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　６ 年　 ５ 月 ２３ 日現在

機関番号：１４６０３
研究種目：基盤研究(B)（一般）
研究期間：2021～2023
課題番号：２１Ｈ０３５２２

研究課題名（和文）安全性と信頼性を備えたロボット強化学習の技術基盤の創出

研究課題名（英文）Creation of a technological platform for robot reinforcement learning with safety and reliability

研究代表者

松原　崇充（Matsubara, Takamitsu）

奈良先端科学技術大学院大学・先端科学技術研究科・教授

研究者番号：２０５０８０５６

交付決定額（研究期間全体）：（直接経費）　　13,400,000円

研究成果の概要（和文）：本研究では、ロボットが環境や道具との接触を伴う物理的作業を学習する際に要求される安全性と信頼性を備えた強化学習技術基盤を提案した。特に、試行錯誤時における衝突リスクの低減による安全性や、経験サンプル不足等による方策改善時の方策振動を抑制する信頼性を実現するための理論やアルゴリズムを開発した。さらに実ロボットを用いた複数の物理接触を含む作業タスクに適用し、その有効性を検証した。

研究成果の学術的意義や社会的意義
本研究では、労働力不足の問題が深刻化する人口減少や超高齢社会において、ロボットを効果的に活用するための強化学習技術基盤を開発した。その成果により、ロボットが環境や道具との物理的接触を伴う作業を、より安全かつ効率的に学習可能なった。今後は、部品組み立てや調理など、実世界の様々な産業やサービスへの応用が期待される。この技術は、ロボットの普及と実用化を促進し、社会的にも大きな意義を持つと考えられる。

研究成果の概要（英文）：In this study, we introduced a reinforcement learning framework that provides the necessary safety and reliability for robots to learn physical tasks involving contact with the environment and tools. Specifically, we developed theories and algorithms to enhance safety by reducing collision risks during trial and error, and to ensure reliability by alleviating policy oscillations due to insufficient experience samples. Additionally, we applied this framework to various tasks involving physical contact using actual robots and validated its effectiveness.

研究分野：知能ロボティクス

キーワード：強化学習　試行錯誤　安全性　信頼性　エントロピー正則化　単調方策改善　ロボットラーニング

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

人口減少や超高齢社会における労働力不足の問題に対して、産業とサービスの両面でロボット活用に期待が高まっている。ロボットを人作業者の代替として活用するためには、状況認識や行動選択に関する知能化が不可欠である。そこで、環境との相互作用により収集される自身の経験サンプルから、最適な行動ルール（以降、方策）を学習可能な人工知能の枠組みである強化学習が注目されている。実世界の産業・サービスにおける人作業の多くは、環境や道具との物理的な接触を伴う接触豊富（コンタクトリッチ）な作業である。しかし、ビデオゲームやシミュレーションなどのサイバー世界で発展した現行の強化学習の枠組みは、実世界における物理的作業の学習には適さない。その具体的な理由として、本研究では以下の 2 点に焦点を当てる。

1) 試行錯誤中の安全性：強化学習が目覚ましい発展を遂げたサイバー世界では、物理的接触によるロボットや環境の故障・破損の心配がない。よって、試行錯誤の際、ロボットが確率的な探索行動を自由に実行し、大量かつ豊富な経験サンプルを収集しても何ら問題とならない。一方、実世界ロボットが接触豊富な作業スキルを学習する場合、行動の安全性が担保されなければ確率的に選択される探索行動を安易に実行できない。

2) 学習の信頼性：一般的な強化学習は方策を徐々に改善する「単調改善」を目標に設計されているが「膨大な経験サンプル」による学習を前提とする。一方、実世界ロボットが収集可能な「少量の経験サンプル」から学習する際には、データ不足により様々な誤差が生じるため、方策改善は確率的な挙動を示す（方策振動現象）。時には方策を大幅に性能劣化させる。実世界での応用、特に物理的接触を伴う作業では、先述の安全性の問題にも影響するため、そのような信頼性の乏しい学習過程を安易に実行できない。

２．研究の目的

本研究の目的は、実世界のロボットシステムが自らの試行錯誤によって収集する経験サンプルから、環境や道具との物理的接触を伴う作業を学習可能にする技術基盤の確立である。そのために、「安全性」と「信頼性」を備えた新しい強化学習の理論および技術基盤の確立を目指す。物理接触豊富な作業として「調理作業」などの実用的な作業タスクに焦点をあて、ロボットを含めた実験環境を構築する。学習実験を通じて、提案する理論および技術基盤の有効性を示すとともに、潜在する問題点や今後の発展性についても明確化する。

３．研究の方法

主に次の３つの課題解決に取り組んだ。

(1) 試行錯誤の安全性を担保する理論・アルゴリズム導出
ロボットが試行錯誤する際、物理接触によるロボット・環境の破損・故障リスクを見積って、危険な接触を回避する理論的枠組みを構築する。未知の経験を取得する探索行動を実行しなければ学習に有益な経験サンプルを収集できないが、破損・故障のリスクが存在するというジレンマがある。このジレンマを解消する本研究のアイデアは、「故障リスク＝制御入力強度」および「未経験の領域＝環境不確実性が高い」と解釈し、環境不確実性に応じた制御入力強度の動的制約を導入することである。環境不確実性が高ければ行動を制限し安全性を確保し、一方、不確実性が低ければ制限なく自由に行動させる。このような方針は、不確実環境下における生物の行動原理としても示唆されている。環境不確実性を経験サンプルから取り込むためにガウス過程などの確率モデルをベースに、計算量・メモリ量に配慮して、実ロボットのリアルタイム制御に適した枠組みを検討する。

(2) 学習の信頼性を担保する理論・アルゴリズム導出
学習の信頼性の指標として、更新された方策の改善性を実験試行無しで予測する指標：Expected Policy Advantage: EPA に注目する。ただし現行の EPA は、更新された方策が持つ期待累積報酬値や状態定常分布など、未知環境では知り得ない変量を必要とするため実用向きの枠組みではない。そこで本研究では、実世界ロボット応用に適した新しい EPA の導出を試みる。その鍵となるアイデアは、方策の更新時にその更新幅に応じたコスト関数を付与し、許容される方策更新幅に制限を加えることである。更新幅制限は「少数の経験サンプル」で学習する際に、学習を安定化する工夫として知られている。ここではその性質を期待累積報酬や状態定常分布の変化を捉える理論の導出に活用し、計算容易な EPA の導出を狙う。

(3) 開発した強化学習基盤の複数ロボットタスクにおける実験検証
物理接触を豊富に伴う作業の例として「調理作業」などの複数の作業タスクを対象に、実ロボッ

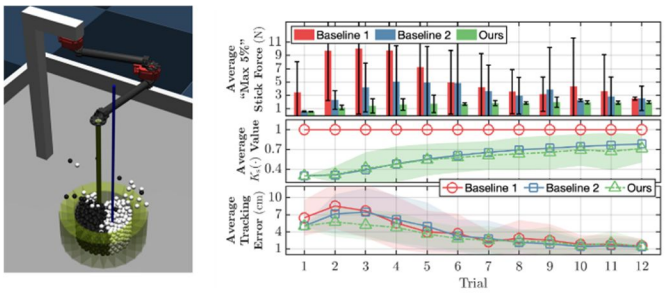
トの作業学習課題を設定し、実世界学習実験を通じて有効性を検証する。

４．研究成果

(1) に関する成果：

接触が頻繁に発生する環境でのロボット応用を目的としたモデルベース強化学習（MBRL）アプローチを開発した。従来の MBRL アプローチがモデルの不正確さにより安全でない制御を適用する可能性があるのに対し、この枠組みでは学習過程での安全性を確保しながら効率的にタスクを学習できるように設計されている。モデル予測制御による行動計画時に、ロボットの行動制限をモデルの不確実性と関連付けることで、学習進行に応じて制御行動を制約する。このアプローチにより、モデルの不確実性が高い場合には小さく安全な行動を取り、予期しない接触の強度を低減できる。その他、LGM-FF（Linear Gaussian Model with Fastfood features）を使用して、ロボットの状態遷移モデルを学習する。この工夫により、サンプルサイズが増加しても計算効率を維持しながら高頻度での制御が可能となった。

図１にシミュレーション結果の一部を示す。対象タスクとして、コンタクトリッチな調理タスク（ボウル内容物の混合タスクや掬いタスク）を設定した。シミュレーションを通じて、開発した安全性を担保する工夫が、ロボットの学習プロセスと接触強度に及ぼす影響を調査した。結果として、学習する状態遷移モデルの予測分散に応じて行動上限を制限することで、学習の進展を阻害することなく、学習初期に見られる危険な行動を回避できることを確認した。



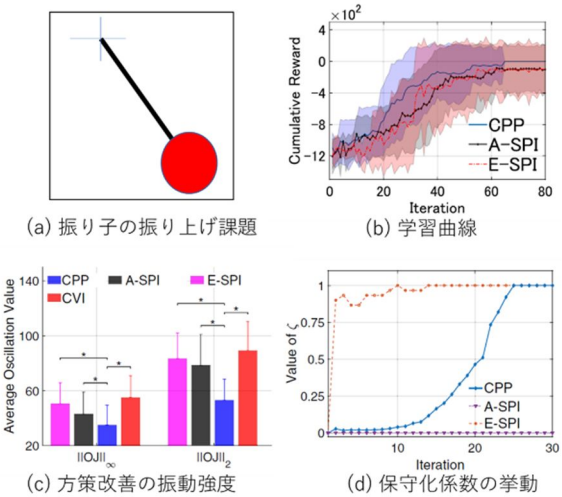
(a) 混合作業タスク　　(b) 方策改善試行と危険行動割合の推移

図 1 安全性を考慮した学習の効果

(2) に関する成果：

学習の信頼性を考慮した強化学習手法として「Cautious Policy Programming (CPP)」と呼ぶ新しい強化学習アルゴリズムを開発した。このアルゴリズムでは、学習中の方策改善を保証することに焦点を当て、方策改善性を見積もる新しい基準（従来型の下界）を導出した。具体的には、エントロピー正則化を取り入れた方策クラスに限定することで、従来の基準における計算困難であった変量を簡略化し、より扱いやすい下界を導出している。さらに、方策振動の度合いを調整するために、この新しい下界を基準として利用して振動を軽減できるメカニズムを採用した。これにより、少量の経験サンプルから方策改善を実行する場合にも信頼性の高い方策改善が期待できる。

単振り子の振り上げ課題に提案手法を適用し、その有効性を検証した。その結果を図２に示す。学習曲線より、ベースラインと比べ学習が効率化され、性能分散が低減されていることが確認できる。さらに、２つの基準により方策振動を評価したところ、ベースラインと比べて振動が抑制されたことを確認した。その他、方策振動の度合いを調整するメカニズムの調整係数の挙動も期待通りであったことを合わせて確認した。



(a) 振り子の振り上げ課題　(b) 学習曲線

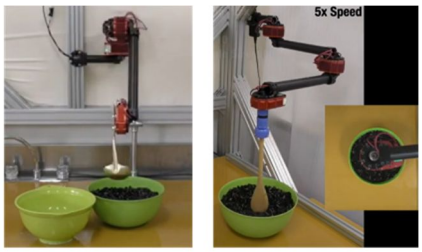
(c) 方策改善の振動強度　(d) 保守化係数の挙動

図 2　信頼性を考慮した学習の効果

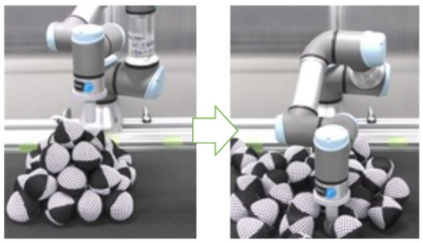
(3) に関する成果：

開発したロボット強化学習基盤の実環境における有効性を検証するために、物理接触を伴う作業用のロボットアーム環境を用いた実験を実施した。特に、キッチン環境を模した環境における簡易調理作業や、ビーズボールの山崩し作業の学習タスクを設計し、各強化学習アルゴリズムを適用した（図３参照）。両環境・タスクにおける実験結果より、物理シミュレーション実験と同様に、開発した強化学習基盤を適用することで、試行錯誤中の予期せぬ物理的接触や衝突に対処する「安全性」、学習により方策をより確実に改善する「信頼性」が得られることを確認した。さらに、複数の環境・タスクへの適用を通じて、一定の汎用性についても確認できた。

その他、計画時には予期していなかった成果として、学習の信頼性を向上させるために導出した Expected Policy Advantage（EPA）を、近年注目を浴びる sim-to-real 方策転移のためのドメインランダム化強化学習の学習効率化に利用可能であることを見出した。単調増加性を指向した方策学習則をベースに、物理パラメータの変更による異なる環境下で学習された方策・価値を活用する強化学習の安定化アルゴリズムを開発した。様々な sim-to-sim および sim-to-real タスクへの適用実験を通じて、開発したアルゴリズムの有効性（許容されるランダム化範囲の拡大、学習の安定化など）を確認した。



(a) キッチン作業タスク（左：掬い、右：混合）



(b) ボール山崩しタスク

図 3　実環境実験の様子

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　1件／うちオープンアクセス　1件）

１．著者名 Yuki Kadokawa, Lingwei Zhu, Yoshihisa Tsurumine, Takamitsu Matsubara	４．巻 165
２．論文標題 Cyclic policy distillation: Sample-efficient sim-to-real reinforcement learning with domain randomization	５．発行年 2023年
３．雑誌名 Robotics and Autonomous Systems	６．最初と最後の頁 104425
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.robot.2023.104425	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Lingwei Zhu and Takamitsu Matsubara	４．巻 112
２．論文標題 Cautious policy programming: exploiting KL regularization for monotonic policy improvement in reinforcement learning	５．発行年 2023年
３．雑誌名 Machine Learning	６．最初と最後の頁 4527-4562
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10994-023-06368-z	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Tsurumine Yoshihisa、Matsubara Takamitsu	４．巻 158
２．論文標題 Goal-aware generative adversarial imitation learning from imperfect demonstration for robotic cloth manipulation	５．発行年 2022年
３．雑誌名 Robotics and Autonomous Systems	６．最初と最後の頁 104264～104264
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.robot.2022.104264	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Lingwei Zhu, Toshinori Kitamura, Takamitsu Matsubara
２．発表標題 Cautious Actor-Critic
３．学会等名 The 13th Asian Conference on Machine Learning (ACML)（国際学会）
４．発表年 2021年

１．発表者名 Toshinori Kitamura, Lingwei Zhu, Takamitsu Matsubara
２．発表標題 Geometric Value Iteration: Dynamic Error-Aware KL Regularization for Reinforcement Learning
３．学会等名 The 13th Asian Conference on Machine Learning (ACML)（国際学会）
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関