公開

研 究 終 了 報 告 書

「 頑健性と安全性の性能限界を明らかにする深層強化学習 」

研究期間:2020 年 12 月～2024 年 3 月
研 究 者:小林 泰介

1. 研究のねらい

深層強化学習の学習アルゴリズム自体は成熟してきており，単一の目的であれば安定して達成可能になってきている．そのため，付加価値，例えばサンプル効率などの追求が進められている．本研究では，信頼性に関与する頑健性（外乱への補償能力）と安全性（制約条件を満たす能力）の2つに注目する．これら単体では，頑健性ならば，システムを攻撃する外乱にも耐えうる制御器を獲得することで，安全性ならば，制御器から生成された行動が制約条件を破るようなケースを回避することで，それぞれ独立に実現されてきた．しかし，これら2要素はどちらも制御器の信頼性には必要にも関わらず，実際には双方を同時に実現させる試みは為されていない．これは，頑健性と安全性を獲得する技術を闇雲に組み合わせてしまうと，互いに干渉し合ってしまうことで制御性能の劣化を引き起こし本来達成すべき目的自体を失敗する恐れすら出てくるからである．

そこで本研究では，頑健性と安全性を両立可能な制御器を獲得する全く新しい深層強化学習の枠組みを要素技術から研究開発する．また，2つの付加価値の水準を定量評価することで，提案する制御器に安心して任せられる問題集合の明確化を目指す．

2. 研究成果

(1)概要

モデルベース強化学習を基礎とし，大きく 3 つの観点から計 6 つの要素技術を開発した．

1. 確率的予測モデルの表現力向上

1.1. 予測の最悪ケースを考慮可能な最適化問題を導出し，そこに潜むバイアス・バリアンスのトレードオフを高効率に調整可能なメタ最適化手法を開発した．

1.2. 確率分布が持つ表現能力を効率的に高めつつも，分布の平均的挙動を解析的に算出可能な制約付き確率分布モデルを開発した．

2. 実時間内での安全性保障

2.1. 高次元空間での低効率な予測を避けるため，制御可能な限界まで低次元空間に圧縮するためのスパース表現学習を開発した．

2.2. 高速に制約条件を破綻させない準最適解を獲得するため，行動系列候補から失敗を積極的に回避する更新則を持つモデル予測制御を開発した．

3. 頑健性の制約付き最大化

3.1. 不等式制約付きの最適化問題を適切に扱うため，学習可能なスラック変数を活用した不等式制約の正則化項への変換手法を開発した．

3.2. 予測モデルの頑健性を最大化するための敵対的学習において，予測精度の低下率に不等式制約を設ける新たな枠組みを開発した．

(2)詳細

1.1. 予測モデル学習における最悪ケースの考慮・メタ最適化

一般的な予測モデルの学習では，データセットを生成した分布に関する予測誤差の期待値を最小化する．これは平均的な挙動を示す上では適切だが，安全性・頑健性のどちらの観点においても最悪ケースの考慮も重要となるため，この方針は不適である．

そこで本研究では，平均・最悪ケースのどちらも考慮する新しい予測モデルの学習則として，データの1つ1つを目的とみなした多目的最適化問題に対する拡大チェビシェフ法を駆使して見出した．結果的に，予測誤差の期待値の最小化とデータ内で予測誤差が最悪となるものの最小化を重み付けして考慮可能となった．

ここで，予測誤差の期待値はバイアス，最悪ケースはバリアンスに対応するため，これら2つの最小化問題はバイアス・バリアンスのトレードオフが大きく関与する．そこで，これらの重みをバイアス・バリアンスが適度なバランスになるように自動調整するメタ最適化手法を検討した．具体的には，予測モデルは学習時には1ステップ分の予測精度を評価するものの，実用上は複数ステップの予測が求められるため，メタ目的を複数ステップの予測精度と定めた．その上で，重みを確率変数とみなし，その生成分布を方策勾配法に当てはめて最適化する枠組みを開発した．これは，生成分布の平均的挙動と乱択された重みの間で生じるメタ目的の差を比較することで，メタ目的が改善するよう生成分布を更新するものである．

数値シミュレーションにより，予測問題の部分観測性や長期予測性に合わせて最悪ケースを優先するよう重みが自動調整され，複数ステップに渡る予測精度の安定的な向上を確認した．なお，基礎検証の後に，実ロボットでの実証実験を実施したところ，獲得した予測モデルを駆使した制御性能において，従来の平均的挙動のみを考慮した予測モデルよりも，提案する最悪ケースまで考慮した予測モデルのほうがタスクの失敗状況を回避できることを確認した．

1.2. 平均的挙動を解析可能な確率分布モデルの表現力向上

予測モデルは確率分布でモデル化して学習されるが，利用する確率分布モデルの表現力次第では表せない分布特性（予測の多峰性や非対称な歪みなど）があり，予測精度の低下に繋がる．混合分布は任意表現が可能とされるが計算コストが大きく，近年注目されるNormalizing Flowは効率的なものの解析的に分布特性を得ることが困難となる．

確率分布の平均的挙動は予測の効率化や再現性の向上に有効であるため，本研究ではNormalizing Flowでも分布の平均を解析的に得られる設計手法を導いた．具体的には，確率分布が偶関数に限定されるようNormalizing Flowによる変換を奇関数で制約することで，変換前の確率分布と共通の平均を得ることが可能となった．

検証として，予測モデルより直接的に制御性能に寄与する行動方策を本技術でモデル化し，モデルフリー強化学習でベンチマークシミュレーションを解いたところ，定番とされるガウス分布よりも効率的かつ汎用的な学習に成功した．また，同程度の性能を発揮した混合分布より低計算コストであり，実ロボットの制御周期に十分に収まることを確認した．実際，実ロボットに問題なく適用可能であり，ガウス分布より高い制御性能を獲得できた．また，本技術を用いることで，条件付き確率分布と周辺化確率分布を設計的に繋げた学習を達成した．

2.1. スパース表現学習による制御に必要十分な低次元空間の抽出

予測モデルを制御に活用する汎用的な手法としては，複数の行動系列候補から予測される将来の状態・収益を比較することで，より良い行動系列候補を選定していき最適解を見出す，サンプリングベースのモデル予測制御が挙げられる．これは汎用的な反面，非常に多くの計算コストがかかってしまい実時間でのロボット制御の障害となる．計算コストが嵩む一要因として，予測モデルが扱う状態の次元数がある．予測精度を十分に維持しつつも，必要最低限の次元まで状態空間を圧縮・抽出することで，高速な制御が可能となる．

そこで本研究では，高次元データに潜む低次元潜在変数を抽出する技術である変分オートエンコーダを発展させ，さらなる低次元化を達成した．具体的には，その定式化をツァリス統計に従って始めることで，予測精度を高める項と潜在変数をスパース化する項のバランスが自動的に調整される新たな最適化問題を導出した．また，このバランスの自動調整が機能するためのパラメータ条件を解析的に明らかにした．

ロボットマニピュレータによる実証実験において，観測画像および手先位置・速度を本技術で圧縮したところ，制御に必要最小限である 6 次元への明確なスパース化に成功した．これにより，モデル予測制御の計算速度が倍以上に向上するとともに，十分に収束した解を得ることで制御性能自体の向上も果たした．

2.2. 制約条件を破綻させる行動を積極的に回避する安全なモデル予測制御

予測モデルを用いるモデル予測制御の計算コストが大きくなるのは，繰り返し最適化問題を解くからである．つまり，十分数最適化を繰り返していない段階で計算時間制限の都合で行動を確定してしまうと，それは最適なものとは遠く安全性に欠ける．不十分な繰り返し数でも安全な行動を得るには，制約条件を破綻させる行動は早期に回避しつつ収束を早めればよい．

そこで本研究では，モデル予測制御が扱う最適化問題が順方向カルバックライブラ情報量の最小化問題であることで，大域解を得ようと安全性の低い行動を取り続けてしまう挙動となる点に着目して，新たに逆方向カルバックライブラ情報量の最小化問題としてモデル予測制御を再導出した．その結果，ソルバーとして鏡像降下法を駆使すれば，従来と類似した行動系列候補の生成分布の更新則を得ながらも，新たに悪い予測結果をもたらす候補を避ける機能を加えることに成功した．ただし，良い予測結果を生成するよう強化する機能と干渉し得るため，各機能に対応する分布を個別に更新しつつ，それらを棄却サンプリングに基づいて合成する手法を開発した．

もう一点の課題である収束速度の向上について，ソルバーとして用いた鏡像降下法に Nesterov の加速法を導入する最新研究を参考にして，本問題に見合うように実装した．特に，モデル予測制御では鏡像が動的鏡像となる点を踏まえた近似と，サンプル依存の疑似勾配が引き起こす更新の不安定さを考慮した適応的な加速度の調整を行った．

提案手法を GPU による十分な並列計算が可能な条件と CPU のみに計算資源が限定される条件で検証したところ，提案手法だけが複数タスクに対して安定して実時間制御に成功した．また，実機検証ではロボットに搭載された CPU のみで 50Hz での実時間インピーダンス制御を可能にし，人とロボットのインタラクションをより滑らかなものにできた．

3.1. 学習可能なスラック変数を活用した不等式制約付き最適化

機械学習の最適化問題において，例えば外乱の強さ制限といった不等式制約を扱うには，従来手法ではラグランジアンを活用する．すなわち，制約式を正則化項へと変換して主問題と合わせて最小化しつつ，正則化項への係数を双対問題に基づき自動調整する．しかし，この方針を実装する際に，不等式制約を等式制約とみなしたラグランジュの未定乗数法と一致してしまっており，適切に不等式制約を扱えていないのが現状である．

そこで本研究では，先んじて不等式制約を等式制約へと変換しておくことで，この実装上の問題を回避する．具体的には，不等式制約における両辺の差を補償するスラック変数を陽に導入した．さらに，このスラック変数を等式制約が概ね満たされるよう与えるべく，制約に関わる入力に依存して学習可能な形で関数近似した．この学習には，等式制約を一定水準まで達成させつつ，一定水準以降はスラック変数を除外するよう振る舞う損失関数を設計した．

モデルフリー強化学習で代表的な Soft Actor-Critic においても，方策エントロピーに関する不等式制約が扱われていたため，それを提案手法で取り扱うように修正した．その結果，数値シミュレーション・実機実験ともに，方策エントロピーを適切に不等式制約の範囲で最大化することに成功した．また，最大化された方策エントロピーの恩恵として，学習効率の向上や頑健性の向上を確認できた．特に，実機実験ではロボットアームが学習時に経験しなかった未知物体を把持した場合や人とのインタラクション時にさえ，適応的に振る舞い制御性能を維持できた．

3.2. 予測精度の低下率を制限した敵対的学習による頑健性の最大化

予測モデルの頑健性を最大化させるには，学習可能な外乱生成器との敵対的学習問題を扱えば良い．ただし，ここで外乱生成器の性能を過剰にしてしまうと，安全性を保障するためにタスク自体が失敗するほど予測が保守的になり得る．この問題を回避するには，外乱生成器による予測精度の低下率に関する不等式制約を加味した敵対的学習により，Light Robustness の最大化を図る必要がある．

そこで本研究では，不等式制約を扱える敵対的学習構造を新たに設計した．具体的には，外乱を知覚しない予測モデルから，外乱で条件付けられた制約付き Normalizing Flow で変換した外乱を知覚可能な予測モデルを設けた．その上で，敵対的学習問題を予測モデルの対数尤度に対する変分下界から導出し，そこに外乱の知覚の有無で生じる予測精度の低下率に関する不等式制約を与えた．この問題を，成果 3.1 を活用して解く．なお，学習の効率化のために収集したデータは可能な限り再利用するが，その場合は外乱生成器への勾配算出用の計算グラフを保持できないため，過去の外乱が現在の外乱生成器から生成されたものとして計算グラフを後付けする手法を合わせて開発した．また，攻撃側にとっての最悪ケースは予測モデルにとっての最良ケースとなるため，新たに最良・最悪ケースの中心に相当する Midrange の最小化を期待値の最小化とバランス良く行う手法を導入した．

数値シミュレーションの結果，提案手法により外乱生成器は適度に調整されて，予測精度の低下率を不等式制約に収めることに成功した．これに伴い，過度な外乱による予測性能ひいては制御性能の低下を抑制しつつも，通常よりも頑健な予測モデルに仕上がることで，訓練した範囲の外乱を付与しても制御性能をある程度維持することに成功した．

3. 今後の展開

まず，直近の展開としては上記の開発した要素技術を全て統合した新たなモデルベース強化学習の枠組みを開発して，その性能限界を明らかにすることが挙げられる．その際，より実践的なロボットタスクでの概念実証も同時に進めることで，本研究の実用性をアピールしたい．

一方で，本研究は基本的に，事前に設計者・ユーザーにより適切に定められたロボットタスクを，安全性を最優先しながら頑健性を可能な限り付与するものであるが，実用的には，安全性を多少犠牲にしてでも任意の状況でも頑健に振る舞うロボットや，曖昧に指示されたタスクに柔軟に対応可能なロボットも求められる．特に後者は，急発展を遂げている大規模基盤モデルと融合することで，曖昧な言語指示に対応する報酬関数を生成して強化学習に利活用可能になってきている．そこで今後は，開発技術を基盤モデルと融合させることで，言語指示に応じて汎用的かつ頑健に振る舞う安全なロボット制御技術への発展が期待できる．

4. 自己評価

当初の想定を上回る技術課題を発見するとともに，各課題を一般化することで，多くの汎用的な要素技術を新開発できた．また，どの要素技術も最終的には実ロボットでの概念実証実験に成功しており，机上の空論に留まらない実用的な要素技術を開発できたと評価している．

一方で，当初予定していた安全性・頑健性の定量評価に関する研究項目に関しては，定性的には開発技術のパラメータとして調整可能にできたものの，厳密な定義には踏み込めなかった．これは必要な要素技術が当初の想定以上に不足しており自ら開発しなければならなかったことで時間を要してしまったのが主な理由のため，本研究終了後も継続して取り組んでいきたい．

5. 主な研究成果リスト

(1)代表的な論文(原著論文)発表

研究期間累積件数:16件

1. T. Aotani, T. Kobayashi, and K. Sugimoto: “Meta-Optimization of Bias-Variance Trade-off in Stochastic Model Learning,” IEEE Access, Vol. 9, pp. 148783-148799, 2021.
2. T. Kobayashi and T. Aotani: “Design of Restricted Normalizing Flow towards Arbitrary Stochastic Policy with Computational Efficiency,” Advanced Robotics, Vol. 37, No. 12, pp. 719-736, 2023.
3. T. Kobayashi and R. Watanuki: “Sparse Representation Learning with Modified q-VAE towards Minimal Realization of World Model,” Advanced Robotics, Vol. 37, No. 13, pp. 807-827, 2023.
4. T. Kobayashi and K. Fukumoto: “Real-time Sampling-based Model Predictive Control based on Reverse Kullback-Leibler Divergence and Its Adaptive Acceleration,” arXiv, 2022 (under review at Robotics and Autonomous Systems)
5. T. Kobayashi: “Soft Actor-Critic Algorithm with Truly-satisfied Inequality Constraint,” arXiv, 2023 (under review at Robotics and Autonomous Systems)

(2)特許出願

研究期間全出願件数:0 件(特許公開前のものも含む)

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

1. T. Aotani and T. Kobayashi: “Cooperative Transport by Manipulators with Uncertainty-Aware Model-Based Reinforcement Learning,” IEEE/SICE International Symposium on System Integration, pp. 959-964, 2024
2. T. Kobayashi: “LiRA: Light-Robust Adversary for Model-based Reinforcement Learning,” ICRA 2024 Workshop Back to the Future: Robot Learning Going Probabilistic, 2024 (accepted for presentation)
3. 第 69 回自律分散システム部会研究会「若手研究者による模倣学習・強化学習の新展開」で招待講演
4. 日本機械学会より日本機械学会奨励賞(研究)を受賞
5. 科学情報出版より“詳解 強化学習の発展と応用 ロボット制御・ゲーム開発のための実践的理論”を出版