Regularized Anderson Acceleration for Off-Policy Deep Reinforcement Learning

Wenjie Shi[*], **Shiji Song**, **Hui Wu**, **Ya-Chu Hsu**, **Cheng Wu**, **Gao Huang**
Beijing National Research Center for Information Science and Technology (BNRist)
Department of Automation, Tsinghua University, Beijing, China
{shiwj16, wuhui14, xuyz17}@mails.tsinghua.edu.cn
{shijis, wuc, gaohuang}@tsinghua.edu.cn

Abstract

Model-free deep reinforcement learning (RL) algorithms have been widely used for a range of complex control tasks. However, slow convergence and sample inefficiency remain challenging problems in RL, especially when handling continuous and high-dimensional state spaces. To tackle this problem, we propose a general acceleration method for model-free, off-policy deep RL algorithms by drawing the idea underlying regularized Anderson acceleration (RAA), which is an effective approach to accelerating the solving of fixed point problems with perturbations. Specifically, we first explain how policy iteration can be applied directly with Anderson acceleration. Then we extend RAA to the case of deep RL by introducing a regularization term to control the impact of perturbation induced by function approximation errors. We further propose two strategies, i.e., progressive update and adaptive restart, to enhance the performance. The effectiveness of our method [2] is evaluated on a variety of benchmark tasks, including Atari 2600 and MuJoCo. Experimental results show that our approach substantially improves both the learning speed and final performance of state-of-the-art deep RL algorithms.

1 Introduction

Reinforcement learning (RL) is a principled mathematical framework for experience-based autonomous learning of policies. In recent years, model-free deep RL algorithms have been applied in a variety of challenging domains, from game playing [1, 2] to robot navigation [3, 4]. However, sample inefficiency, i.e., the required number of interactions with the environment is impractically high, remains a major limitation of current RL algorithms for problems with continuous and high-dimensional state spaces. For example, many RL approaches on tasks with low-dimensional state spaces and fairly benign dynamics may even require thousands of trials to learn. Sample inefficiency makes learning in real physical systems impractical and severely prohibits the applicability of RL approaches in more challenging scenarios.

A promising way to improve the sample efficiency of RL is to learn models of the underlying system dynamics. However, learning models of the underlying transition dynamics is difficult and inevitably leads to modelling errors. Alternatively, off-policy algorithms such as deep Q-learning (DQN) [1] and its variants [5, 6], deep deterministic policy gradient (DDPG) [7], soft actor-critic (SAC) [8] and off-policy hierarchical RL [9], which instead aim to reuse past experience, are commonly used to alleviate the sample inefficiency problem. Unfortunately, off-policy algorithms are typically based on policy iteration or value iteration, which repeatedly apply the Bellman operator of interest and generally require an infinite number of iterations to converge exactly to the optima. Moreover, the Bellman iteration constructs a contraction mapping which converges asymptotically to the optimal

[*] Contact auther.

[2] The code and models are available at: https://github.com/shiwj16/raa-drl

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.