A Note on Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor

Author names
Department of Computer Science
National Chiao Tung University
{xxx, yyy, zzz}@nctu.edu.tw

1 Problem Overview

Please provide a brief overview of the selected paper. You may want to discuss the following aspects:

- The main research problem tackled by the paper Demonstrating that we can devise an off-policy maximum entropy actor-critic algorithm, which we call soft actor-critic (SAC), we focus on sample efficiency and learning stability.
- High-level description of the proposed method Creating an actor-critic architecture with separate policy and value function networks. Off-policy enables reuse of previously collected data for efficiency, entropy maximization enables stability and exploration.

2 Background and The Algorithm

Please present the essential background knowledge and the algorithm in this section. You may also describe the notations and the optimization problem of interest. $\pi^* = \arg\max_\pi E_{(s_t,a_t)\ \rho_\pi}[\sum R(s_t, a_t)]$

Besides maximum entropy, we also want to find the maximum action entropy of each trajectory

$\pi^* = \arg\max_\pi E_{(s_t,a_t)\ \rho_\pi}[\sum R(s_t, a_t) + \alpha H(\pi(.|_t))]$ In order to randomize policy, so the action distribution is randomized Finding maximum entropy let us to construct a neural network which can explore existing better path,

3 Detailed Implementation

Please explain your implementation in detail. You may do this with the help of pseudo code or a figure of system architecture. Please also highlight which parts of the algorithm lead to the most difficulty in your implementation.

$q_\pi(s, a) = r(s, a) + \gamma \sum_{s' \in S} P^a_{ss'} \sum \pi(a'|s') q_\pi(s', a')$

seeing entropy as one part of the reward.

4 Empirical Evaluation

Please showcase your empirical results in this section. Please clearly specify which sets of experiments of the original paper are considered in your report. Please also report the corresponding hyperparameters of each experiment.

Preprint. Under review.

5 Conclusion

Please provide succinct concluding remarks for your report. You may discuss the following aspects:

- The potential future research directions
- Any technical limitations
- Any latest results on the problem of interest Off-policy maximum entropy in deep reinforcement learning algorithm that provides sample-efficient learning while accessing the benefits of entropy maximization and stability. Our theoretical results derive soft policy iteration, which we show to converge to the optimal policy. From this result, we can formulate a soft actor-critic algorithm, and we empirically show that it outperforms state-of-the-art model-free deep RL methods.