OUKA
Osaka University Knowledge Archive

Title	Safe Natural Policy Gradient Algorithms
Author(s)	岩城, 諒
Citation	大阪大学, 2019, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/72376
rights	
Note	

Osaka University Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

Osaka University

様式３

論　文　内　容　の　要　旨

氏　名　（　岩城　諒　）	
論文題名	Safe Natural Policy Gradient Algorithms （安全な自然方策勾配法）

論文内容の要旨

強化学習は機械学習手法の一種であり，方策と呼ばれる意思決定則を，未知環境と相互作用しながら最適化することを目的とする．意思決定主体であるエージェントは，意思決定の即時的な評価値として環境から与えられる報酬が長期的に最大化されるよう，方策を更新する．長期的な報酬の期待値は，価値関数と呼ばれる．強化学習は，ロボット制御や債権回収，囲碁などに適用され，成功を収めてきた．自然方策勾配法は強化学習手法の一種であり，自然な計量に基づく最適化手法である自然勾配法を利用して，方策のパラメータを最適化する．近年，方策勾配型強化学習手法の発展の多くが，自然方策勾配法に基づいている．一般の機械学習手法においては，自然勾配の推定にはFisher情報行列の逆行列を推定することが求められ，最適化したいパラメータの次元に対し二乗のオーダーの計算量とメモリが必要となる．一方，強化学習のための自然方策勾配は，方策パラメータの次元に対し線形のオーダーの計算量とメモリで推定できることが示されている．この手法において，TD誤差と呼ばれる価値関数の予測誤差を，方策の対数勾配を基底とする線形関数近似器で回帰することで，そのパラメータが自然方策勾配に収束する．しかし，そのような自然方策勾配の推定法は計算機的に不安定であり，学習率などメタパラメータの設定値に非常に敏感である．本論文では，自然方策勾配の逐次推定法を改良し，推定量が発散することのない安全な学習手法を構築することを目的として，2つの手法を提案した．

一つ目の手法として，自然方策勾配を逐次推定するための適応的学習率を提案した．まず，適応的学習率を提案するための準備として，自然方策勾配を逐次推定するための学習率が満たすべき上限を導出した．導出した上限を学習率が満たせば，線形近似器の近似誤差は局所的に必ず減少することが保証され，安定した学習が可能である．次に，無限小の学習率で無限回更新する極限として，自然方策勾配を推定するための適応的学習率を導出した．この適応的学習率は，導出した学習率の上限を常に満たすことが保証される．提案法を古典的な制御問題に適用し，従来法よりも安定した学習が可能であり，学習率などメタパラメータの設定値に対し頑健であることを示した．

二つ目の手法では，陰的な確率的勾配降下法を基にして，自然方策勾配を推定するための更新則そのものを拡張した．一つ目の手法では，学習率を制御することで自然方策勾配の更新量の大きさが適応的に変化するのに対し，二つ目の手法では，勾配に正定値行列を掛け合わせることで，更新量の大きさだけでなく勾配方向も適応的に変化する．提案法は，計算の順番に注意することでベクトルの内積計算・スカラ倍・足し合わせのみで実行できるため，その計算量とメモリは方策パラメータの次元に対し線形のオーダーに保たれる．提案法を理論解析することで，その安定性を示した．まず，Lyapunovの安定性原理に基づく理論解析により，いくつかの技術的な仮定のもとで，提案法は確率1で自然方策勾配を推定できることを示した．次に，推定される自然方策勾配のノルムの上限を解析的に求めた．この理論解析結果から，従来法は学習が成功してもその後に推定量が発散しうること，一方で提案法ではノルムの上限が有界に抑えられることが示された．一つ目の手法と同様に，提案法を古典的な制御問題に適用し，従来法よりも安定した学習が可能であり，学習率などメタパラメータの設定値に対し頑健であることが示された．さらに，従来法であれば推定量が発散する実験設定であっても，提案法は頑健に学習できることが示された．

様式 7

論文審査の結果の要旨及び担当者

氏　名．（　岩城　諒　）		
	(職)	氏　名
論文審査担当者	主 査　教授	浅田 稔
	副 査　教授	細田 耕（基礎工学研究科）
	副 査　教授	大須賀 公一
	副 査　教授	平田 勝弘
	副 査　教授	中谷 彰宏
	副 査　教授	南埜 宜俊

論文審査の結果の要旨

本論文は，強化学習において自然方策勾配を安全に逐次推定するための手法を提案し，既存の逐次推定法が抱える課題である計算機的な不安定性を解決できることを示した．提案手法は推定される自然方策勾配のノルムが有界に抑えられることを理論的に保証することで，自然方策勾配を安全に逐次推定できる．

本論文の第 1 章では，強化学習と自然方策勾配法の概要を示し，さらに安全な強化学習アルゴリズムを構築する試みに関する，既存の研究成果を概説した．第 2 章では，マルコフ決定過程を定式化し，その古典的かつ効率的な解法として方策反復法を概説した後，方策反復法の近似解法として強化学習を導入した．第 3 章では，自然方策勾配法について詳述し，その有用性について述べた．さらに，自然方策勾配の既存の逐次推定法を導出し，それらの利点について述べた後，本論文が対象とする課題である，逐次推定法の不安定さを示した．

第 4 章では，1 つ目の提案手法として，自然方策勾配を安全に逐次推定するための適応的学習率を示している．まず，線形近似器の近似誤差が局所的に必ず減少することを安全であるとして，安全な逐次推定のために学習率が満たすべき上限を導出した．次に，導出した学習率の上限を満たすことが保証されるような適応的学習率を提案した．この適応的学習率は，無限小の学習率で無限回更新する極限として導出された．計算機実験において，提案手法は既存手法と比較して安定した学習が可能であり，さらに設計者が調節すべきメタパラメータの設定値に対し頑健であることが示された．

第 5 章では，2 つ目の提案手法として，陰的な確率的勾配降下法を基にした，自然方策勾配の逐次推定法を提案している．2 通りの理論解析によって，提案手法の安定性が示された．まず，提案手法による自然方策勾配の推定値が既存手法と同じ不動点に収束することを，Lyapunov の安定性原理に基づいて示した．この収束の証明方法は非常に一般的であるが，実際での学習中には必ずしも成立しない技術的な仮定が必要である．2 つめの理論解析では，これらの仮定をおかず，推定される自然方策勾配のノルムの上界を解析的に求めた．この理論解析結果から，既存手法はマルコフ決定過程での学習において勾配の推定量が発散しうることと，提案手法では推定量の発散を回避できることが示された．ここでも計算機実験を行い，既存手法であれば推定量が発散する実験設定であっても，提案手法は安定した学習が可能であり，メタパラメータの設定値に対しても頑健であることが示された．

以上のように、本論文は自然方策勾配を逐次推定するための既存手法が不安定である理由を理論的に明らかにし，安全性が理論的に保証されるに逐次推定法を提案した．よって本論文は博士論文として価値あるものと認める．