論文 / 著書情報
Article / Book Information

題目(和文)	Online decision making in non-stationary Markovian environments
Title(English)	Online decision making in non-stationary Markovian environments
著者(和文)	MaYao
Author(English)	Yao Ma
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10040号, 授与年月日:2015年12月31日, 学位の種別:課程博士, 審査員:杉山 将,徳永 健伸,篠田 浩一,村田 剛志,藤井 敦
Citation(English)	Degree:, Conferring organization: Tokyo Institute of Technology, Report number:甲第10040号, Conferred date:2015/12/31, 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		Yao Ma	
論文審査 審査員		氏　名	職　名		氏　名	職　名
	主査	杉山　将	連携教授	審査員	藤井 敦	准教授
	審査員	徳永 健伸	教授			
		篠田 浩一	教授			
		村田 剛志	准教授			

論文審査の要旨（2000字程度）

本論文は「Online Decision Making in Non-stationary Markovian Environments」と題し，英文５章から成っている．

第１章「Introduction」では，本論文の対象であるマルコフ性を持つ非定常的な環境におけるオンライン意思決定問題について論じ，さらに本論文の全体構成を示している．オンライン意思決定問題とは，意思決定者が環境との相互作用を通じて最適な意思決定政策を見つける問題である．環境が定常的な場合，意思決定問題はマルコフ決定過程として定式化されるのが一般的であり，報酬和を最大化する政策を見つけることが目的である．これまでに動的計画法や強化学習法などが提案されてきた．一方，環境が非定常的な場合，意思決定問題はオンラインマルコフ決定過程として定式化される．オンラインマルコフ決定過程の枠組みでは，リグレットとよばれる，最適な政策によって得られる報酬和からの差を漸近的に最小化する，すなわち，ステップ数Tに対するリグレットを$o(T)$にする政策を求めることが目的である．これまでに提案されてきたオンラインマルコフ決定過程の学習アルゴリズムは，計算量が多いため大規模な状態・行動空間を持つ問題に適用することが困難であり，また，連続的な状態・行動空間を持つ問題に直接適用できないという制限があった．このような背景のもと，本論文の目的は，大規模あるいは連続的な空間を持つオンラインマルコフ決定過程における新しい意思決定アルゴリズムを提案することであると述べている．

第２章「Background Knowledge and Related Work」では，オンラインマルコフ決定過程に対する既存のアルゴリズムを概観している．具体的には，まず，エキスパートアルゴリズムを用いた古典的なオンライン意思決定手法として，重み付き多数決法，乱択重み付き多数決法，摂動リーダー追跡法のアルゴリズム，および，それらのリグレット解析を紹介している．そして，これらをオンラインマルコフ決定過程に拡張した，マルコフ決定過程エキスパート法と怠惰摂動リーダー追跡法のアルゴリズムとリグレット解析を紹介している．次に，オンライン凸最適化法であるオンライン勾配降下法とオンライン鏡面降下法のアルゴリズムとリグレット解析を紹介し，それらをマルコフ決定過程に拡張したアルゴリズムとそのリグレット解析を紹介している．

第３章「Online Policy Gradient for Continuous States and Actions Online MDP」では，オンライン政策勾配法を提案し，そのリグレット上界を与えている．オンライン政策勾配法は，パラメトリックな政策モデルを用いることにより，連続的な状態・行動空間を直接扱うことができるという長所を持っている．適当な凸仮定のもとでリグレットが$O(\sqrt{T})$であることを証明するとともに，さらに強凸仮定のもとではリグレットが$O(\log T)$まで少なくできることも証明している．また，報酬関数全体でなく現在の状態と行動での報酬値のみしか観測できないという状況でも，リグレットが$O(\sqrt{T})$であることを証明している．最後に，これらの理論解析の妥当性を数値実験により確認している．

第４章「Online MDPs with Policy Iteration」では，政策反復法に基づくオンライン意思決定アルゴリズムを提案し，そのリグレット上界を与えている．この手法では政策反復法に関数近似を組合せることにより，オンライン政策勾配法で必要だった凸条件を仮定しなくても，$o(T)$のリグレットが達成できることを証明している．また，提案アルゴリズムの計算量が従来法よりも少ないことも理論的に証明している．さらに，これらの理論解析の妥当性を数値実験により確認している．

第５章「Conclusions and Future Work」では，本論文の成果を総括し，今後の課題を述べている．

以上を要するに本論文は，機械学習分野における非定常的マルコフ環境下でのオンライン意思決定技術を発展させるものであり，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．