Softmax Deep Double Deterministic Policy Gradients

Kakhaev Ivan
Department of Computer Science
National Yang Ming Chiao Tung University
kakhaev.c@nycu.edu.tw

1 Introduction

DDPG, suffers from the overestimation problem. TD3, algorithm mitigates the overestimation issue, it can lead to a large underestimation bias. TD3) method leveraging double estimators for the critic.Pan et al. [2020]

In this paper, they propose to use the Boltzmann softmax operator for value function estimation. Also design two new algorithms, Softmax Deep Deterministic Policy Gradients (SD2) and Softmax Deep Double Deterministic Policy Gradients (SD3).

2 Problem Formulation

The reinforcement learning problem can be formulated by a Markov decision process(MDP). They consider a continuous action space, and assume it is bounded. They also assume the reward function r is continuous and bounded.

The objective is to maximize the expected long-term rewards.

3 Theoretical Analysis

First, they theoretically analyze the properties of the softmax operator in continuous action space Second, propose to incorporate the softmax operator into actor-critic for continuous control

They formally define value iteration with the softmax operator. Which updates the value function using the softmax operator iteratively.

The error between the value function induced by the softmax operator and the optimal can be bounded, and can be arbitrarily close to 0.

First show that the softmax operator can help to smooth the optimization landscapes. SD2 is a variant of DDPG that leverages the softmax operator to update the value function, which is the only difference between the two algorithms. Specifically, SD2 estimates the value function using the softmax operator, and the update of the critic of SD2, where the the actor aims to optimize a soft estimation of the return. However, the softmax operator involves the integral, and is intractable in continuous action space. We express the Q-function induced by the softmax operator in expectation by importance sampling. SD2 enables a better value estimation by reducing the overestimation bias in DDPG, for which it is known that the critic estimate can cause significant overestimation.

They propose a novel method to leverage the softmax operator with double estimators, called Softmax Deep Double Deterministic Policy Gradients (SD3). Clipped Double Q-learning is proposed in TD3, which clips the Q-value from the double estimator of the critic by the original Q-value itself. Specifically, TD3 estimates the value function by taking the minimum of value estimates from the two critics. With TD3, i.e., apply the softmax operator to the Q-value from the double critic estimator and then clip it by the original Q-value.

4 Conclusion

Therefore, according to our SD2 and SD3 algorithms, we conclude that the softmax operator can not only reduce the overestimation bias when built upon DDPG, but also improve the underestimation bias when built upon TD3

References

Ling Pan, Qingpeng Cai, and Longbo Huang. Softmax deep double deterministic policy gradients, 2020.