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)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	Computer Science 専攻	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of (Engineering)
学生氏名： Student's Name	Ma Yao	指導教員（主）： Academic Advisor(main)	Masashi Sugiyama
		指導教員（副）： Academic Advisor(sub)	Takenobu Tokunaga

要旨（英文 800 語程度）
Thesis Summary (approx.800 English Words)

Decision making in non-stationary environments gains extensive attention since environments are not always fixed in reality. The tasks are more challenging due to the lack of information about the future evolution of the environments. In this thesis, we considered the online decision making problem which is a sequential interaction between a decision maker and a non-stationary environment. How can we learn the best strategy which gains the maximum benefit (minimum cost) in this unknown changing environment without any priority knowledge? Such problems with some specific environments are mathematically formalized as the online Markov decision processes (MDPs) in this thesis. Since this online MDP could be applied to many optimization problems from robotics to finance with nonstationary environments, a growing number of researchers proposed online MDP algorithms and analyzed their online performance with theoretical guarantees. At the same time, the demand of solving online MDP problems with large (continuous) state space has arisen naturally in reality. Especially, the online MDP problem with continuous state and action spaces is a challenging research direction since we cannot perform greedy methods searching for the best policy. In recent years, some online MDP algorithms have achieved exciting results based on two kinds of main ideas: expert algorithm based methods and online linear optimization based methods. However, these algorithms are not extensible to the continuous setting without additional assumptions since they learn the action distribution for each state individually. In order to handle large (continuous) state space online MDP problems, we proposed two algorithms which are aimed at learning action distributions for every state jointly.

Firstly, we proposed the Online Policy Gradient (OPG) algorithm that can be implemented in a straightforward manner for online MDP problems with continuous state and action spaces. The proposed algorithm utilizes the parameterized policy model which is natural for handling continuous state and action spaces. Through regret analysis, we showed that the proposed algorithm achieves a sublinear regret, which means our algorithm performs asymptotically equal to the best fixed policy. More precisely, the regret against the best fixed policy of the proposed OPG algorithm is $O(\sqrt{T})$ with full information of reward functions under a concavity assumption. Furthermore, the OPG algorithm achieves $O(\log T)$ regret with full information feedback under a strong concavity assumption. In practice, full information of reward functions may be hard to acquire, but only the value of the reward function for the current state and action is available. Such a setup, called the bandit feedback scenario, has attracted a great deal of attention recently. We showed that the OPG algorithm also achieves $O(\sqrt{T})$ regret with the bandit feedback under a concavity assumption. We also demonstrated the performance of our algorithm with the target tracking and the linear-quadratic regulator experiments, which verify that our algorithm improves the performance for continuous tasks and substantiates the theoretical results. To the best of our knowledge, this is the first work to give an online MDP algorithm that can handle continuous state and action spaces with guarantee.

Secondly, we considered that the concavity assumption in the OPG algorithm may not be fulfilled in some real applications. For such a situation, we investigated the large (possibly continuous) state space online MDP problems and proposed the Online Markov Decision Processes with Policy Iteration (OMDP-PI) algorithm. Compared with the state-of-the-art algorithms, the OMDP-PI algorithm aims at large state space online MDP problems where less computational complexity and the function approximation are necessary. To this end, the proposed OMDP-PI algorithm is motivated by the idea of combining the function approximation with policy iteration, which parameterizes the value function and constructs the policy directly. Moreover, OMDP-PI has a close form update rule and could be performed in $O(|S|^2.3728639+|S|^2 |A|)$ at each time step, which is more efficient than existing methods. With full information of reward functions, the proposed OMDP-PI algorithm is proved to achieve a sublinear regret against the best fixed policy. A significant benefit of the OMDP-PI algorithm is that a linear approximation could be used together with the OMDP-PI

algorithm for large (continuous) state space problems, where the convergence is guaranteed. Furthermore, the OMDP-PI algorithm could be extended to a more general algorithm called the online Markov decision processes with stochastic iteration (OMDP-SI) algorithm. Under some additional assumptions, the OMDP-SI algorithm achieves a sublinear regret as well. Through a grid world experiment, we illustrated the experimental performance of the OMDP-PI algorithm, which verifies the theoretical regret analysis.

By the solid theoretical results, we concluded that the proposed algorithms could handle online MDP problems with large (continuous) state space. Regardless of the change in reward functions, the proposed algorithms always asymptotically perform equally to the best time independent policy in hindsight.