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)	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 (Engineering) Doctor of
学生氏名： Student's Name	Tingting Zhao	指導教員（主）： Academic Advisor(main)	Masashi Sugiyama
		指導教員（副）： Academic Advisor(sub)	

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

Reinforcement learning (RL) is concerned with how an agent ought to take actions in an unknown environment so that expected future rewards are maximized, which offers a framework to robotics such that a robot can autonomously discover the optimal action through the interaction with the underlying environment. Model-free reinforcement learning is a flexible framework in which decision making policies are directly learned without going through explicit modeling of the environment. The RL methods developed so far can be categorized into two types: Policy iteration where policies are learned based on value function approximation and policy search where policies are learned directly to maximize expected future rewards.

In the policy iteration framework, approximation of the value function for the current policy and improvement of the policy based on the learned value function are iteratively performed until an optimal policy is found. Thus, accurately approximating the value function is a challenge in the value function based approach. However, because policy functions are learned indirectly via value functions in policy iteration, improving the quality of value function approximation does not necessarily yield a better policy function. Furthermore, because a small change in value functions can cause a big change in policy functions, it is not safe to use the value function based approach for controlling expensive dynamic systems such as a humanoid robot. Another weakness of the value function approach is that it is difficult to handle continuous actions because a maximizer of the value function with respect to an action needs to be found for policy improvement. Therefore, policy iteration algorithms in the robotics context is not directly applicable.

On the other hand, in the policy search approach, policy functions are determined so that expected future rewards are directly maximized. Policy search can handle continuous states and actions naturally, it is very suitable for solving the robot control tasks. Among policy search methods, gradient-based methods are popular in physical control tasks because policies are changed gradually and thus steady performance improvement is ensured. However, a classic policy gradient method called REINFORCE tends to produce gradient estimates with large variance, which results in unreliable policy improvement. To cope with this problem, a method called policy gradients with parameter-based exploration (PGPE) was proposed. The experimental success of PGPE was demonstrated; however,

theoretical properties were not clear.

In this thesis, we first proved that the variance of gradient estimates in PGPE is smaller than that of REINFORCE under mild assumptions. We then derived the optimal baseline for PGPE, which contributes to further reducing the variance. We also theoretically showed that PGPE with the optimal baseline is more preferable than REINFORCE with the optimal baseline in terms of the variance of gradient estimates. In addition to the solid theoretical analyses, the proposed methods were experimentally shown to yield state-of-the-art results on a variety of problems.

The standard PGPE still requires a relatively large number of samples to obtain accurate gradient estimates, which can be a critical bottleneck in real world applications that require large costs and time in data collection. In order to solve this problem, we combined the following three ideas and gave a highly data effective policy gradient method: (a) PGPE, which is a policy search method with the low variance of gradient estimates, (b) an importance collection sampling technique, which allows us to effectively reuse previously gathered data, and (c) an optimal baseline technique, which minimizes the variance of gradient estimates while the unbiasedness of the gradient estimates is maintained. For the proposed method, we gave theoretical analyses of the variance of gradient estimates and showed its usefulness through extensive experiments. Moreover, we also investigated the benefit of the proposed method in complex high-dimensional humanoid robotic experiments, and the results showed that the proposed method yield state-of-the-art results.

Overall, this thesis contributed to developing statistical reinforcement learning algorithms, which enable the robot to autonomously discover the optimal behavior in the unknown environment. Given the solid theoretical analyses and the encouraging experimental results, we conclude that the proposed methods compare favorably with the corresponding state-of-the-art methods. Therefore, they can be applied to real-world robot control tasks and worth a further study in the future.