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　３ 年　 ６ 月 ２２ 日現在

機関番号：１２１０２
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ１１４２４

研究課題名（和文）一次遅れ要素やむだ時間要素でモデル化可能なアクチュエータに対応する強化学習法

研究課題名（英文）Reinforcement learning method for environment with actuators that can be modeled with first-order lag elements or dead time elements

研究代表者

澁谷　長史（Shibuya, Takeshi）

筑波大学・システム情報系・助教

研究者番号：９０５８２７７６

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：つぎの３通りの方法で補償器設計に関する研究を進め、さらに応用に関する研究も行った。１つ目の方法は、一次遅れ要素やむだ時間の有無によって生じる遷移先の状態の差を小さくするという基準で補償器を設計するというものである。２つ目の方法は、一次遅れ要素の有無によって生じる一次遅れ要素の出力の差を小さくするという基準で補償器を設計するというものである。３つ目の方法は、一次遅れ要素に対する拡張状態を一次遅れの特性を利用した低次元表現によって設計するものである。最後に、アクチュエータを持つロボットが様々な床面を走行する場合を想定し、行動戦略を切り替える強化学習法について研究を行った。

研究成果の学術的意義や社会的意義
本研究の成果は大きく２つの学術的意義を有する。本研究の意義の1つ目は、補償器をあとから追加する方式をとる場合でもそれらの再学習を不要にできる点である。一次遅れ要素やむだ時間要素を含まない環境で学習を行い、あとからこれらを追加した環境で学習しようとする場合に生じる再学習を避けることができる。また、2つ目は、一次遅れ要素やむだ時間要素の出力値に関する情報を直接的には利用しないため、環境の情報を新たにセンシングする必要もない点である。この性質により、環境から見込んだ先を不変のものとして扱うことができる。

研究成果の概要（英文）：In this study, the compensator was designed by the following three methods. The first method is to design the compensator to reduce the difference in the successive state caused by the presence or absence of the first-order lag element and the dead time. The second method is to design the compensator to reduce the difference in the output of the first-order lag element caused by the presence or absence of the first-order lag element. The third method is to design the extended state for the first-order lag element by a low-dimensional representation using the characteristics of the first-order lag. Numerical simulations using a two-link manipulator or an inverted pendulum were performed to confirm its effectiveness. Lastly, we studied reinforcement learning method which switches control strategy adaptively for environment conditions.

研究分野：機械学習

キーワード：機械学習　強化学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

強化学習は、多段階の意思決定問題(行動決定問題) を解くことでシステムを最適に運用する機械学習の有力な枠組みである。具体的には、エージェント(制御主体) と環境(制御対象) とが、状態・行動・報酬といった情報を通じて離散時間ごとに相互作用する系を対象とし、試行錯誤によって望ましい行動を決定するエージェントを構築する。近年では、Google DeepMind 社による囲碁プログラム AlphaGo[1] やビデオゲームの自動操縦[2]などの、計算機のなかで理想化された題材に対して、強化学習の枠組みが採用され、成功を収めてきた。

政府がとりまとめた人工知能技術戦略には、人工知能技術を活用した知能ロボティクスの実現が掲げられている。強化学習はこのようなロボットの知的な振る舞いを実現するための中核的な技術となる。しかし、実機への適用には、現実的なアクチュエータに起因する大きな課題がある。

実機を扱う際には、エージェントは計算機、環境はロボットの身体やロボットの存在する外界となる。そして、状態遷移には、アクチュエータのダイナミクスが反映される。ここでいうアクチュエータとは、エージェントが決定した行動を、状態遷移に対して作用させる要素である。実機では、モータや人工筋肉などの現実的なアクチュエータを用いることになる。これらの例から分かるように、現実的なアクチュエータは、遅延が発生し、本来的に指令値をすぐ達成することができない。制御工学では、このような遅延を発生する現実的なアクチュエータを、一次遅れ要素とむだ時間要素によって近似的にモデル化できることが知られている。一階の微分方程式で記述できる遅延を表す要素は一次遅れ要素、時間推移で記述できる遅延を表す要素はむだ時間とよばれる。

申請者らは、アクチュエータの一次遅れ要素の時定数やむだ時間の大きさが制御周期に対して大きくなるという条件で、強化学習による制御性能が著しく低下するということを突き止めた。この予備実験は、2-リンクマニュピレータを対象とし、強化学習における標準的な評価基準(エージェントが獲得した報酬からなる基準) を用いて数値実験により実施したものである。

申請者らは、一次遅れ要素やむだ時間要素が存在する環境において標準的な強化学習のアルゴリズムが有効に機能しなくなるのは、マルコフ性の欠如に原因があると考察している。強化学習のアルゴリズムは、環境がマルコフ性を持つことを前提としている。すなわち、次なる状態は、現在の状態と、エージェントが現在選択した行動により決定される、という前提である。ところが、一次遅れ要素やむだ時間要素が存在すると、エージェントの状態遷移に過去の行動が影響を与えるようになってくる。したがって、強化学習の前提条件がくずれてしまい、学習が成立しなくなってしまう。たしかに、強化学習による実機の制御での成功例もある。これらは静特性のみを扱っていたか、動特性を対象としていても時定数と制御周期の兼ね合いがほぼ問題ない範囲に収まっていたなどして、マルコフ性が近似的に成立していたものと推察される。

マルコフ性が崩れる場合に対応する一般的な対処法として、未知の環境に適用可能な拡張法[3]がある。拡張法は、過去の状態の履歴もあわせて拡張状態とすることでマルコフ性の回復を図る、原理的にもっとも有効な手法である。しかし現実的には、単純に過去の履歴を利用するだけでは、状態数が指数的に増加してしまい、学習を困難にするという問題が発生する。特に一次遅れは、無限の過去まで遡る必要があるため、特にこの手法では十分な近似を得ようとするだけでも現実的ではない。このように、環境の挙動がブラックボックスであるという前提にたちながらマルコフ性の回復を図る従来手法の適用では、一次遅れ要素やむだ時間要素でモデル化可能なアクチュエータを含む対象を、理論的には扱えても、実際に対象を制御することはできない。

そこで、単純に拡張法を適用すると学習が困難になってしまうという問題に対して、申請者は、アクチュエータ部分についてのみ予め特性を同定しておき、その結果を事前知識として組み込むことで、完全ではなくともアクチュエータの影響をなるべく打ち消し、マルコフ性の欠如を低減する補償器を用いるアプローチが有効ではないかと考えた。

２．研究の目的

本申請課題では、一次遅れ要素やむだ時間要素でモデル化可能なアクチュエータが存在する環境で有効に機能する強化学習法を明らかにすることを目的として、研究を実施した。

３．研究の方法

つぎの３通りの方法で補償器設計に関する研究を進め、さらに応用に関する研究も行った。

(1) 補償器設計に関する１つ目の方法として、一次遅れ要素やむだ時間の有無によって生じる遷移先の状態の差を小さくするという基準での補償器設計方式について研究を行った。

(2) 補償器設計に関する２つ目の方法は、一次遅れ要素の有無によって生じる一次遅れ要素の出力の差を小さくするという基準での補償器設計方式について研究を行った。
(3) 補償器設計に関する３つ目の方法は、従来法では理論的には無限の過去の行動までさかのぼって考慮する必要のあった一次遅れ要素に対する拡張状態を一次遅れの特性を利用した低次元表現による補償器設計方式について研究を行った
(4) アクチュエータを持つロボットが様々な床面を走行する場合を想定し、行動戦略を切り替える強化学習法について研究を行った。

４．研究成果
本研究における具体的な成果は以下の通りである。

(1) 行動にむだ時間要素または一次遅れ要素を持つ環境において強化学習の性能を下げる原因はマルコフ性が欠如することにあった。そこで、完全ではないもののその欠如の程度を低減するために計算コストを抑えつつ得られる報酬を従来手法よりも向上させるような強化学習手法として、状態と行動の履歴を用いた遅れによる状態遷移先の差を小さくする補償器を強化学習器の行動出力に追加する手法を提案した。２リンクマニピュレータを題材とした数値実験において、本手法が行動にむだ時間要素または一次遅れ要素を持つ環境において、時定数が一定の範囲内であれば学習性能を向上させることが可能であることを確認した。

(2) 一次遅れ要素の有無によって生じる一次遅れ要素の出力の差を小さくするという基準で補償器を設計する手法を提案した。行動に一次遅れ要素を持つ２リンクマニピュレータを題材とした数値実験を行った。補償器を行わない方法、申請者らが以前提案した補償器を用いる方法、本手法の３つで比較したところ、制御周期に対する時定数の比が 1.5 程度までの範囲で、本手法の性能がもっとも高くなったことを確認した。

(3) 存在する内部変数を事前に同定した時定数をもとに逐次的に推定することにより、拡張状態の次元の拡張数を 1 アクチュエータあたり 1 次元に抑える手法を提案した。加えて、この方式で構成された補償器が、従来の拡張法によって無限の行動履歴を用いて状態の拡張を行った場合と等価な制御性能を持つことも示した。これにより、従来の拡張法が抱えていた状態数の指数的増加に伴う計算爆発を回避することができる。倒立振り子を題材とした数値実験において、提案手法が一次遅れを考慮していない手法や Augmented MDP よりもより良い制御則を獲得でき、計算コストが小さいことが分かった。

(4) 一般には一次遅れ要素やむだ時間要素のパラメータは、アクチュエータ以外の要素の影響を受けて変化すると考えられる。そこで、環境のパラメータ変動を検知して制御戦略を適応的に使い分ける強化学習法を提案した。複数の床面を走行するロボットを題材にしてその有効性を確認した。

参考文献
[1] David Silver, “Mastering the game of Go with deep neural networks and tree search," Nature 529, pp.484-489, 2016.
[2] Volodymyr Mnih et.al, “Human-level control through deep reinforcement learning," Nature 518, pp.529{533, 2015.
[3] Thomas J. Walsh et.al, “Planning and learning in environments with delayed feedback,", Vol.18, pp.83-105, 2009.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Teppei Iwata and Takeshi Shibuya	４．巻 -
２．論文標題 Adaptive Modular Reinforcement Learning for Robot Controlled in Multiple Environments (in press)	５．発行年 2021年
３．雑誌名 IEEE Access	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2021.3070704	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 －

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 小林翔樹、澁谷長史
２．発表標題 行動出力に大きな一時遅れを持つ環境における強化学習のための補償器の設計
３．学会等名 第76回知的システム研究会（SIC2019-2）論文集
４．発表年 2019年

１．発表者名 Masaki Yotsukura, Takeshi Shibuya
２．発表標題 Reinforcement Learning Method for Cases Where the State Observation Period Is Larger Than the Action Decision Period
３．学会等名 Proceedings of the SICE Annual Conference 2018（国際学会）
４．発表年 2018年

１．発表者名 四ツ倉 昌輝, 澁谷 長史
２．発表標題 行動出力に一次遅れ要素がかかる環境のための拡張状態を用いた強化学習法
３．学会等名 第78回知的システム研究会（SIC2020-2）論文集
４．発表年 2020年

１．発表者名 Shoki Kobayashi and Takeshi Shibuya
２．発表標題 Reinforcement Learning Compensator Robust to the Constants of First Order Delay Elements
３．学会等名 Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics（国際学会）
４．発表年 2020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