様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月 １７ 日現在

機関番号：９４３０１
研究種目：挑戦的萌芽研究
研究期間：2016～2018
課題番号：１６Ｋ１２５０４

研究課題名（和文）カルバックライブラー制御法と内的報酬を統合した強化学習

研究課題名（英文）Integration of Kullback-Leibler control and intrinsic rewards for reinforcement learning

研究代表者
内部　英治（UCHIBE, Eiji）

株式会社国際電気通信基礎技術研究所・脳情報通信総合研究所・主幹研究員

研究者番号：２０４２６５７１

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：サンプル効率の良い二つの強化学習アルゴリズム（適応的ベースラインを持つEMアルゴリズムによるハイパーパラメータ探査法EPHEと方策探査のための適応的多重重点サンプリングAMIS）を開発した。EPHEは決定論的方策を探査し、倒立二輪型移動ロボットに実装した。実験結果よりEPHEは標準的な方策探査法よりもサンプル効率が良いことが示された。AMISは方策探査法が過去に収集したデータを再利用する際に多重重点サンプリングを用いた推定量の分散を削減する。AMISはEPHEを含む様々な方策探査法に適用可能で、通常よりもさらにサンプル効率を改善できることを示した。さらにスマートフォンロボットを開発した。

研究成果の学術的意義や社会的意義
学術的意義はロボットの制御器の学習に適した強化学習アルゴリズムを開発したことである。通常の強化学習アルゴリズムでは確率的な制御則を用いることが多いが、生成行動系列は滑らかではないためロボットの制御には適さない。EPHEは決定論的な制御則を学習するため滑らかな行動系列を生成でき、スマートフォンロボットのような高精度のアクチュエータを持たないシステムにも適用できる。
社会的意義はデータの高効率性を実現したことである。現実的な問題設定では学習に利用できるデータは限られており、様々なアルゴリズムと組み合わせて使用可能なAMISは強化学習アルゴリズムを実問題に適用する際に重要な構成要素となると期待できる。

研究成果の概要（英文）：We have developed sample-efficient reinforcement learning algorithms: EM-based Policy Hyperparameter Exploration (EPHE) with adaptive baseline and Adaptive Multiple Importance Sampling (AMIS) for Policy Search. EPHE optimizes deterministic policies by EM algorithm and it was implemented in a wheeled inverted pendulum mobile robot. Experimental results showed that EPHE outperformed standard policy search methods. AMIS reduces the variance of the estimator based on multiple importance sampling when policy search algorithms tries to reuse samples that are collected in previous iteration steps. AMIS is evaluated with several policy search methods such as EPHE, REINFORCE, REPS, CMA-ES, and NES and experimental results showed that AMIS improved sample efficiency for all the algorithms. Besides we developed experimental platform based on smartphone and some basic behaviors such as battery foraging and mating based on visual communication are implemented by reinforcement learning.

研究分野：強化学習

キーワード：強化学習　EMアルゴリズム　ロボット学習　スマートフォンロボット　逆強化学習　進化計算

１．研究開始当初の背景

研究開始時において、最適制御と確率推論の双対性に関して理論的な進捗があり、あるクラスの確率最適制御問題をグラフィカルモデル上での推論問題に変換できることが示された。カルバックライブラー（KL）制御と呼ばれるこの手法の特徴は、方策（制御則）の即時的な評価を与える報酬に情報理論的な制約を与えた点である。研究代表者はこの理論の実ロボットへの応用に関して研究してきた。KL 制御はベルマン方程式を線形化するために報酬の一部を KL ダイバージェンスによって制約し、確率推論を用いて近似された最適方策を推定する。線形化の利点により方策は効率よく求めることができるが、データ収集に関しては特に考慮されていないため学習時の探査効率は悪い。一方で、内的報酬を用いた強化学習と呼ばれる研究分野では、タスクに依存しない環境探査のための報酬（目的関数）の設計論について議論している。報酬は環境とのインタラクションから得られる「現在の状態の目新しさ」や環境モデルの不確実性などを情報理論によって定量化したものである。研究代表者は研究開始時までに進化的手法と組み合わせた手法を提案し、内的報酬の利用が有効であることを示した。ただし無駄な探索を省くことによりデータ収集は効率化されるが、方策を求めるには従来と同様に非線形ベルマン方程式を解く必要があり、計算効率は悪いままである。これら二つの研究分野は最適制御において、目的関数を情報理論の観点から設計する点において類似しているが、両者の関係は明確ではなかった。

２．研究の目的

そこで本研究では両者を統合した新しい制御則の学習方法を提案することを目的とした。本研究で提案する手法は KL 制御と内的報酬による強化学習がともに情報理論の観点から報酬を修正することによって導出されたことに着目する。つまり、KL 制御を導出するために必要な報酬の表現を拡張し、内的報酬として計算される幾つかの報酬を追加し、KL 制御の枠組みを用いて定式化する。これにより図１のようにデータ収集の探査効率と制御則を求めるための計算効率の両方を改善した、新しい行動学習法が提案できる。

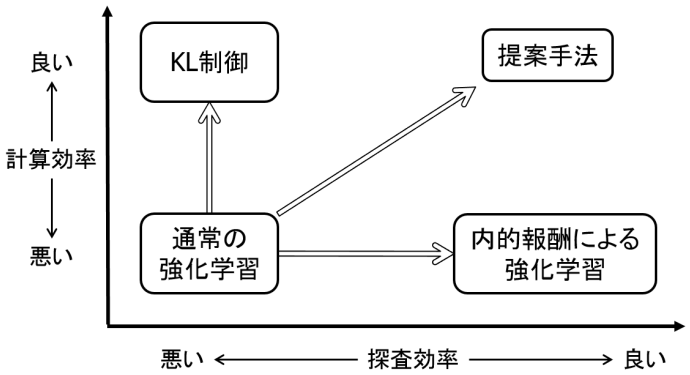

図 1　KL 制御と内的報酬の統合による提案手法

KL 制御は確率最適制御に端を発し、機械学習や物理学の考えを導入した手法である。一方、内的報酬を用いた強化学習は認知科学や心理学、神経科学といった分野から着想を得ている。意思決定のメカニズムを明らかにする、といった点で言えば、これらの分野は良く似た問題を研究対象としているが、それぞれ異なる分野で研究されているため、研究者間の交流もあまり盛んではなく、両者の類似性を指摘した研究は申請者の知る限り存在しない。両者を統合することで、効率の良い制御則の学習法を提案するだけでなく、異分野の研究者をつなぐ新しい研究分野を創発できると期待した。

３．研究の方法

研究開始時当初は線形可解マルコフ決定過程（LMDP）に基づいた KL 制御を想定していた。このときベルマン最適方程式が簡単化され、密度比推定による逆強化学習を用いてエキスパートからのデータから報酬を推定する枠組みを提供し、衰退された報酬をもとに強化学習を適用することで効率的な学習を実現する。これは研究成果（4）に示すもので、今後敵対的模倣学習として発展が望まれるものである。内的報酬の利用のためには環境モデルの情報が必要で、通常の強化学習が想定する問題設定ではモデル学習が必要になる。いくつかのモデル学習法を用いて検証し、モデル化誤差が及ぼす影響を調査する。

４．研究成果

(1) KL 制御のロバスト化

通常の KL 制御では環境のダイナミクスから導出される受動ダイナミクスが重要な役割を果たすが、これまではダイナミクスのモデル化誤差が最終的に導出される方策に与える影響について調査されていなかった。Hamilton-Jacobi-Bellman 方程式を拡張した Hamilton-Jacobi-Isaacs (HJI)方程式を用いることでゲーム理論の考えを KL 制御に導入し、KL ダイバージェンスの代わりに Renyi ダイバージェンスを用いることで HJI 方程式を線形化する方式において、環境のモデルパラメータの変動や遷移確率の分散が及ぼす影響について調査した。その結果、離散状態･行動の場合はモデル化誤差の影響を軽減するパラメータαの設計が容易である一方、連続状態・行動の場合は価値関数の近似誤差のためにパラメータαの設計に別の指標を導入する必要があることが判明した。図 2 はその一例で、αと方策の性能（ここでは倒立振子課題に

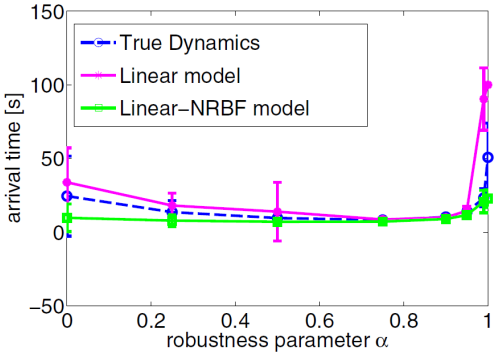

図 2　ロバストネスパラメータと制御性能の関係

おける初期状態から倒立状態までの経過時間）の関係を示したもので、真のダイナミクス、線形モデルを使った場合、線形-NRBF モデルを使った場合を比較している。線形モデルのような近似誤差が大きいモデルを使ってもαを大きくすることで制御を達成しているが、αが１に近づくにつれ急激に性能が悪化している。離散問題の場合には見られなかったこの現象を確認できた点が大きな成果の一つである。

(2) 環境探査のための報酬の符号の分離

報酬値を符号に応じて分離する強化学習法 MaxPain を深層強化学習化した Deep MaxPain を開発した。Deep MaxPain は最下層の畳み込み層は共有するが価値関数を学習する全結合層は独立している。そのため正の報酬から学習するための経験と負の報酬から学習するための経験を個別に保存し、学習時に同じ重みで混合することで学習が安定になり、ニューラルネットワークを用いた関数近似と MaxPain を統合することに成功した。

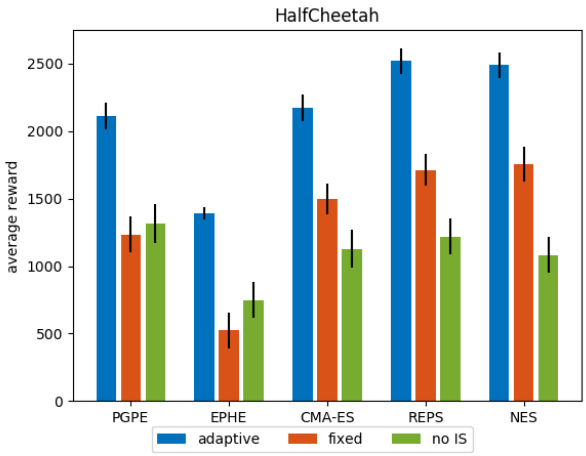

図 3　重点サンプリングに用いる混合分布の混合係数を適応的にした場合、固定した場合、および経験を再利用しない場合の性能比較

(3) 適応的多重重点サンプリングによる経験の再利用

方策探査法は多くのアルゴリズムが方策オン型であり過去の経験を再利用するためには重点サンプリングを用いた補正が必要になるが、単純な重点サンプリングの使用は学習を不安定化させる。そこで重点サンプリングによる推定値の分散を最小にするように過去のデータ収集分布の結合重みを修正する適応的重点サンプリング法を開発した。PGPE、EPHE、CMA-ES、REPS、NES という 5 種類の代表的な方策探査法に適用しデータ効率が改善できることを示した。図 3 はベンチマーク課題 HalfCheetah をタスクとして選んだ時の、結合重みを適応的にした場合、固定した場合、および過去の経験を再利用しない場合の制御性能を比較したものである。すべての方策探査法において提案手法で結合重みを適応的にした場合、性能が大幅に改善できていることがわかる。とくに EPHE の場合は過去に取集したデータを単純に利用するよりも利用しない場合もあることが確認でき、重みを適切に設定することは極めて重要であることも確認できた。

(4) 密度比推定による逆強化学習法

これまでに開発されてきた逆強化学習システムは報酬関数だけを推定するものであったこと、逆強化学習と（順）強化学習が異なる仮定の下で導出されていたため、両者の間で学習結果を共有することができなかった。開発した手法では報酬関数だけでなく状態価値関数も同時に推定するため、推定した状態価値関数を強化学習の学習時の高速化に利用できる。同時に強化学習の結果を逆強化学習の報酬と状態価値の推定時に相互に利用することが可能となった。開発した手法も敵対的模倣学習（Generative Adversarial Imitation Learning; GAIL）の拡張と解釈できるが、GAIL よりも順強化学習時における環境とのインタラクション回数を大幅に削減できることを示した。

(5) EM アルゴリズムを用いた強化学習におけるベースラインの調節

決定論的方策のパラメータを最適化する方策探査法において、パラメータを更新する際にすべてのサンプルを用いると学習効率が悪くなることが知られていた。そこで本研究では方策改善にあまり貢献しないサンプルを切り捨てるための閾値を動的に調整する方法を開発した。実験結果を解析したところ、学習の初期段階ではほとんどタスクを達成できない多数のサンプルから形成されるピークと、少しだけ達成できる少数のサンプルから形成されるピークの二つのピークを持つことが多く、提案手法はこのようなケースで特に有効に働くことを確認した。スマートフォンロボットを用いた倒立課題、目的地までのナビゲーション課題ともにシミュレー

タを使用することなく学習できることを示した。

(6) マルチエージェント強化学習におけるコミュニケーションの創発

マルチエージェント強化学習ではエージェント数に応じて状態数が指数的に増加し、学習が極めて困難になるという問題がある。他のエージェントの状態を直接表現する代わりに、低次元に圧縮した情報を通信することで状態数を大幅に削減するモジュール強化学習法を開発した。図 4 に開発したマルチエージェント強化学習の情報の流れを示す。i 番目のエージェント(i=1,2)は j 番目のエージェントにメッセージを送信するための行動価値 Q_i^c と行動を決定する行動価値 Q_i^p の二つを持ち、Q_i^c によってえらばれた信号 a_i^c が Q_j^p の状態料として利用される。他者の行動価値各エージェント内の構造もマルチエージェント強化学習になっている。標準的なマルチエージェント強化学習と比較し、学習に必要なデータ数を削減することができ、学習過程も安定化させることができた。

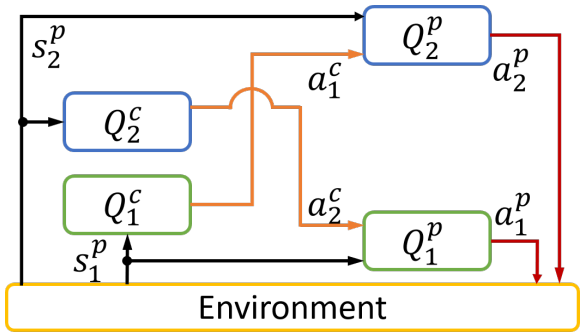

図 4　マルチエージェント強化学習における情報の流れ

(7) スマートフォンロボットを用いた自律分散協調ロボットシステムの開発

強化学習のメタパラメータの影響を調査するためには、メタパラメータの値の異なる学習システムを並列に実行し学習過程を比較する方法が有効である。また、実ロボットにおける学習のサンプル効率を改善するために、複数学習システムのためのアルゴリズムの開発が重要で、検証のためにロボット実験システムを改良した。本年度は実ロボット上で外部バッテリからの充電行動、および交配行動実現のためのロボット間での視覚情報を通した情報交換の行動を方策探査法によって実現した。図 5 は実際の交配行動の例である。Receiver は受信可能であることを示すハートを液晶画面に表示し、Sender はそれに応じてメッセージを QR コードに変換し液晶画面に表示する。Receiver は QR コードを認識することでメッセージを受け取る。このような局所的な通信は Embodied evolution を研究する上で必要不可欠な構成要素であり、開発したロボットは Embodied evolution を研究するためのプラットフォームとして有効であることを確認した。

図 5　スマートフォンロボットにおける交配行動の例

５．主な発表論文等

〔雑誌論文〕（計 4 件）

E. Uchibe. Cooperative and Competitive Reinforcement and Imitation Learning for a Mixture of Heterogeneous Learning Modules. Frontiers in Neurorobotics, 2018. DOI: 10.3389/fnbot.2018.00061

K. Kinjo, E. Uchibe, and K. Doya. Robustness of linearly solvable Markov games employing inaccurate dynamics model. Artificial Life and Robotics. Vol. 23, pp. 1-9, 2018. DOI: 10.1007/s10015-017-0401-2

内部英治, 王潔心. ロボット制御のための決定論的方策探査法. 日本神経回路学会誌. Vol. 24, pp. 195-203, 2017. DOI: 10.3902/jnns.24.195

J. Wang, E. Uchibe, and K. Doya. Adaptive Baseline Enhances EM-Based Policy Search: Validation in a View-Based Positioning Task of a Smartphone Balancer. Frontiers in Neurorobotics, 2017. DOI: 10.3389/fnbot.2017.00001

〔学会発表〕（計 8 件）

E. Uchibe. Imitation learning under entropy regularization. Workshop on Reinforcement Learning & Biological Intelligence（招待講演）, 2019.

E. Uchibe. Cooperative and competitive reinforcement and imitation learning. The 8th Joint IEEE International Conference on Development and Learning and Epigenetic Robotics, 2018.

J. Wang, S. Elfwing, and E. Uchibe. Deep reinforcement learning by parallelizing reward and punishment using MaxPain architecture. The 8th Joint IEEE International Conference on Development and Learning and Epigenetic Robotics, 2018.

E. Uchibe. Efficient sample reuse in policy search by multiple importance sampling. Genetic and Evolutionary Computation Conference. 2018. DOI: 10.1145/3205455.3205564

内部英治. 方策探査法のための多重重点サンプリングを用いた経験再利用. ロボティクス・メカトロニクス講演会. 2018. DOI: 10.1299/jsmermd.2018.1A1-E13

J. Wang, and E. Uchibe. EM-based policy search for learning foraging and mating behaviors. ロボティクス・メカトロニクス講演会. 2018. DOI: 10.1299/jsmermd.2018.1A1-E17

E. Uchibe. Forward and inverse reinforcement learning and generative adversarial formulation. NC/IBISML/IPSJ-MPS/IPSJ-BIO 合同研究会（招待講演）, 2018.

Q. Huang, E. Uchibe, and K. Doya. Emergence of communication among reinforcement learning agents under coordination environment. The 6th Joint IEEE International Conference on Development and Learning and Epigenetic Robotics, 2016. DOI: 10.1109/DEVLRN.2016.7846790

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）
○取得状況（計　0　件）

〔その他〕（計　2　件）
S. Elfwing, E. Uchibe, and K. Doya. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. arXiv: 1702.03118, 2017.

S. Elfwing, E. Uchibe, and K. Doya. Online Meta-learning by Parallel Algorithm Competition. arXiv: 1702.07490, 2017.

6．研究組織
(1)研究分担者　なし
(2)研究協力者　なし