tag{46}
\end{aligned}$$

この時、ハイパーパラメータの更新則は

$$\Theta_{os}^L \leftarrow \Phi_{os}^L + \sum_{t=\tau-T}^{\tau} Q(s_t=s) \mathbb{1}(o_t, o), \tag{47}$$

$$\Theta_{s'sa}^{M^x} \leftarrow \Phi_{s'sa}^{M^x} + \sum_{t=\tau-T}^{\tau} Q(s_t=s', s_{t-1}=s, M_t=M^x) \mathbb{1}(a_{t-1}, a), \tag{48}$$

$$\Theta_{rs}^{N^z} \leftarrow \Phi_{rs}^{N^z} + \sum_{t=\tau-T}^{\tau} Q(s_t=s, N_t=N^z) \mathbb{1}(r_t, r), \tag{49}$$

$$\Theta_x^X \leftarrow \Phi_x^X + \sum_{t=\tau-T}^{\tau} Q(M_t=M^x), \tag{50}$$

$$\Theta_z^Z \leftarrow \Phi_z^Z + \sum_{t=\tau-T}^{\tau} Q(N_t=N^z), \tag{51}$$

となる．ただし， $o_t = o$  の時  $1(o_t, o) = 1$  となり，それ以外の時は 0 となる．

式 (40) を式 (33) に代入することで，E ステップは

$$Q(\tilde{s}, \tilde{M}, \tilde{N}) = \prod_{t=\tau-T}^{\tau} Q(s_t, M_t, N_t | s_{t-1}), \quad (52)$$

$$Q(s_t, M_t, N_t | s_{t-1}) \propto \exp \left\{ \mathbb{E}[\ln L_{o_t s_t}]_{Q(L)} + \mathbb{E}[\ln M_{s_t s_{t-1} a_{t-1}}^{x_t}]_{Q(M^x)} \right. \\ \left. + \mathbb{E}[\ln N_{r_t s_t}^{z_t}]_{Q(N^z)} + \mathbb{E}[\ln \text{SBP}(x_t)]_{Q(B_M)} + \mathbb{E}[\ln \text{SBP}(z_t)]_{Q(B_N)} \right\}, \quad (53)$$

のように分解できる．したがって，時刻  $\tau - T$  の状態の事後分布は，事前分布  $P(s_{\tau-T-1})$  を用いて  $Q(s_{\tau-T}, M_{\tau-T}, N_{\tau-T}) = \sum_{s_{\tau-T-1}} Q(s_{\tau-T}, M_{\tau-T}, N_{\tau-T} | s_{\tau-T-1}) P(s_{\tau-T-1})$  で計算できる．次に，同時確率  $Q(s_{t-1}, s_t, M_t, N_t) = Q(s_t, M_t, N_t | s_{t-1}) Q(s_{t-1})$  と周辺化確率  $Q(s_t) = \sum_{s_{t-1}, M_t, N_t} Q(s_{t-1}, s_t, M_t, N_t)$  は各時刻  $t$  について， $\tau - T + 1$  から  $\tau$  まです順に得られる．ディリクレ分布に従う変数の対数の期待値は，次のようにディガンマ関数  $\psi(\cdot)$  を用いて計算できる：

$$\mathbb{E}[\ln L_{os}]_{Q(L)} = \psi(\Theta_{os}^L) - \psi\left(\sum_o \Theta_{os}^L\right) \quad (54)$$

$$\mathbb{E}[\ln M_{s' sa}^x]_{Q(M^x)} = \psi(\Theta_{s' sa}^{M^x}) - \psi\left(\sum_{s'} \Theta_{s' sa}^{M^x}\right), \quad (55)$$

$$\mathbb{E}[\ln N_{rs}^z]_{Q(N^z)} = \psi(\Theta_{rs}^{N^z}) - \psi\left(\sum_r \Theta_{rs}^{N^z}\right), \quad (56)$$

$$\mathbb{E}[\ln \text{SBP}(x)]_{Q(B_M)} = [\psi(\Theta_x^V) - \psi(\Theta_x^V + \bar{\Theta}_x^V)] \\ \times \sum_{k=1}^{x-1} [\psi(\bar{\Theta}_k^V) - \psi(\Theta_k^V + \bar{\Theta}_k^V)], \quad (57)$$

$$\mathbb{E}[\ln \text{SBP}(z)]_{Q(B_N)} = [\psi(\Theta_z^V) - \psi(\Theta_z^V + \bar{\Theta}_z^V)] \\ \times \sum_{k=1}^{z-1} [\psi(\bar{\Theta}_k^V) - \psi(\Theta_k^V + \bar{\Theta}_k^V)]. \quad (58)$$

E ステップ（式 (53)）と M ステップ（式 (47) - 式 (51)）を交互に適用することで，自由エネルギーからの更新規則の導出に対する近似事後分布が得られる．

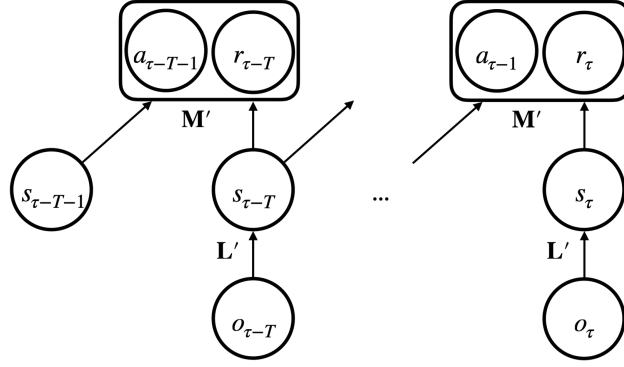


図 9: ROMDP における GOEI のグラフィカルモデル. このモデルは, 状態  $s_{t-1}$  および  $s_t$ , 行動  $a_t$ , 報酬  $r_t$  の関係に基づいて観測をクラスタリングする. モデルの構造は, 行動が状態遷移から生成されることを示しており, これは実際の因果関係とは一致しない. この点が, 目的指向のコア状態を効果的に識別するために重要である.

#### 4.5 目的指向な環境推論

ROMDP アプローチでは, 式 (3) の状態削減モデル  $p(s_t | \mathcal{A}_{t-1}, \mathcal{R}_t, \mathcal{O}_t)$  を仮定 (28) に従って  $P(s_t | o_t)$  に置き換える. 式 (3) のベイズ推論には, 式 (3) に現れる確率分布と経験  $\mathcal{A}_{t-1}, \mathcal{R}_t$ , および時刻  $t$  での  $\mathcal{O}_t$  を含む生成モデルが必要である. しかし, 仮定 (28) に従った式 (3) のグラフ構造は  $o_t$  から始まるため, 異なる時刻  $t, t-1, \dots$  から始まる式 (3) の構造は互いに矛盾する. 一方, 時刻  $t$  で始まるグラフ構造では, 状態  $s_t$  は  $P(s_t | o_t)$  により  $o_t$  から生成される. 他方, 同じ  $s_t$  は, 一時刻前の  $o_{t-1}$  で始まる構造においては,  $p(s_t | a_{t-1}, s_{t-1})$  により  $a_{t-1}$  と  $s_{t-1}$  から生成されるべきである. したがって, 時刻  $t$  と  $t-1$  から始まるグラフ構造は統合できない. 他の時刻でも同様の矛盾が生じるため, 観測系列  $o_t, o_{t-1}, \dots$  を含む同時生成モデルを記述することはできない. 異なる開始点に対して異なる生成モデルを構築することはできるが, ベイズ推論が非常に複雑になる. 以下では, 異なる手順を提案する.

ベイズの定理は, 条件付き確率分布の変数がそれらの真の因果関係に関係なく可逆であることを示している. これは, ベイズ推論に使用される条件付き確率分布の方向が推論対象の環境の方向と一致する必要がないことを意味する. この柔軟性を活用して, 行動, 報酬, および隣接する時刻の状態間の相関を維持しつつ, 式 (3) とは異なるグラフィカルモデルを考



える (図 9). よって, エージェントモデルは

$$P(\mathcal{R}, \mathcal{A}, \mathcal{S} | \mathcal{O}, s_{\tau-T-1}, L', \tilde{M}') = \prod_{t=\tau-T}^{\tau} P(r_t, a_{t-1} | s_t, s_{t-1}, M'_t) P(s_t | o_t, L'), \quad (59)$$

で記述される.

状態削減モデル  $P(s_t | o_t)$  のパラメータを  $L'$  とし, 行動-報酬モデル  $P(r_t, a_{t-1} | s_t, s_{t-1})$  のパラメータを  $M't \in \bar{M}' = [M'^1, \dots]$  とする.  $M't$  はハイパーパラメータ  $\alpha_y (> 0)$  と基底分布  $BM'$  に従うディリクレ過程に従う. このエージェントモデルは, 現在の行動とその結果としての報酬を予測できるように観測値を状態に分類するために, 現在と前回の状態  $[s_{t-1}, s_t]$  を組み合わせる.

$P(a_{t-1} | s_t, s_{t-1})$  のモデルは一見すると不自然に見えるかもしれないが, これは状態  $s_{t-1}$  から  $s_t$  への遷移が行動  $a_{t-1}$  によって引き起こされた可能性を表しているだけであり, 未来の状態から行動を予測することを意味しているわけではない. GOEI (Goal-Oriented Environment Inference) は環境を学習するだけであり, 行動制御には直接関与しない. 重要な点は, GOEI が事後分布を更新することで行動と報酬の尤度を最大化することである. この最大化により, 観測が予測不可能な場合でも尤度が高い状態を保つことができる. これは, 観測の予測不可能性が尤度を低下させる CEI (Complete Environment Inference) とは対照的である.

行動, 観測, および報酬が前述の POMDP の定式化のように離散的である場合,  $L'$  および  $M't$  はカテゴリカル分布に従う. これらの事前分布は, それぞれ  $\alpha L', \Phi^{L'}, \Phi^{M'y}, \alpha_y$ , および  $\Phi^Y$  というハイパーパラメータを持つ独立したディリクレ分布に従う:

$$P(L', B_{M'}) = P(L')P(\bar{M}', Y) = \text{SBP}(L'; \alpha_{L'}, \Phi_{L'}) \prod_y^{\infty} \text{Dir}(M'^y; \Phi^{M'}) \text{SBP}(Y; \alpha_y, \Phi^Y). \quad (60)$$

ここで,  $Y$  は  $M'$  に対する割り当て指標  $y \sim \text{Cat}(Y)$  の確率であり, 棒折過程に従う. この過程では, 基底分布  $B_{M'}$  が  $\bar{M}'$  と割り当て指標  $Y$  に分解される. 変分ベイズ法における近似事後分布は, ハイパーパラメータ  $\Theta^{L'}, \Theta^{M'y}$ , および  $\Theta^Y$  を持つディリクレ分布の積と

なる：

$$\begin{aligned}
Q(\Omega_{\text{GOEI}}) &\equiv Q(L', B_{M'}) = Q(L', \bar{M}', Y) = Q(L'; \Theta^{L'}) \prod_y^{\infty} Q(M'^y; \Theta^{M'^y}) Q(Y; \alpha_y, \Theta^Y) \\
&= \text{SBP}(L'; \alpha_{L'}, \Theta_{L'}) \prod_y^{\infty} \text{Dir}(M'^y; \Theta^{M'}) \text{SBP}(Y; \alpha_y, \Theta^Y). \quad (61)
\end{aligned}$$

GOEI は報酬  $r$  と行動  $a$  を予測し、観測  $o$  は予測しない。要約すると、GOEI の自由エネルギーは、環境推論モデルにおいて式 (59) と式 (60) を式 (23) に代入し、 $D = \{\tilde{a}, \tilde{r}\}$ ,  $\bar{D} = \{\tilde{o}\}$ , そして  $\Omega = \{L', B_{M'}\}$  と設定することで得られる。

GOEI の M-step はエージェントモデル (59) を FE の式 (34) に代入することで、

$$Q(L', B_{M'}) = \exp \left[ \sum_{t=\tau-T}^{\tau} Q(s_t, M'_t) \left\{ \ln L'^{o_t}_{s_t} + \ln M'^{y_t}_{a_{t-1} r_t s_t s_{t-1}} + \ln \text{SBP}(y_t) \right\} \right] P(\hat{L}', B_{M'}) \quad (62)$$

となる。よって、ハイパーパラメータの更新則は、

$$\Theta_{so}^{L'} \leftarrow \Phi_{so}^{L'} + \sum_{t=\tau-T}^{\tau} Q(s_t=s) \mathbb{1}(o_t, o), \quad (63)$$

$$\Theta_{ars's}^{M'^y} \leftarrow \Phi_{ars's}^{M'^y} + \sum_{t=\tau-T}^{\tau} Q(s_t=s', s_{t-1}=s, M'_t=M'^y) \mathbb{1}(a_{t-1}, a) \mathbb{1}(r_t, r), \quad (64)$$

$$\Theta_y^Y \leftarrow \Phi_y^Y + \sum_{t=\tau-T}^{\tau} Q(M'_t=M'^y), \quad (65)$$

で得られる。また、CEI と同様に、GOEI の E-step はエージェントモデル (59) を式 (33)

に代入することで、 $Q(\tilde{s}, \tilde{M}') = \prod_{t=\tau-T}^{\tau} Q(s_t, M'_t | s_{t-1})$  と分解でき、

$$Q(s_t, M'_t | s_{t-1}) \propto \exp \left[ \mathbb{E} \left[ \ln L'^{o_t}_{s_t} \right]_{Q(L')} + \mathbb{E} \left[ \ln M'^{y_t}_{a_{t-1} r_t s_t s_{t-1}} \right]_{Q(M')} + \mathbb{E} \left[ \ln \text{SBP}(y_t) \right]_{Q(B_{M'})} \right], \quad (66)$$

で計算できる．E-step の計算に必要な確率分布の対数の期待値は，それぞれ

$$\begin{aligned} \mathbb{E}[\ln L'^o_s]_{Q(L')} &= [\psi(V_{so}^{L'}) - \psi(V_{so}^{L'} + \bar{V}_{so}^{L'})] \\ &\quad \times \sum_{k=1}^{s-1} [\psi(\bar{V}_{ko}^{L'}) - \psi(V_{ko}^{L'} + \bar{V}_{ko}^{L'})], \end{aligned} \quad (67)$$

$$\mathbb{E}[\ln M'^y_{ars's}]_{Q(M')} = \psi(\Theta_{ars's}^{M'^y}) - \psi\left(\sum_{a,r} \Theta_{ars's}^{M'^y}\right), \quad (68)$$

$$\begin{aligned} \mathbb{E}[\ln \text{SBP}(y)]_{Q(B_{M'})} &= [\psi(\Theta_y^V) - \psi(\Theta_y^V + \bar{\Theta}_y^V)] \\ &\quad \times \sum_{k=1}^{y-1} [\psi(\bar{\Theta}_k^V) - \psi(\Theta_k^V + \bar{\Theta}_k^V)], \end{aligned} \quad (69)$$

で与えられる．

もし 2 つの解が同等の予測性能を持つ場合，GOEI はより少ない状態を持つ解を好む．GOEI では，状態変数  $s$  はハイパーパラメータ  $\alpha_{L'} > 0$  を持つディリクレ過程に従う（すなわち， $s \sim \text{DP}(L', \alpha_{L'})$ ），これによりシミュレーション中に任意の数の状態が許容される．変分ベイズ法は，エビデンス下限（ELBO）を最大化する．この ELBO は，一般化された対数尤度から事前分布と事後分布の間のカルバック・ライブラー（KL）ダイバージェンスを引いたものである [61]．ディリクレ過程において，KL ダイバージェンスは一般に状態が少ないほど小さくなるため，対数尤度（すなわち，性能レベル）が同じであれば，状態が少ない方が好まれる．これにより，GOEI は予測性能を保ちながら，モデルの複雑さを最小限に抑えることができる．

各推定された状態において最適な行動を選択するためには，最適ベルマン方程式 (2) に従い，環境の遷移モデル  $P(s_{t+1} | s_t, a_t)$  と報酬モデル  $P(r_t | s_t)$  が必要である．しかし，GOEI は遷移モデル  $P(s_{t+1} | a_t, s_t)$  の代わりに不一致なエージェントモデル  $P(a_t | s_{t+1}, s_t)$  を使用するため，これらのモデルを直接推定することはない．したがって，遷移モデルと報酬モデルは GOEI のグラフィカルモデルとは独立して推論しなければならない．この困難を解決するために，遷移モデルおよび報酬モデルの割り当て指標の事後分布をそれぞれ，

$$Q(M_t | s_t, s_{t-1}) \propto \exp \{ \mathbb{E}_{Q(\bar{M}, X)} [\ln P(s_t, M_t | s_{t-1}, a_{t-1}, \bar{M}, X)] \}, \quad (70)$$

$$Q(N_t | s_t) \propto \exp \{ \mathbb{E}_{Q(\bar{N}, Z)} [\ln P(r_t, N_t | s_t, \bar{N}, Z)] \}, \quad (71)$$

で計算し、式 (63) ～ (66) の反復によって推定された状態分布  $Q(S)$  を式 (48) ～ (51) に代入することで、それらの規則を更新する。

#### 4.6 忘却によるオンライン処理

環境をオンラインで推定するために、GOEI の更新則を毎  $T$  ステップごとに適用する [62]。更新されたパラメータの近似事後分布  $Q(L', B_{M'})$  を  $T$  ステップ後の事前分布  $P(L', B_{M'})$  として使用する。更新範囲  $T$  を小さくするほど、試行錯誤による状態の推定が早くなる。しかし、更新範囲  $T$  を小さくするのは簡単ではない。主な理由は次の 2 つである。

まず、事後分布のオンライン更新は切り捨て誤差を導入する。過去の逐次更新を切り捨てた後、現在の推定を改善するために切り捨てられた過去の状態を再推定することはできない。事後分布  $Q(L', B_{M'})$  の更新は推定された状態  $Q(s)$  に依存するため、 $T$  の値が小さいほど、過去の状態推定における不確実性が事後分布に速く蓄積する。結果として、後のステップでの状態推定における切り捨て誤差の修正がますます困難になる。

第二に、更新の各ステップで近似事後分布に近似誤差が導入される。変分ベイズ法による近似事後分布の更新は、パラメータと状態の独立性を仮定する。しかし、この仮定は一般的には保証されていない [60]。その結果、近似事後分布は近似誤差を蓄積していき、時間が経つにつれて推定結果の精度が劣化する。この蓄積効果は、特に長期間にわたる推定における行動制御に重大な影響を及ぼす可能性がある。

更新範囲を縮小するために、更新則 (63) に忘却効果を導入する。ここで言う忘却は、事後分布を更新する際に、範囲  $T$  内の観測の影響を強化するものである。言い換えれば、忘却は過去の記憶である事前分布の影響を割り引くことにより、パラメータ推定における累積した近似誤差を除去する。パラメータ  $\eta_t$  は、予測された観測  $o_t$  および推定された状態  $s_t$  の正確さに応じて、忘却の度合いを動的に調整する：

$$\eta_t = (1 - \rho) (1 - q(s_t = s)) \mathbb{1}(o_t, o), \quad (72)$$

$$\Theta_{so}^{L'} \leftarrow (1 - \eta_t) \Theta_{so}^{L'} + Q(s_t) \mathbb{1}(o_t, o). \quad (73)$$

ここで、 $\eta_t \in [0, 1 - \rho]$  であり、 $\rho \in [0, 1]$  は最大忘却率である。パラメータ  $\Theta_{so}^{L'}$  の初期値は  $\Phi_{so}^{L'}$  であり、 $\tau - T$  から  $\tau$  まで順次更新される。時刻  $t$  での観測が実際に  $o$  であるとき、 $1(o_t, o) = 1$  かつ  $Q(s_t) = 0$  の場合、最大の忘却が  $1 - \eta_t = \rho$  で発生する。

#### 4.7 行動価値トンプソンサンプリングによる行動制御

最適ベルマン方程式 (2) は価値反復法 [19] によって解くことができる。しかし、遷移モデル、報酬モデル、それらの割り当て指標、および真の状態はエージェントには未知である。そのため、エージェントは探索行動を通じてこれらを推定する必要があるが、過度の探索は報酬を獲得する機会を失う結果になる。最適な行動を学ぶためには、探索と活用間の適切なバランスが重要である。そこで、モデル  $M^x$  と  $N^z$  からサンプリングすることにより、最適ベルマン方程式 (2) を解いて行動  $a_\tau = \arg \max_a \mathbb{E}_{Q(s_\tau)}[Q(s_\tau, a)]$  を得ることを提案する。ここで、CEI と GOEI の両方が、遷移モデル、報酬モデル、それらの割り当て指標、および真の状態の確率分布  $Q(B_M)$ ,  $Q(B_N)$ , および  $\vec{Q}(s_\tau)$  を推定するために変分ベイズ推論を使用することに注意する。これにより、推定された分布からモデル  $M^x$  と  $N^z$  をサンプリングすることが可能になる。

提案した行動制御の方法は、環境推定に対するエージェントの信頼度に応じて探索と活用を動的に制御する。推定の精度が低い場合、広い分布からサンプリングされた行動価値は各タイムステップで大きく変動し、探索的な行動をもたらす。これに対して、推定の精度が高い場合、狭い分布からサンプリングされた行動価値はあまり変動せず、活用的な行動となる。提案されたサンプリングベースの行動制御は、マルチアームバンディット問題における推定された報酬分布からのサンプリングであるトンプソンサンプリング [63] を拡張したものである。したがって、提案した方法を行動価値トンプソンサンプリング (Action-value Thompson sampling: ATS)」と呼ぶ。

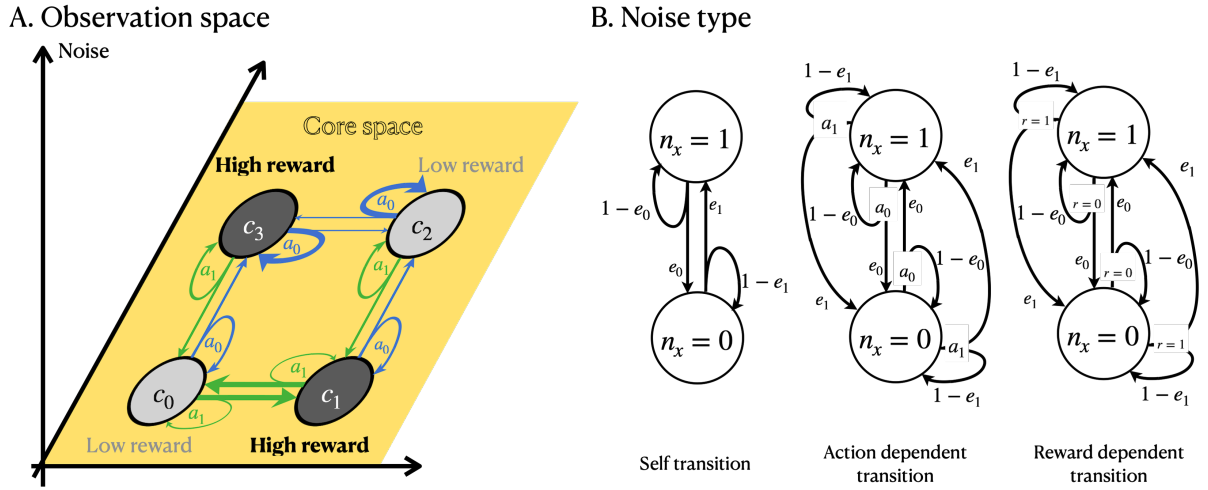


図 10: 数値実験のテスト環境. (A) 観測空間はコア空間とそれに直交するノイズ空間の直積として与えられる. コア状態は高報酬状態（黒）と低報酬状態（灰色）からなり、エージェントは2つの行動 ( $a_0, a_1$ ) を選択できる. 矢印の太さは対応する状態遷移の確率に比例する. (B) 自己遷移型  $p(n_{t+1} | n_t)$ , 行動依存型  $p(n_{t+1} | n_t, a_t)$ , 報酬依存型  $p(n_{t+1} | n_t, r_t)$  の3種類のノイズを考えた.  $e_0$  と  $e_1$  はノイズの遷移則の確率のパラメータである.

## 5 数値実験

### 5.1 コアとノイズによるテスト環境

表 2: 行動  $a_0$  によるコアの遷移確率

from \ to	$c_0$	$c_1$	$c_2$	$c_3$
$c_0$	0.5	0	0	0.5
$c_1$	0	0.5	0.5	0
$c_2$	0	0	$1 - \varepsilon$	$\varepsilon$
$c_3$	0	0	$\varepsilon$	$1 - \varepsilon$

表 3: 行動  $a_1$  によるコアの遷移確率

from \ to	$c_0$	$c_1$	$c_2$	$c_3$
$c_0$	$\varepsilon$	$1 - \varepsilon$	0	0
$c_1$	$1 - \varepsilon$	$\varepsilon$	0	0
$c_2$	0	0.5	0.5	0
$c_3$	0.5	0	0	0.5

ROMDP アプローチのために提案した GOEI をテストするために, ROMDP の仮定 (28) が成り立つテスト環境を構築した. 排除すべき情報を分離するために, 観測  $o_t$  をコア  $c_t \in C$  とノイズ  $n_t \in N$  の直積として表現した. 報酬獲得に無関係な情報は全てノイズと

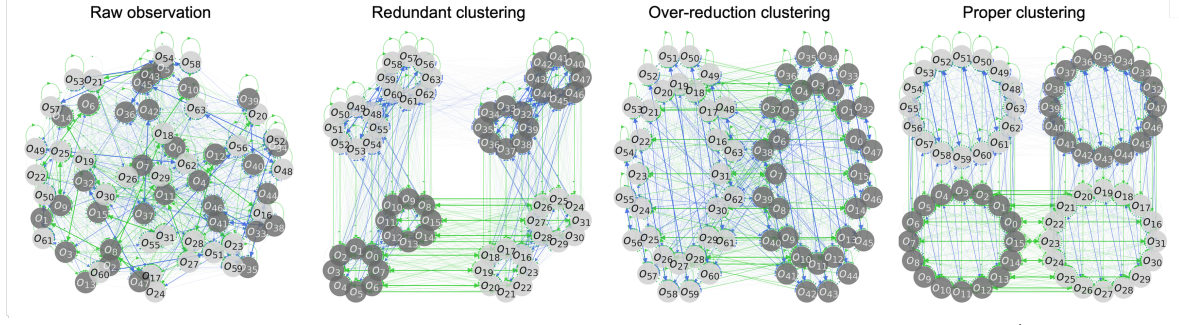


図 11: 観測のクラスタリングの模式図. このダイアグラムの例は, 遷移則  $M^1$  と報酬則  $N^1$  に従う 4 つのコア状態と 4-bit の行動依存型のノイズで構成された観測に対する例を示している. 報酬を手がかりとして, エージェントは複雑な遷移ダイアグラムに従う生の観測 (一番左の例) をコアの構造を保持するようにクラスタリングしなければならない. 遷移ダイアグラムには, 高報酬 (濃い灰色の円) と低報酬 (薄い灰色の円) の状態があり, エージェントは 2 つの行動 (緑と青) を選択できる. 一番右のダイアグラムは 4 つのコアに正しくクラスタリングした例であり, 一方, 真ん中の左と右のダイアグラムは, 冗長な 8 つの状態と過剰な 2 つの状態へクラスタリングした例である. クラスタリングされた観測値は円状に配置している.

見なされるため, 与えられた行動に対する報酬の生成モデルは, その遷移  $p(c_{t+1} | c_t, a_t)$  および  $p(r_t | c_t)$  を通じて, コアに依存する. ノイズの生成規則は多数考えられるが, テストではノイズがランダムではなく, 選択された行動と結果として得られた報酬に依存するマルコフ遷移則  $p(n_{t+1} | n_t, a_t, r_t)$  に従う場合を考慮した.

テスト環境では, コアが報酬の大きさや確率から容易に推測できないようにする必要がある. こうした環境には, 少なくとも 4 つのコア状態  $c_0, c_1, c_2$ , および  $c_3$  が必要で, 報酬確率が同一のコア状態のペアを 2 組合むようにする. 各コア状態での報酬は二値 ( $r = 1$  もしくは  $r = 0$ ) であり, 状態に依存する報酬則  $N^1$  を設定する. 高報酬コア状態  $c_1$  と  $c_3$  では  $p_{N^1}(r = 1 | c_1) = p_{N^1}(r = 1 | c_3) = 0.9$  とし, 低報酬コア状態  $c_0$  と  $c_2$  では  $p_{N^1}(r = 1 | c_0) = p_{N^1}(r = 1 | c_2) = 0.1$  とする. さらに, 行動に依存する遷移則  $M^1$  を使用し, 高報酬のコア状態に留まることが常に最適ではないように設定する (図 10A). エージェントが最も効率的に報酬を得るためには, 高報酬の状態  $c_3$  に留まる行動  $a_0$  を選択する必要があるが, もう一方の高報酬の状態  $c_1$  からは離れる行動  $a_1$  を選択する必要がある.

行動  $a_0$  を選択した際, 状態  $c_3$  から  $c_2$  への遷移は稀にしか起こらない. しかし, この稀

なケースでは、 $c_2$  から  $c_3$  への復帰は、 $c_1$  および  $c_0$  を経由するルートを取る方が直接戻るよりも速い。この環境での最適な戦略は  $(c_0, a_0)$ ,  $(c_1, a_1)$ ,  $(c_2, a_1)$ , および  $(c_3, a_0)$  の通りである。同じ報酬確率を持つ状態ペア、すなわち  $c_0 - c_2$  および  $c_1 - c_3$  が異なる最適な行動を必要とするため、エージェントは報酬確率のみに基づいてコア状態を識別しても最適な戦略を学習できない。コア状態間の遷移確率は表 2 および表 3 に示されている。

環境には  $m$  ビットのノイズ  $(n_1, n_2, \dots, n_m)$  を導入し、各ノイズビットは 2 つの状態  $n_x \in 0, 1$  の間を遷移する。さらに、3 つの遷移則のタイプを考慮する (図 10B)。自己遷移タイプは遷移則  $p(n_{t+1} | n_t)$  に従って遷移する。行動依存型は、選択された行動に依存する規則  $p(n_{t+1} | n_t, a_t)$  で遷移を生成する。報酬依存型は、獲得した報酬に依存する規則  $p(n_{t+1} | n_t, r_t)$  で遷移を生成する。 $e_0 = e_1 = 0.5$  の場合、すべてのノイズタイプはランダムノイズとなる。この場合、観測の遷移則はランダムノイズとコアの直積として与えられ、ノイズの方向において構造を持たない。したがって、観測の不確実性のみを考慮すればコアの推定が可能な簡単なタスクとなる。そこでコアの推定をより困難にするために、数値実験では、遷移確率を以下のように設定した： $e_0 = e_1 = 0.1$  (自己遷移型),  $e_0 = 0.1$  および  $e_1 = 0.9$  (行動依存型),  $e_0 = 0.1, e_1 = 0.9$  (報酬依存型)。

各観測は、コアビットとノイズビットの直積として与えられる。したがって、可能な観測の数は  $4 \times 2^m$  となる。観測からコア状態の推定を難しくするために、観測  $o$  を  $(4 \times 2^m)$  次元の one-hot ベクトルで表現し、観測の構造からコア状態を復元できないようにした。数値実験では、 $m = 4$  で、エージェントのタスクは 64 の観測を 4 つのコアにクラスタリングすることとなる。図 11 は、生の観測をクラスタリングした場合の異なる結果を例示している。適切なクラスタリングは、4 つのコア状態とそれらの間の遷移則を提供する。しかし、エージェントがコア推定を失敗した場合には、4 つ以上の (冗長なクラスタリング) または 4 つ未満の (過剰削減な) クラスタを生成する可能性がある。

## 5.2 非定常なテスト環境

前節で定義した基本環境に、2 種類の非定常性を導入する。非定常な報酬則環境では、遷移則を  $M^1$  に固定し、報酬則を 5000 ステップごとに  $N^1$  と  $N^2$  (図 12 の中央と左端のパネ



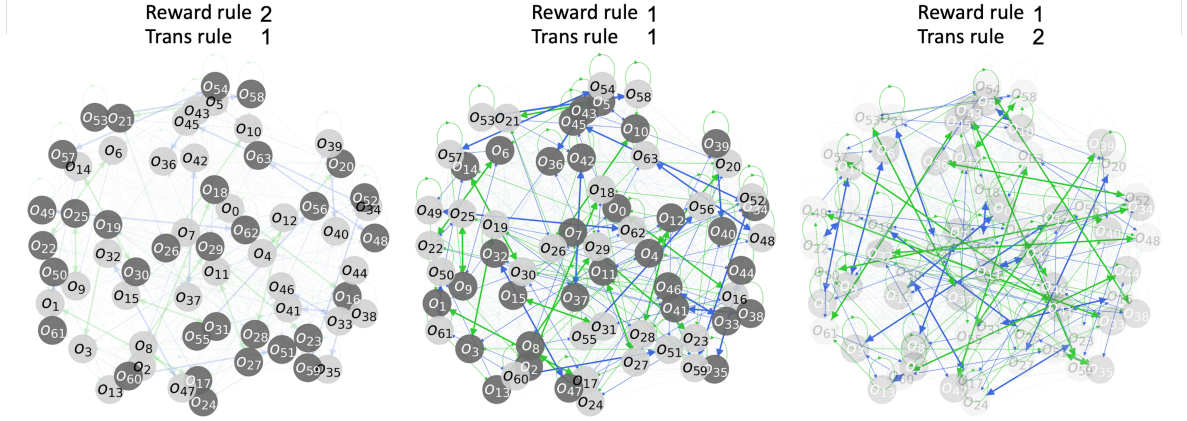


図 12: 非定常なテスト環境の図. 4 ビットの行動依存ノイズを含む観測は、高報酬（濃い灰色）または低報酬（薄い灰色）に関連するノードで示される. 中央の図は、初期条件の報酬則  $N^1$  と遷移則  $M^1$  に従う観測間の遷移ダイアグラムを示している. 一番左の図は、コアの報酬則のみが  $N^2$  に変化する場合の観測のダイアグラムを表し、中央のパネルと同様の遷移則を意味する矢印が薄く描かれている. 一番右の図は、コアの遷移則のみが  $M^2$  に変化したときの遷移図であり、中央の図と同様の報酬則を意味するノードは薄く描かれている.

ル) の間を交互に変更する. 変更後の報酬則は  $p_{N^2}(r = 1 | c) = p_{N^1}(r = 0 | c)$  であり、報酬は  $c_2$  で  $a_0$  を選択することで最も効率的に得られる.

これに対して、非定常な遷移則環境では、報酬則  $N^1$  を変更せずに、5000 ステップごとに遷移則を  $M^1$  と  $M^2$  の間で交互に切り替える (図 12 の中央と右端のパネル). 変更後の遷移則は、 $p_{M^2}(c_k | c_i, a_j) = 1 - p_{M^1}(c_k | c_i, a_{1-j})$  であり、 $p_{M^1}(c_k | c_i, a_{1-j}) \neq 0$  のとき、報酬を獲得するための最適戦略は  $c_1$  で  $a_0$  を選択することである. 両方の非定常環境で共通の最適な戦略は  $(c_0, a_1)$ ,  $(c_1, a_0)$ ,  $(c_2, a_0)$ , および  $(c_3, a_1)$  として与えられ、変更前と最適な戦略が一致しないようになっている.

### 5.3 状態削減の評価指標

ROMDP の目的は最適戦略が可能な最小の状態集合を推定することである. 言い換えれば、コアの情報を保持しながら観測の情報を削減することに等しい. この情報の削減度を評価するために、コアに対する情報損失度と観測に対する情報削減度を条件付きエントロピー

$H(\cdot|s) \geq 0$  を用いて、それぞれ、

Information loss:

$$H(c|s) = - \sum_{c \in C} \sum_{s \in S} p(c, s) \ln p(c|s), \quad (74)$$

Information reduction:

$$H(o|s) = - \sum_{o \in O} \sum_{s \in S} p(o, s) \ln p(o|s), \quad (75)$$

と定義する．ここで、 $s$  はエージェントが推定している状態であり、 $c$  はエージェントが知ることができない環境の真のコア状態である．

推定された状態  $s$  がすべてのコア情報を持っているとき、情報損失  $H(c|s) = 0$  となり、核心情報が失われるにつれて増加する．推定された状態がすべての観測を表す場合、情報削減  $H(o|s) = 0$  となり、一部の冗長な観測に関する情報が削減されるにつれて増加する．コア状態は、 $H(o|s)$  をさらに最大化しつつ  $H(c|s) = 0$  を維持する場合、より少ない状態で表現することができる． $H(o|s)$  の最大値は観測の分布に依存する．図 10 に示すテスト環境では、ノイズ空間の次元は  $m = 4$  であり、理論的な上限は  $H(o|s) \doteq 4.15$  である．

## 5.4 コア推定による戦略学習の効果検証

表 4: CEI の初期事前分布

$\Phi_{os}^L$	$\Phi_{s'sa}^{M^x}$	$\Phi_{rs}^{N^z}$
$0.1 + 0.9 \times \mathbf{1}(o=s)$	0.5	0.5

表 5: GOEI の初期事前分布

$\alpha_{L'}$	$\Phi_{so}^{L'}$	$\alpha_y$	$\Phi_y^Y$	$\Phi_{ars's}^{M'^y}$
4	0	0.5	0	0.5

状態推定を GOEI と組み合わせ、行動制御を ATS で扱うことで、状態の削減度と戦略学習の速度との関係を検討した．比較対象には CEI と ATS の組み合わせによる結果を用いた．更新頻度と最大忘却率は、長期報酬を最大化できる状態を最も効率的に推定できるように、 $T = 500$ 、 $\rho = 0.95$ 、 $\gamma = 0.95$  と設定した．国立研究開発法人産業技術総合研究所 (AIST) の NVIDIA A100 GPU を使用して、GOEI と CEI は平均して 30 分で 10,000 ステップを計算した．

エージェントは以下のシミュレーションにおいて観測の構造に関する事前知識を持っ

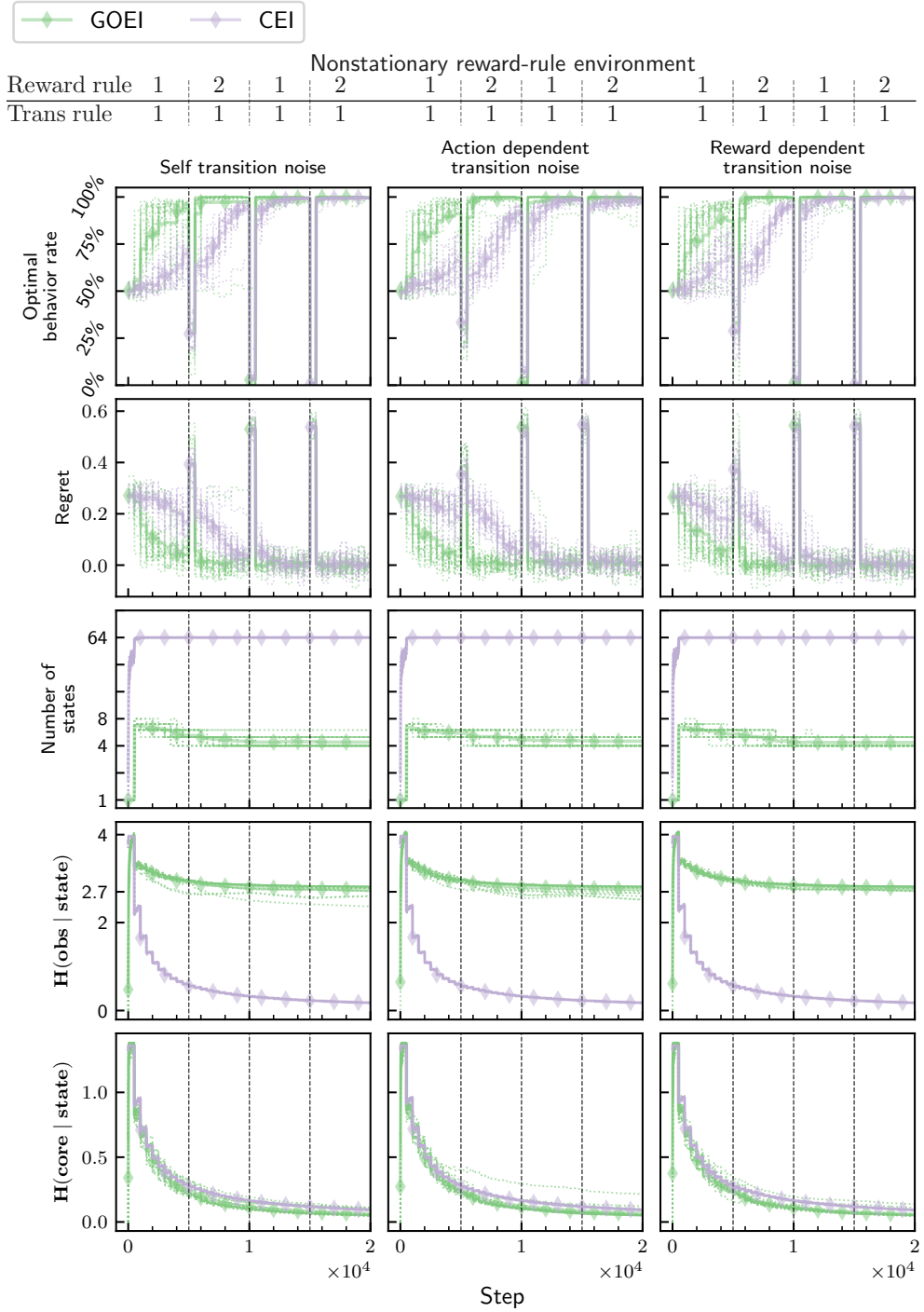


図 13: 報酬則が非定常な環境における GOEI と CEI の比較. 上部の表に示すように, 2 つの報酬則を 5000 ステップごとに交互に適用した. コア遷移の不確実性  $\varepsilon$  は 0.05 に設定した. シミュレーションは 3 種類のノイズに対して行った. 全ての可能な観測とコアの数はそれぞれ 64 と 4 であり, 最も冗長な状態数は 64 である. 最適行動率とリグレットは更新間隔  $T = 500$  毎の平均である. 各図において, 破線は個々のサンプルの結果を示し, 菱形の実線は 20 サンプルの平均を示す.

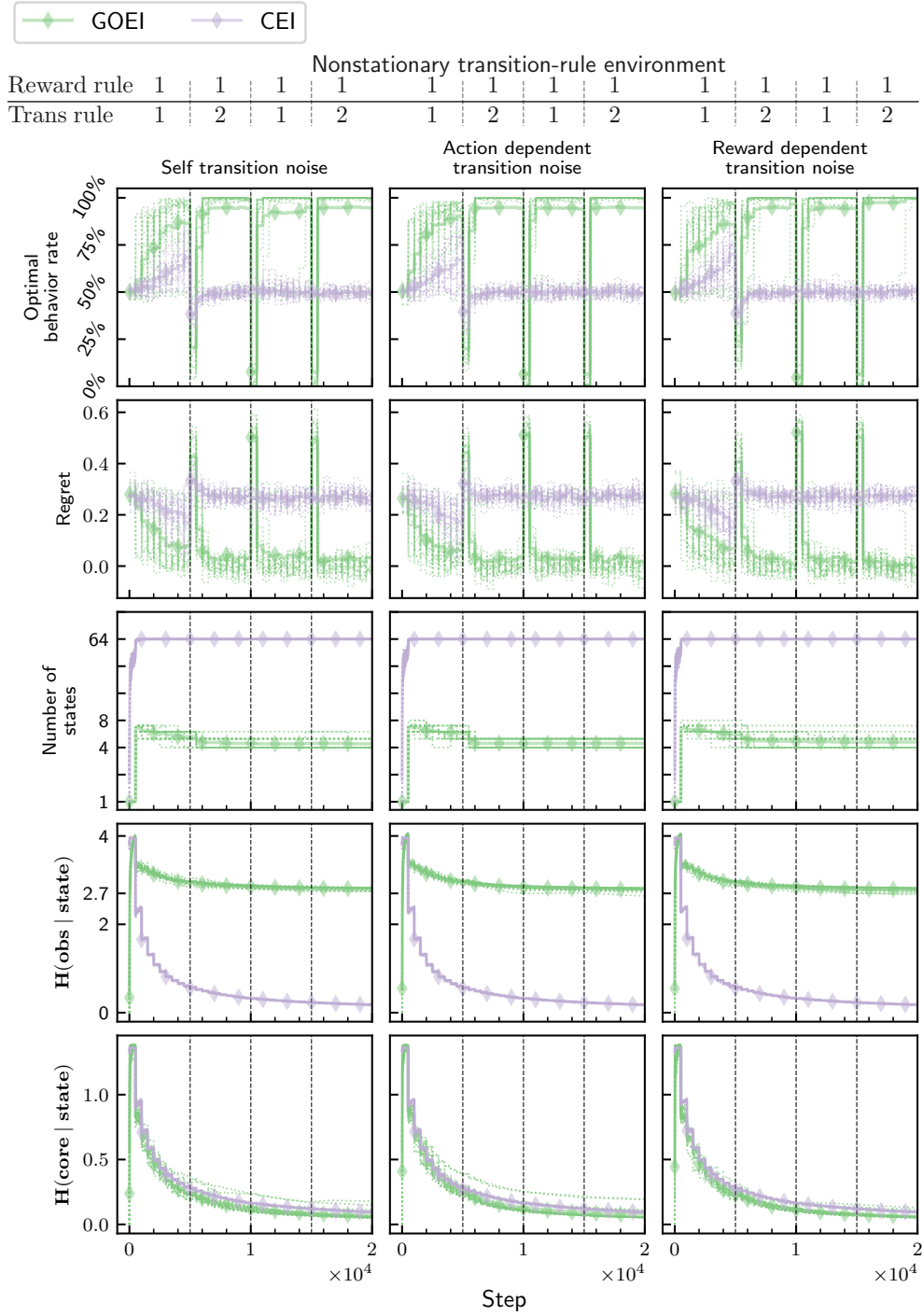


図 14: 遷移則が非定常な環境における GOEI と CEI の比較. 上部の表に示すように, 2 つの状態間の遷移則を 5000 ステップごとに交互に適用した. コア遷移の不確か性  $\varepsilon$  は 0.05 に設定した. 他のシミュレーション設定は図と同じである.

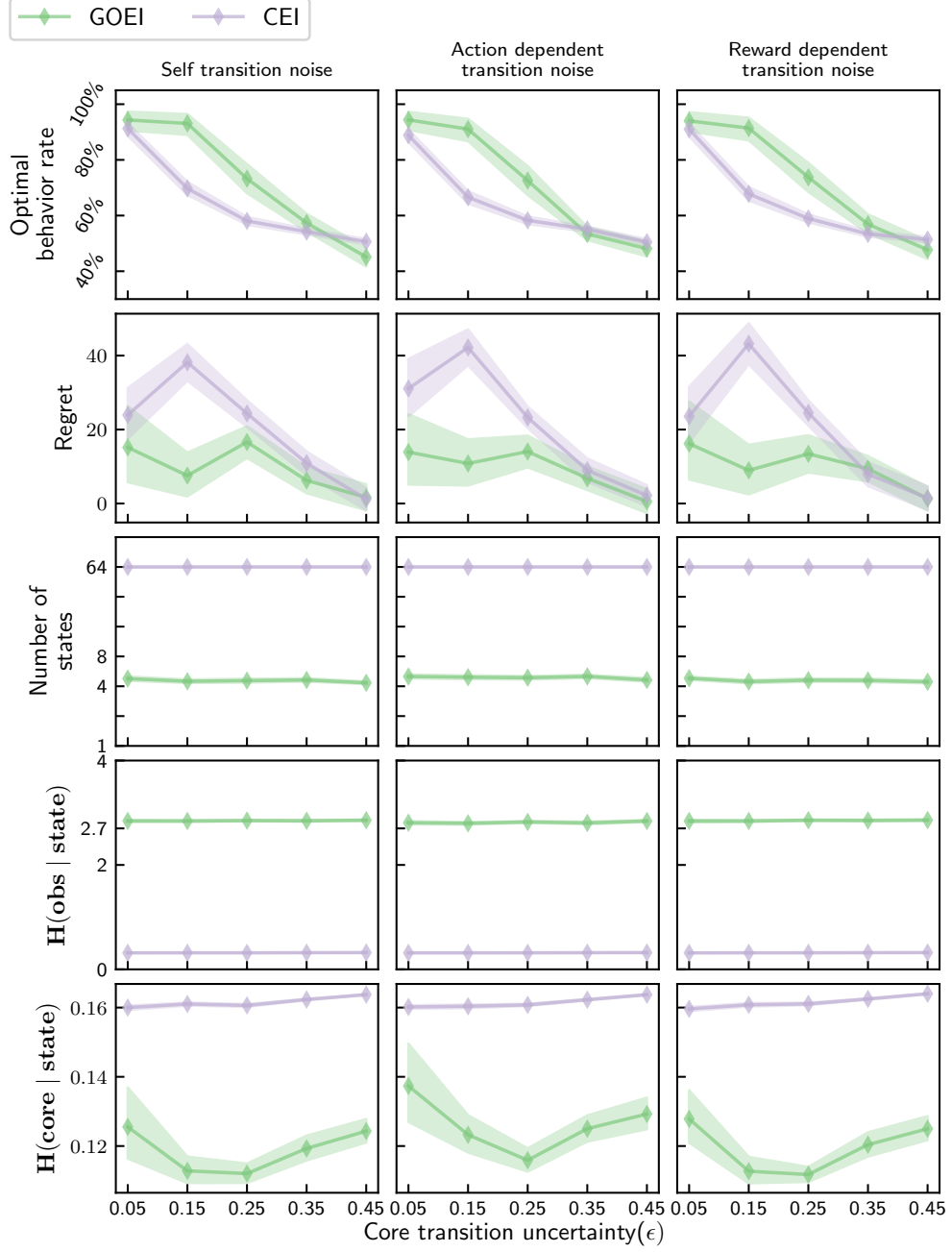


図 15: 報酬則  $N^1$  と遷移則  $M^1$  のテスト環境において、コア遷移の不確実性を様々な値に設定した場合の 10000 学習ステップ後の GOEI と CEI の比較. 最適行動率は、最後の  $T = 500$  ステップの間にエージェントがとった全ての行動に対する最適行動の数の割合を意味する. リグレットは、最適戦略による報酬の合計から、最後の  $T = 500$  ステップにおいてエージェントが実際に獲得した報酬の合計を引いたものである. 菱形の実線は 100 サンプルの平均, エラーバーはその 95% 信頼区間である.

ていなかったことに注意する．よって、全てのノイズタイプ、コアの不確実性、および非定常性を含む実験環境において共通の事前分布を設定した．具体的には、総報酬を最大化するように初期パラメータを設定した．CEI と GOEI の両方に共通するパラメータは  $\Phi^X x = \Phi^Z z = 0$ ,  $\alpha_x = \alpha_z = 1$  である．他のパラメータの詳細は表 4 および表 5 に示されている．パラメータの小さな変動が結果に大きな影響を与えないことを確認した．

報酬則と遷移則の変化は図 13 および図 14 の上部の表に示している．非定常な報酬則（図 13 の三段目）および遷移則（図 14 の三段目）の環境において、GOEI は全てのノイズタイプに対して、状態数は真のコア数 4 に近く、常にコア空間に十分近い状態空間を推定することができた．推定された状態空間の情報削減度は、 $H(o|s) \approx 2.7$ （図 13 と 14 の四段目）および情報損失度は  $H(c|s) = 0$ （五段目）であった．これらの結果は、コア状態に関する情報を失うことなく状態空間が適切に縮小されたことを示している．一方、CEI は全てのノイズタイプに対して正しく状態空間を推定することができず、推定された状態数は 4 ではなく、観測の総数である 64 であった．したがって、GOEI はコア状態の推定において CEI よりも優れていることが分かった．

GOEI では、適切に状態空間を縮小することで、オンライン学習の速度と非定常な遷移則の推定や報酬の獲得性能が向上する．図 13 と 14 の最上段では、GOEI と CEI の間で最適行動の割合を比較している．GOEI の最初の 5000 ステップでは、状態数が減少するにつれて最適行動の割合が着実に 100 % に近づいた．GOEI は非定常な報酬則環境と遷移則環境の両方で、環境構造に関する学習した知識を忘れることなくダイナミクスの変化に追従することができた．同様に、CEI も報酬則の環境変化には追従することができた（図 13）．しかし、CEI は遷移則の変化には追従できなかった（図 14）．非定常で不確かな環境において、エージェントはモデルの推定と得られたデータの間で観察された不一致が実際の環境ダイナミクスの変化を示しているのか、それとも単なる観察の不確かさによるものかを判断しなければならない．更新幅  $T = 500$  の場合、GOEI では各推定状態が平均して  $500/4 = 125$  の手がかりを提供するのに対し、CEI では各状態が  $500/64 \approx 7.8$  の手がかりしか提供できない．手がかりの数が少ないため、CEI では環境の変化と観察の不確かさを区別することが難しくなる．したがって、コンパクトな状態空間は、動的に変化する環境を頑強に推定するた

めに重要である。

状態集合の適切な縮小は，コア遷移の不確実性に直面した際の戦略学習の頑強性に寄与する（図 15）。コア遷移の不確実性は， $\varepsilon$  が 0 から 0.5 に近づくにつれて増加する。図 15 の最上段と二段目では，GOEI と CEI の間で，最適行動の割合と報酬の損失を示すリグレット [64] を比較している。GOEI は，コア遷移の不確実性が増加しても，一貫して高い最適行動率を維持し，リグレットの増加も無視できる程度に抑えられている。対照的に，CEI はコア遷移の不確実性が増加するにつれて最適行動率が急速に低下し，対応するリグレットの増加も GOEI よりも大きい。GOEI のコア遷移の不確実性に対する頑強性は，現実世界の問題における戦略学習の実現可能性を向上させる。

## 6 考察とまとめ

本論文では、報酬とは無関係な情報を含んだ冗長な観測をもつ強化学習環境において、報酬に関連する有益な情報を抽出する状態削減問題に取り組んだ。実環境における強化学習は、目的の達成に必要なかつ十分な状態空間がエージェントにとって未知であり、そのような状態空間の推定は困難で時間がかかる。逆に、エージェントが限られた経験からこの状態空間を正確に推測することができれば、最適な戦略の学習は容易となる。

DNN は、推定された状態に対して行われる行動の結果を予測することで、複雑な情報をもつ環境下でも戦略学習を実現する。しかし、DNN が生成する状態空間はブラックボックスである。このようなブラックボックスは、大量のデータがあれば学習可能であるが、タスクによっては大量のデータを用意するのは容易ではない。また、ブラックボックスでは、なぜ学習した戦略がうまく機能するのかを説明できず、DNN は説明可能性に欠ける。新奇な環境のように経験したデータ量が限られている場合、過去の学習効果の説明可能性が特に重要になる。

エージェントが目的を達成するのに必要な最小の状態空間を明示的に学習すれば、説明可能性はほとんど単純になる。このことが、強化学習における状態削減問題を定式化する動機となった。状態削減問題の一般的なクラスを解くことは非常に困難であるため、情報の構造を利用したアプローチと、情報の生成過程を利用したアプローチを導入した。1 つ目のアプローチは分離度に基づく次元削減を類似度によって教師なしに拡張したものである。もう 1 つは強化学習における環境推論の新しいクラスである ROMDP を提案するものである。2 つのアプローチは、状態と観測の関係に異なる制約の下で状態の削減を試みるため、次元削減による前処理の後に ROMDP による状態削減を組み合わせることで応用することができる。ROMDP の効率的な手法はあまり研究されておらず、このような手法を開発することは、一般的な強化学習課題を解く上で極めて重要である。

3 章では、観測値の間の類似度によって近似した分離度に基づいて状態削減を行う DLE を提案した。DLE は教師あり次元削減手法である DA と教師なし次元削減手法である LE を組み合わせた手法である。これらの手法を比較することで、LE には平均化制約が不足し



ていることが明らかとなり、分離度を最大にする状態へ削減できない課題が明らかとなった。LE が平均化制約を満たすように改良した DLE は、目的関数の最適化する固有値分解の安定性があり、LE において実用上の課題である解の不安定性を回避できる。

丸め誤差を利用することで、類似度に付加されているノイズを除去する前処理を導入し、DLE と組み合わせることで、分離度がさらに高くなるように状態削減が可能となる。この類似度に対する前処理は、LE に対しては理論的に効果がなく、実験的にも効果がないことが示された。

DLE は状態削減問題に対する情報の構造を利用したアプローチとしてだけでなく、教師なし学習の領域におけるクラスタリングやデータ分析に広く適用することができる。特に、画像領域のセグメンテーションのタスクに応用される Spectral clustering[65] は、データの前処理に LE を用いることが一般的である。その代わりに、この前処理に DLE を採用することで、分離度を最大化するようにクラスタが形成されるので、類似度に付加されたノイズに頑健な性能となることが期待される。

4 章では、ROMDP アプローチを実現するための具体的なアルゴリズムである GOEI を提案した。GOEI の特徴は、ベイズ推論に用いるエージェントモデルの構造が、エージェントが想定する環境構造、すなわち ROMDP とは異なることである。このモデル化の柔軟性はベイズの定理によって保証されており、実際の因果関係とは無関係に条件付き確率の方向を入れ替えることができる。さらに、変分ベイズ法を並行して行い、推定された状態分布を用いて、不一致な状態遷移モデルと望ましい状態遷移モデル、すなわち、それぞれ  $P(a_t | s_{t+1}, s_t)$  と  $P(s_{t+1} | a_t, s_t)$  の確率モデルを推論した。並列した変分ベイズ法は、特定の目的のために設計された様々なグラフィカルモデルに適用可能である。

GOEI のもう一つの特徴は、ディリクレ過程を利用することである。この過程は、環境推論のための初期状態集合を大幅に小さくことができ、ROMDP アプローチで採用されている推論方向  $o_t \rightarrow s_t$  と互換性がある  $P(s_t | o_t)$ 。これに対して、POMDP アプローチでは逆の方向  $s_t \rightarrow o_t$  を仮定しており、標準的な形のディリクレ過程と両立できない。POMDP アプローチで最小状態集合を探索するには、異なる状態数の推論結果をモデル比較する必要がある。

数値実験の結果は、限られた経験から学習するためには、本質的な最小状態空間（コア空間）において観測間の因果関係をモデル化することが有効であることを示している．実際に、冗長な観測を含むマルチアームドバンディット問題において、報酬が被験者の状態空間を削減させる事例が知られている [66]．このような最小状態空間を推定する方法は、不一致モデルに基づく知覚と行動の自由エネルギー原理と解釈することができる．

自由エネルギー原理と新たに提案した目的指向な環境推定手法には、もうひとつ興味深い違いがある．前者はしばしば Active inference を採用し、推論指向で行動を生成する．これに対し、GOEI は目的指向で最適ベルマン方程式を解くことにより行動を決定する．さらに、探索と活用のトレードオフを制御するために、ATS は推定された環境モデルから状態と行動をサンプリングする．重要なことは、GOEI と ATS の併用により、環境のオンライン推定が可能になることである．この手法は自由エネルギー原理とは行動制御の仕方が異なっているが、両者とも環境モデルに依存して探索と活用のバランスをとっているため、ATS と Active inference の間に理論的なつながりが存在する可能性がある．

行動制御に ATS 法を導入し、CEI と GOEI の推定環境モデルに対する最適ベルマン方程式を解いた．環境モデルをサンプリングすることで、エージェントはベイズ推論の精度に依存して探索と活用のバランスを調整することができる．サンプリング法はベイズ推論にのみ適用され、推論された遷移則を持たないブラックボックス環境モデルには適用できない．

目的指向な環境推定による状態削減とその削減された状態に基づいた行動制御のこれらの手法の組み合わせは、コア推定の正確性と戦略学習のサンプル効率性に新たなトレードオフの問題を引き起こす．戦略学習が局所解に陥るケースとして考えられるのは、探索が不十分なときであり、このケースは強化学習の一般的な全ての手法で想定される．このケースに加えて、目的指向な環境推定ではコア推定が失敗したときにも戦略学習が局所界に陥る．従って、コア推定の正確性が十分担保できないようなタスクには、GOEI と ATS の組み合わせは効果的な戦略学習の手法とはならない．特に、報酬確率の分布が統計的に有意に違うことを判別できないタスクの場合では、コアを正しく推定することは困難である．

その他のコア推定を困難にする要因としては、探索的行動により得られる観測系列がコアに対して不均一になっている場合が考えられる．この問題は不均一データのクラスタリング

問題 [67] としても解釈できる．数値実験に用いたテスト環境は，ランダムな探索行動によって全てのコアを満遍なく訪問することができる．一方，現実のタスクでは，意図的に探索しなければ辿りつかないコアが存在する場合が考えられる．動機づけ強化学習 [68] では，環境から与えられる報酬に加えてエージェント内部で意図的な行動を促す興味を報酬として足すことで，意図的な探索行動を促そうとする試みが行われている．不均一な観測によるコア推定の課題には，動機づけ強化学習と不均一データのクラスタリングの手法が解決の糸口になると思われる．

実世界でのエージェントの開発には，オンラインでの処理速度が課題となる．変分ベイズによるモデルの解法は，E-step と M-step を繰り返し計算する．遷移則の学習には状態数の 2 乗に比例する計算量が必要となり，遷移則のメモリ要求が変分ベイズの計算に最も時間がかかる．GOEI は，少ない状態数で遷移則を表現できるため，メモリ要求が少なく，この計算を高速に実行することが可能となる．ただし，5 章における数値実験では，メモリ上には使用していない状態も残っていたため，CEI と GOEI で実行時間による大きな差は見られなかった．推定している状態数に応じて，動的にメモリを使用する工夫を行うことで，オンラインでの処理速度が向上すると考えられる．

今後の展望としては，冗長観測性と部分観測性の両方が存在する環境で効率的に動作するエージェントを開発することが挙げられる．観測空間が構造化された実世界のタスクでは，DLE などの次元削減によるアプローチであらかじめ観測空間を低次元化することが可能である．さらに ROMDP を Factored-MDP や Low-rank MDP と組み合わせることで，観測のスケーラビリティと状態空間の説明可能性を向上させることができる．効率的な状態削減のために必要な履歴の長さを明らかにすることで，コア推定の困難さを軽減し，効率的に解ける強化学習課題のクラスを POMDP と ROMDP の混合に拡張することへの貢献が期待される．

## 謝辞

本論文を完成させるにあたり，多くの方々のご支援とご指導を賜りましたことに深く感謝申し上げます。

まず初めに，工学院大学大学院の竹川高志教授には，研究の指導のみならず，常に温かい励ましと貴重な助言をいただきましたことを心より感謝申し上げます。教授の洞察力と指導力により，本研究の質の高める大きな助けとなりました。そして，本論文の執筆にあたり様々なご助言をいただいた，工学院大学の大和淳司教授，真鍋義文教授，田中久弥教授，並びに玉川大学の鮫島和行教授に深く感謝申し上げます。

また，本研究において共同研究者としてご協力いただいた沖縄科学技術大学院大学の深井朋樹教授と玉川大学の酒井裕教授には，深い知識と経験に基づいたご指導をいただきましたことに感謝いたします。お二人のご支援とご助言が本研究の発展に大きく寄与しました。

最後に，最も身近で支えてくださった家族に心より感謝申し上げます。

## 参考文献

- [1] Helga Kolb, Eduardo Fernandez, and Ralph Nelson. Webvision: the organization of the retina and visual system [internet]. 1995.
- [2] Rufin VanRullen and Simon J Thorpe. The time course of visual processing: from early perception to decision-making. *Journal of cognitive neuroscience*, 13(4):454–461, 2001.
- [3] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2493–2500, 2020.
- [4] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214:106685, 2021.
- [5] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.
- [6] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.
- [7] Kuniyiko Fukushima, Sei Miyake, and Takayuki Ito. Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (5):826–834, 1983.
- [8] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.
- [9] Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*, 2017.

- [10] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.
- [11] A Asuntha and Andy Srinivasan. Deep learning for lung cancer detection and classification. *Multimedia Tools and Applications*, 79(11):7731–7762, 2020.
- [12] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [13] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.
- [14] Carl Rasmussen. The infinite gaussian mixture model. *Advances in neural information processing systems*, 12, 1999.
- [15] Huan Wan, Hui Wang, Bryan Scotney, and Jun Liu. A novel gaussian mixture model for classification. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 3298–3303. IEEE, 2019.
- [16] Stephane Ross, Brahim Chaib-draa, and Joelle Pineau. Bayes-adaptive pomdps. *Advances in neural information processing systems*, 20:1225–1232, 2007.
- [17] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.
- [18] Karl Friston, Lancelot Da Costa, Danijar Hafner, Casper Hesp, and Thomas Parr. Sophisticated inference. *Neural Computation*, 33(3):713–763, 2021.
- [19] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [20] Paul W Glimcher. Understanding dopamine and reinforcement learning: the

- dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, 108(supplement\_3):15647–15654, 2011.
- [21] Sunyong Kim and Hyuk Lim. Reinforcement learning based energy management algorithm for smart energy buildings. *Energies*, 11(8):2010, 2018.
  - [22] Bibhas Chakraborty and Susan A Murphy. Dynamic treatment regimes. *Annual review of statistics and its application*, 1:447–464, 2014.
  - [23] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
  - [24] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pages 737–744. IEEE, 2020.
  - [25] Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 40(4-5):698–721, 2021.
  - [26] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sal-lab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.
  - [27] Yevgen Chebotar, Ankur Handa, Viktor Makoviyshuk, Miles Macklin, Jan Issac, Nathan Ratliff, and Dieter Fox. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8973–8979. IEEE, 2019.
  - [28] Nick Jakobi, Phil Husbands, and Inman Harvey. Noise and the reality gap: The use of simulation in evolutionary robotics. In *Advances in Artificial Life: Third European Conference on Artificial Life Granada, Spain, June 4–6, 1995 Proceedings*

- 3, pages 704–720. Springer, 1995.
- [29] Kazuki Takahashi and Takashi Takekawa. Discriminant laplacian eigenmaps by the approximation of discriminant analysis using similarity. *Nonlinear Theory and Its Applications, IEICE*, 13(2):300–305, 2022.
  - [30] Kazuki Takahashi, Tomoki Fukai, Yutaka Sakai, and Takashi Takekawa. Goal-oriented inference of environment from redundant observations. *Neural Networks*, 174:106246, 2024.
  - [31] Haiying Wan, Xiaoli Luan, Vladimir Stojanovic, and Fei Liu. Self-triggered finite-time control for discrete-time markov jump systems. *Information Sciences*, 634:101–121, 2023.
  - [32] Zhiyuan Chen and Bing Liu. *Lifelong machine learning*, volume 1. Springer, 2018.
  - [33] Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022.
  - [34] Zhihe Zhuang, Hongfeng Tao, Yiyang Chen, Vladimir Stojanovic, and Wojciech Paszke. An optimal iterative learning control approach for linear systems with nonuniform trial lengths under input constraints. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(6):3461–3473, 2022.
  - [35] Vladimir Stojanović. Fault-tolerant control of a hydraulic servo actuator via adaptive dynamic programming. *Mathematical Modelling and Control*, 3(3):181–191, 2023.
  - [36] Kenji Doya, Kazuyuki Samejima, K Katagiri, and Mitsuo Kawato. Multiple model-based reinforcement learning. *Neural computation*, 14(6):1347–1369, 2002.
  - [37] Zhi Wang, Chunlin Chen, and Daoyi Dong. Lifelong incremental reinforcement learning with online bayesian inference. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8):4003–4016, 2021.
  - [38] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic pro-*



- gramming*. John Wiley & Sons, 2014.
- [39] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.
  - [40] Anthony R Cassandra. A survey of pomdp applications. In *Working notes of AAAI 1998 fall symposium on planning with partially observable Markov decision processes*, volume 1724, 1998.
  - [41] Xiaofei He and Partha Niyogi. Locality preserving projections. *Advances in neural information processing systems*, 16, 2003.
  - [42] Alnour Alharin, Thanh-Nam Doan, and Mina Sartipi. Reinforcement learning interpretation methods: A survey. *IEEE Access*, 8:171058–171077, 2020.
  - [43] George Konidaris. On the necessity of abstraction. *Current Opinion in Behavioral Sciences*, 29:1–7, 2019.
  - [44] Malcolm Strens. Efficient hierarchical MCMC for policy search. In *Proceedings of the twenty-first international conference on Machine learning, ICML '04*, pages 97–104, 2004.
  - [45] Finale Doshi-Velez. The infinite partially observable markov decision process. *Advances in neural information processing systems*, 22:477–485, 2009.
  - [46] Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored mdps. In *IJCAI*, volume 16, pages 740–747, 1999.
  - [47] Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.
  - [48] Rolf A. N. Starre, Marco Loog, Elena Congeduti, and Frans A Oliehoek. An analysis of model-based reinforcement learning from abstracted observations. *Transactions on Machine Learning Research*, 2023.
  - [49] David Abel, Dilip Arumugam, Lucas Lehnert, and Michael Littman. State abstractions for lifelong reinforcement learning. In *Proceedings of the 35th International*

*Conference on Machine Learning*, volume 80, pages 10–19. PMLR, 2018.

- [50] David Abel, Dilip Arumugam, Kavosh Asadi, Yuu Jinnai, Michael L Littman, and Lawson LS Wong. State abstraction as compression in apprenticeship learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3134–3142, 2019.
- [51] Cameron Allen, Neev Parikh, Omer Gottesman, and George Konidaris. Learning markov state abstractions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 34:8229–8241, 2021.
- [52] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.
- [53] Christopher Grimm, André Barreto, Greg Farquhar, David Silver, and Satinder Singh. Proper value equivalence. *Advances in Neural Information Processing Systems*, 34:7773–7786, 2021.
- [54] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.
- [55] Anil K Jain, Robert P. W. Duin, and Jianchang Mao. Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1):4–37, 2000.
- [56] Deng Cai, Xiaofei He, and Jiawei Han. Efficient kernel discriminant analysis via spectral regression. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 427–432. IEEE, 2007.
- [57] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.
- [58] Yasser Ghanbari, Panos E Papamichalis, and Larry Spence. Graph-laplacian features for neural waveform classification. *IEEE transactions on biomedical engineer-*

- ing, 58(5):1365–1372, 2010.
- [59] David M Blei and Michael I Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 2006.
  - [60] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.
  - [61] Takashi Takekawa and Tomoki Fukai. A novel view of the variational bayesian clustering. *Neurocomputing*, 72(13-15):3366–3369, 2009.
  - [62] M Sato. Online model selection based on the variational bayes. *Neural computation*, 13(7):1649–1681, 2001.
  - [63] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 39.1–39.26, Edinburgh, Scotland, 25–27 Jun 2012. PMLR.
  - [64] Yasin Abbasi-Yadkori, András György, and Nevena Lazić. A new look at dynamic regret for non-stationary stochastic bandits. *Journal of Machine Learning Research*, 24(288):1–37, 2023.
  - [65] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17:395–416, 2007.
  - [66] Aurelio Cortese, Asuka Yamamoto, Maryam Hashemzadeh, Pradyumna Sepulveda, Mitsuo Kawato, and Benedetto De Martino. Value signals guide abstraction during learning. *eLife*, 10:e68943, 2021.
  - [67] XU Xiaolong, CHEN Wen, and SUN Yanfei. Over-sampling algorithm for imbalanced data classification. *Journal of Systems Engineering and Electronics*, 30(6):1182–1191, 2019.
  - [68] Nuttapon Chentanez, Andrew Barto, and Satinder Singh. Intrinsically motivated

reinforcement learning. *Advances in neural information processing systems*, 17, 2004.