論 文 の 内 容 の 要 旨

論文題目　Efficient Reinforcement Learning Using Simulation for Robot Control: Application to Logistics Cart Transportation System (シミュレーションを活用した実ロボット制御のための効率的な強化学習: 台車搬送システムへの適用)

氏　　名　松尾　凌輔

近年，ロボットの社会実装が進み，より適応的なロボット制御技術が求められるようになってきている．強化学習は，適応的な制御という観点から，将来有望なロボット制御技術の1つであると考えられるが，一方で，方策の学習とデータ収集を同時並行に行う必要があるために学習が複雑になり，多量の施行によるデータが必要，という課題がある．また，ロボット制御の文脈で考えると，さらなる課題が発生する．例えば，実世界でデータを収集しようとすると，制御の実行や環境のリセットに大きな時間が必要となり，さらなる学習コストがかかること，学習初期の不安定な挙動により，ロボットやその周囲に危険が及ぶ可能性があり，安全管理のためのコストが高いこと，などである．そういった中で，近年，物理シミュレータを利用することで学習効率の改善やリスクの低減を目指す研究が行われている．しかし，物理シミュレータを利用することによって発生する課題もある．具体的には，シミュレーション環境と実環境の物理的な振る舞いの差異によって，シミュレーション環境で学習した強化学習制御則が実環境でうまく機能しない可能性がある．

本論文の目標は，強化学習の学習コスト低減によって，実ロボットシステムに強化学習を適用することを容易にすることである．そのために，物理シミュレータを活用して効率的な強化学習手法を提案する．また，物理シミュレータを利用することによって発生する実環境での性能劣化を緩和するための手法を提案する．これらの提案手法を検証するための実際的なロボットシステムとして，協調台車搬送システムを採用している．

本論文は1章から5章までからなる．1章では，強化学習をロボットに適用する動機，ニューラルネットワークを利用した強化学習の導入，物理シミュレータの発展について

紹介する.

2章では，物理シミュレータ環境で，制御知識事前知識として利用した残差強化学習や分散並列での学習によって学習効率を改善する方法を提案する.制御知識を利用することで学習効率が向上でき，また，安定した制御則の獲得が期待できる．さらに，物理シミュレータを利用することでデータ収集コストを下げることができる．実験として，2台のロボットによる台車の協調運搬制御の学習を行い，効率的に学習できることを確認している．また，物理シミュレーション環境下で学習した強化学習制御則によって，実環境でのロボット制御を実現可能であることを示している.

3章では，実環境でのロボット制御で発生する遅延に着目して，遅延していない情報を効率的に活用して学習効率及び制御則の性能を改善する手法を提案する．これまで，遅延した観測から制御則を学習する方法として，遅延観測+遅延ステップ分の行動履歴の拡張観測を利用する方法や，遅延していない観測を推定し，その情報を利用して強化学習問題を解く方法などが研究されている．提案手法では，Actor Critic強化学習において，訓練時には遅延していない観測情報を状態行動価値関数の学習に利用できることに着目し，効率的な学習を実現している．実験として，いくつかのシミュレーションタスク及び2章と同様の協調台車搬送タスクに提案手法を適用し，学習を効率的に進められること，優れた性能の制御則を安定して得られることを実験的に示している.

4章では，実環境に制御則を転移した際の性能劣化を抑えるための方法を提案する．転移を実現するための方法として，今まで主に2つの手法Domain adaptationとDomain randomizationが提案されている．Domain adaptationでは，シミュレーションと実環境の両方を表せるような表現を獲得する手法であるが，実環境のデータが必要であるという課題がある．また，Domain randomizationはシミュレーション環境下で画像情報や環境変数（摩擦や重量など）をランダムにサンプリングして，その環境で学習することによって，実環境への転移に耐えうる頑強な制御則を獲得しようとするものである．しかしながら，ランダム化することによって問題が複雑になり，そもそも実用に耐えうる制御則を獲得できない可能性がある．本章では，この2手法の融合領域に注目し，Domain randomizationによってランダム化された環境変数を表現する状態表現を時系列データから学習する表現学習とその状態表現を利用して制御則を獲得する強化学習の2ステージの学習アルゴリズムを提案する．また，表現学習においては，状態表現が次の状態を推定できるように学習させることによって，マルコフ決定過程に漸近させ，強化学習にとって望ましい表現を学習させる．いくつかのシミュレーション実験を通して，提案手法が学習効率及び制御則の性能を改善することを示している.

5章では，本論文の貢献やこれまでの研究で残された課題について述べ，今後の展望を示している.