論文 / 著書情報
Article / Book Information

題目(和文)	強化学習の統計的理論とそのロボット制御への応用
Title(English)	Statistical Theory of Reinforcement Learning with Applications to Robot Control
著者(和文)	TINGTING ZHAO
Author(English)	TINGTING ZHAO
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第9557号, 授与年月日:2014年3月26日, 学位の種別:課程博士, 審査員:杉山　将,佐藤　泰介,篠田　浩一,藤井　敦,瀬々　潤
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第9557号, Conferred date:2014/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	審査の要旨
Type(English)	Exam Summary

（博士課程）

論文審査の要旨及び審査員

報告番号	甲第　　号		学位申請者氏名		Tingting Zhao	
論文審査 審査員		氏　名	職　名		氏　名	職　名
	主査	杉山将	准教授	審査員	瀬々潤	准教授
	審査員	佐藤泰介	教授			
		篠田浩一	教授			
		藤井敦	准教授			

論文審査の要旨（2000 字程度）

本論文は「Statistical Theory of Reinforcement Learning with Applications to Robot Control」と題し，英文５章から成っている．

第１章「Introduction」では，研究の背景および本論文の全体構成を示している．まず，機械学習の主要な研究課題である教師付き学習，教師なし学習，および，強化学習の問題設定，更には，ロボット制御における強化学習の役割について説明している．特にロボット制御においては，高次元のシステムを扱う必要があり，また，大量のデータを集めるためには多大なコストがかかるため，少ないデータから制御則を効率良く学習する必要があることを述べている．

第２章「Related Works」では，強化学習の標準的な枠組みであるマルコフ決定過程の定式化を示した後，政策反復法と政策探索法という二つの主要なアプローチを紹介している．政策反復の枠組みでは，価値関数を通した政策の評価，および，価値関数を用いた政策の改善を繰り返し実行することによって，最適な政策を求める．この政策反復の枠組みでは，価値関数を通した政策の評価は直接実行することが難しいため，最小二乗法による関数近似を用いるのが一般的である．一方，政策探索法の例として，政策勾配法，自然政策勾配法，政策事前分布勾配法，EM 政策探索法を紹介している．政策探索法は，価値関数を経由せず直接政策を学習するため，ロボット制御のような高価なシステムの学習には適しているが，政策勾配法では政策の更新が不安定になる問題があると指摘している．そして，その問題を改善すべく近年提案された政策事前分布勾配法が，ロボット制御のための政策探索法として特に有望な方法であると述べている．

第３章「Analysis and Improvement of Policy Gradient Estimation」では，政策事前分布勾配法の理論的な性質を解明するとともに，精度を更に向上させるための改良法を与えている．具体的には，政策勾配法，及び，政策事前分布勾配法で政策の更新に用いる勾配推定量の分散を理論的に評価し，政策勾配法の勾配推定量の分散の上界は意思決定のステップ数に比例するのに対して，政策事前分布勾配法の勾配推定量の分散の上界は意思決定のステップ数に依らないことを証明している．この理論解析により，意思決定に長いステップ数を要するタスクにおいては政策事前分布勾配法の方が安定して政策の更新を行うことができると示唆され，実際に計算機実験により政策事前分布勾配法の安定性を実証している．また，政策事前分布勾配法は，ベースライン減算という技法を用いることにより，勾配推定量の分散を更に低減させられることを理論的に証明するとともに，その有効性を実験的に示している．

第４章「Efficient Sample Reuse in Policy Gradients with Parameter-based Exploration」では，多数のデータを収集することが困難な状況において，過去に別の政策に従って収集したデータを，現在の政策の学習に活用する標本再利用法を提案している．この標本再利用の枠組みでは，政策関数の差異に起因する確率分布の違いを重点サンプリングとよばれる技法により吸収する．一般に重点サンプリングを用いると推定量のバイアスを軽減できるが，一方で分散が増加してしまうことが知られている．本章では，ベースライン減算を重点サンプリングと組み合わせることにより，分散の増加を抑制できる実用的な手法を提案している．更に，提案法によって減らすことのできる分散の量を理論的に評価するとともに，ロボット制御に関する様々な計算機実験を通してその実用性を実証している．

第５章「Conclusions and Future Work」では，本論文の成果をまとめると共に，今後の課題を示している．

以上を要するに本論文は，強化学習における政策探索法で用いられる勾配推定量の安定性の向上，および，標本の再利用に資するものであり，提案手法の有効性が理論的かつ実験的に示されていることから，工学上，及び，工業上貢献するところが大きい．よって我々は，本論文が博士（工学）の学位論文として十分価値あるものと認める．