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　 ７ 日現在

機関番号：１２６０５
研究種目：若手研究(B)
研究期間：2017～2018
課題番号：１７Ｋ１２７３７

研究課題名（和文）オンライン鏡像降下法に基づく高次元強化学習アルゴリズムの構築と応用

研究課題名（英文）Mirror Descent approach for the high dimensional reinforcement learning algorithm

研究代表者
矢野　史朗（Yano, Shiro）

東京農工大学・工学（系）研究科（研究院）・助教

研究者番号：９０６３６７８９

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）：「鏡像降下法を基盤にした強化学習アルゴリズムの設計」「鏡像降下法とベイズ学習の関係理解」「強化学習アルゴリズムの応用」という構成で研究を進めた．
アルゴリズム設計では，鏡像降下法を基盤にderivative freeアルゴリズムを設計した．さらに，鏡像降下法の拡張を行った上で同様の手続きを行うという手順により，強化学習アルゴリズムの拡張が可能であることを示した．鏡像降下法からベイズの定理が導出できることに着目し，統計的推定アルゴリズムと強化学習アルゴリズムの接点について研究を行った．設計した強化学習アルゴリズムの応用事例研究として，多自由度の強化学習問題と，ロボットアーム制御問題を扱った．

研究成果の学術的意義や社会的意義
相手の価値観や競技の採点基準（目的関数）を満たすよう行動を最適化する必要があるとき，初対面の相手や初めての競技で，この目的関数を事前に把握することは困難である．本課題で扱うのは，こうした扱う問題のモデルを持たない状況で現場に臨み行動（方策関数）を最適化していく問題であり，未知環境下で活動する人工物にとって重要な問題である．
より実用的には行動空間も状態空間も高次元かつ連続という状況を考える必要があり，本課題ではこうした高次元な強化学習問題のためのアルゴリズム設計と，いくつかの応用事例を示すものである．

研究成果の概要（英文）：In summary, this project tried the three issues. 1. To provide and extend a direct policy search method on the basis of the Mirror descent method. 2. To study the relationship between the mirror descent method and Bayes' theorem. 3. To apply the proposed reinforcement algorithms for the tasks including locomotion simulation, deep reinforcement learning tasks and robotic arm control.
The project proposed "mirror descent search". Then, accelerated mirror descent method was applied onto the proposed one. The project studied the Bayesian inference algorithms from the viewpoint of mirror descent method. The project was evaluated by the tasks such that
1. Convolutional Neural network training (~5e8 dimensional problem) 2. Locomotion learning in the physics engine 3. Robotic arm control problem in the real world

研究分野：知能情報学

キーワード：強化学習　鏡像降下法　ベイズ推定　Nesterov加速　直接方策探索

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

対人サービスやスポーツ競技などでは，相手の価値観や競技の採点基準（目的関数）やルール（制約条件）を満たすよう行動を最適化する必要がある．しかしこの目的関数や制約条件を事前に把握することは，相手が初対面であったり初めての競技であったりする場合，難しい問題である．人工的なシステムにこうしたタスクを与える場合，まさに目的関数や制約条件が未知であることが障壁となる．このように目的関数のモデルを持たない状況で現場に臨み，行動（方策関数）を最適化しなければならないという問題を，強化学習問題と呼ぶ．

対人サービスを始め様々な産業で重要となるシステムには，例えば映像入力に基づく多リンクロボット制御システムなどが考えられる．こうしたシステムは，視覚情報のように高次元ベクトルとして観測された入力状態を，高自由度アクチュエータの制御信号である高次元ベクトルへ変換し出力しなければならない．このように，行動空間も状態空間も高次元かつ連続という実用的な状況で，さらに目的関数など環境に関するモデルも事前に準備できていない，という強化学習問題への解法に注目が集まっている．

本研究で強化学習アルゴリズムの設計に用いる鏡像降下法(mirror descent)は，最急降下法や指数勾配法といった幾つかの逐次最適化アルゴリズムを，Bregman 擬距離を用いて一般化した勾配法である．また観測値を得るごとに鏡像降下法を１ステップ進める手法をオンライン鏡像降下法（以降，単に鏡像降下法）と呼ぶ．鏡像降下法の Bregman 擬距離を KL 擬距離あるいはそれに類するものにすると，目的関数の微分を用いずに最適解を探索する derivative free アルゴリズムが導かれる．強化学習問題を derivative free アルゴリズムによって解くというアプローチは従来から試みられており，本研究ではこのアプローチを鏡像降下法という数学的技法に基づいて行うことで，強化学習アルゴリズム設計の体系的方法の確立を目指す．

また鏡像降下法からは，ベイズの定理を導出することもできる．強化学習アルゴリズムにはモデルフリーとモデルベースの２種類があり，後者には目的関数や環境のダイナミクスについてのモデルを学習し推定するといった，認知的な技能の獲得と利用が含まれる．モデルベース強化学習のような認知と運動の双方を必要とする手法にとって，鏡像降下法は重要な橋渡しとなると期待される．

２．研究の目的

本研究は，方策探索の主要な 2 つの流れである，方策勾配定理・ 確率推論に基づく手法を，オンライン鏡像降下法から統一的に導く新体系を提供する．これによって近年の鏡像降下法の成果の多くが，方策探索手法に適用可能となる．更にその最近の一例である加速鏡像降下法を応用して，収束速度と最適性の高い強化学習アルゴリズムを構築する．

これをロボットアーム制御問題と，多自由度の強化学習の学習アルゴリズムに適用し，従来手法と性能を比較評価する．

３．研究の方法

本研究は，この鏡像降下法を基盤にして，⑴理論１：強化学習アルゴリズムの設計，⑵理論２：鏡像降下法とベイズ学習の関係理解，⑶応用：強化学習アルゴリズムの深層強化学習課題またはロボットアーム制御への適用という構成で進める．

⑴　①鏡像降下法に基づいてアルゴリズムを設計することの利点は，鏡像降下法について提案されてきた様々な拡張手法を援用できる点にある．そこでまず，鏡像降下法を基盤に derivative free アルゴリズムを設計し，それを強化学習に適用することを目指す．さらに，鏡像降下法の拡張を行った上で同様の手続きを行うという手順により，強化学習アルゴリズムの拡張が可能であることを示す．

②鏡像降下法からベイズの定理が導出できることに着目し，鏡像降下法に基づいて設計したアルゴリズムによって統計的推定・推論問題が解けることを示す．また推定と強化学習問題の接点について研究を行う．

⑵　設計した強化学習アルゴリズムの応用事例研究として，多自由度の強化学習問題と，ロボットアーム制御問題に適用する．多自由度強化学習問題は， MuJoCo や OpenSim 上に構築された二足歩行物理シミュレータの歩行制御器学習を課題とする．ロボットアーム制御問題では，7 軸ロボットアーム(KUKA LBR iiwa 7 R800)を使用し，3 次元空間上でのエンドエフェクタ位置のフォースコンプライアンス制御課題を扱う．

４．研究成果

⑴の成果として，鏡像降下法に基づく直接方策探索アルゴリズムを設計した．また，最急降下法向けの加速手法（Nesterov 加速法）の Bregman 擬距離についての一般化として提案された加速鏡像降下法(W. Krichene, NIPS 2017)をこれに適用することで，直接方策探索の

加速が可能なことを示した（M. Miyashita et al., Robotics and Autonomous Systems, 2018他）. 鏡像降下法に基づいて強化学習問題を解くという道筋を示したことと, 鏡像降下法の拡張手法がこれに適用できることを示したことにより, 今後さらなる理論研究や拡張研究が進むと期待される. 関連する分析として, 鏡像降下法により設計されたアルゴリズムの収束速度についての分析を行った（Y. Murata, et al., ICT ISPC, 2018 他）. また, 同手法を適切に拡張することで, 最適化アルゴリズムである Particle Swarm Optimization が導出できることを示した（S. Yano, ASCC, 2019）.

⑵の成果として, 得られた derivative free アルゴリズムによって 7 軸協調ロボットアームの強化学習課題を扱い, 制御器の学習を行った（国際会議査読中）. また深層ニューラルネットの１つである畳み込みニューラルネットワークによる MNIST 画像分類の課題（7 層, 330 万パラメータ）の訓練が可能なことを示した. MuJoCo や OpenSim 上の二足歩行物理シミュレータの学習が可能なことを示した（投稿中）.

本研究で開発したアルゴリズムの一部は, Github でオープンソースとして公開しており, また手法の設計と評価が進むに従って, 徐々に公開を進めていく予定である (https://github.com/mmilk1231/MirrorDescentSearch).

５．主な発表論文等

〔雑誌論文〕（計　2　件）

1. M. Miyashita, S. Yano, T. Kondo, “Mirror descent search and its acceleration”, *Robotics and Autonomous Systems*, Vol. 106, pp.107-116, 2018（査読有）
2. 矢野史朗, 近藤敏之, 前田貴記, 「運動主体感に着目したリハビリへのモデルベースドアプローチ」日本ロボット学会誌, 35 巻 7 号, pp. 512-517, 2017（解説記事, 査読無）

〔学会発表〕（計　13　件）

1. S. Yano, “Tutorial on Blackbox Optimization Approach for Reinforcement Learning Problems”, The 12th ASian Control Conference (ASCC 2019, 口頭, 査読無).
2. S. Yano, “Statistical Learning formulation of Sense of Agency, From normal subjects to mental disordered subjects”, The 1st Korea-China-Japan international symposium on disability overcome, 2018 (Invited talk, 口頭, 査読無)
3. S. Yano, “Mirror Descent: Bridge Between Bayesian-brain and Reinforcement Learning Algorithm”, The 2018 Japan-America Frontiers of Engineering symposium (JAFOE 2018, invited poster, ポスター, 査読無)
4. 矢野史朗, 近藤敏之, 林叔克「運動主体感とベイズ学習モデルおよび検証用仮説の提案」ロボット学会 2018 （口頭, 査読無）
5. Y. Murata, S. Yano, T. Kondo, H. Imamizu, T. Maeda, “Estimation of the human learning algorithm under the time-delay adaptation task”, The Second International Symposium on Embodied-Brain Systems Science 2018（ポスター, 査読無）
6. R. Philipp, T. Oya, N. Uchida, T. Funato, M. Ishitsubo, S. Yano, T. Kondo, K. Tsuchiya, K. Seki, “Neural adaptation to uni- and cross tendon transfer in the macaque forearm: An EMG and ECoG study”, The Second International Symposium on Embodied-Brain Systems Science 2018（ポスター, 査読無）
7. M. Ishitsubo, S. Yano, R. Philipp, K. Seki, K. Tsuchiya, T. Kondo, “Slow dynamics extraction of Muscle synergy in Macaque monkey”, The Second International Symposium on Embodied-Brain Systems Science 2018（ポスター, 査読無）
8. Y. Murata, M. Miyashita, S. Yano, T. Kondo, “On the Residual of Mirror Descent Search and Scalability via Dimensionality Reduction”, The 7th ICT International Student Project Conference (ICT-ISPC2018), Thailand 2018（口頭およびポスター, 査読有）
9. M. Miyashita, R. Hirotani, S. Yano, T. Kondo, “Direct Policy Search with Extremum Seeking”, SICE Annual Conference 2017, pp. 1539-1540（口頭, 査読有）
10. M. Miyashita, R. Hirotani, S. Yano, T. Kondo, “Experiment of Reinforcement Learning with Extremum Seeking”, The 6th ICT International Student Project Conference (ICT-ISPC2017) Malaysia 2017, pp. 1-4（口頭およびポスター, 査読有）
11. S. Yano, “Accelerated Mirror Descent in Reinforcement Learning”, The 8th International Symposium on Adaptive Motion of Animals and Machines, 2017 (口頭, 査読無)
12. M. Miyashita, S. Yano, T. Kondo, “Mirror Descent based Reinforcement Learning”, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2017), Embodied Brain Systems Science Workshop, 2017（ポスター, 査読無）

13. S. Yano, H. Imamizu, T. Kondo, T. Maeda “Bayesian Learning and Sense of Agency”, IEEE/ RSJ International Conference on Intelligent Robots and Systems (IROS 2017), Embodied Brain Systems Science Workshop, 2017（ポスター，査読無）

〔図書〕（計　1　件）
1. 矢野史朗(分担執筆)“身体性システムとリハビリテーションの科学 2 身体認知”，編：近藤敏之，今水寛，森岡周，東京大学出版会，pp.103-141（第 4 章「身体意識の数理モデル」），2018

〔産業財産権〕
○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等
開発済みアルゴリズムを以下のレポジトリで公開している：
https://github.com/mmilk1231/MirrorDescentSearch

アウトリーチ活動の１つとして，強化学習に関するワークショップを The Asian Control Conference (ASCC2019)で開催し，本研究成果に関するチュートリアルを行った．
国内学会「日本ロボット学会 2018」で特別セッションをオーガナイズし，本課題の成果を含む講演を行った．
国内学会「計測自動制御学会 システム・情報部門 学術講演会 2018」で一般セッションをオーガナイズし，本課題の成果を含む講演を行った．

６．研究組織

(1)研究分担者
研究分担者氏名：
ローマ字氏名：
所属研究機関名：
部局名：
職名：
研究者番号（8 桁）：

(2)研究協力者
研究協力者氏名：
ローマ字氏名：