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

Manan Tomar : ED14B023

Abstract

This work (Haarnoja et al., 2018) proposes an off-policy actor critic algorithm for the maximum entropy reinforcement learning framework. The method is shown to provide better sample complexity and stable convergence in Mujoco based continuous action tasks as compared to established algorithms, both on and off policy ones.

1. Approach

The development of this method is motivated by sample complexity issues that arise from using on-policy algorithms and ”brittle” convergence i.e. stability issues that require fine tuning of hyperparameters in off-policy methods. The authors propose working with an off-policy method as it allows for sample efficient learning, for example by maintaining a replay buffer. Moreover, a maximum entropy based method is proposed mainly in order to incentivize exploration and provide a policy which distributes equal probability to actions which are equally valued. This is not the case in methods which use a deterministic policy such as Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2015). More concretely, the objective function optimized here is as follows for the expected sum of rewards.

$$J(\pi) = \sum_{t=0}^{T} E_{s_t,a_t}[r(s_t,a_t) + \alpha\mathcal{H}(\pi(.|s_t))] \tag{1}$$

,where α is temperature parameter which controls the relative importance of entropy to the reward. The paper then moves to providing proofs for policy evaluation and policy improvement steps for this modified objective in the tabular case. The Q and V value functions are both parameterized by neural networks to implement this for the function approximation case. The policy $\pi(a_t|s_t)$ is a gaussian with the mean and covariance being represented by neural networks as well. The policy is obtained by minimizing the KL divergence between the set of policies obtained by the gaussian parameterization and the exponential of the current Q value estimate. The V value function is written in terms of the Q function as below :

$$V(s_t) = E_{a_t \sim \pi}(Q(s_t,a_t) - \log \pi(a_t|s_t)) \tag{2}$$

2. Experiments

The authors choose benchmarked Mujoco tasks such as Hopper, Walker, Ant etc. to test their method. They primarily compare this with the performance of DDPG, Proximal Policy Optimization (PPO) (Schulman et al., 2017) and soft Q-Learning (Haarnoja et al., 2017) methods. The results show that Soft Actor Critic performs consistently well on all chosen tasks, providing a significant improvement in more challenging environments such as Humanoid.

In the ablation studies, the authors compare their method with and without a stochastic policy i.e. having a deterministic policy and thus not maximizing the entropy. This case then becomes very similar to DDPG. They show that the deterministic policy is highly unstable across multiple runs, while the stochastic policy is consistent, thus proving the importance of the entropy term in the objective.

The authors also show that increasing the temperature parameter α too much (around 0.1) starts to hurt the performance as the reward magnitude becomes negligible in the overall objective term. This means that the method is quite sensitive to the temperature, which is a potential drawback of the method too.

References

Haarnoja, Tuomas, Tang, Haoran, Abbeel, Pieter, and Levine, Sergey. Reinforcement learning with deep energy-based policies. *arXiv preprint arXiv:1702.08165*, 2017.

Haarnoja, Tuomas, Zhou, Aurick, Abbeel, Pieter, and Levine, Sergey. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

Lillicrap, Timothy P, Hunt, Jonathan J, Pritzel, Alexander, Heess, Nicolas, Erez, Tom, Tassa, Yuval, Silver, David, and Wierstra, Daan. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Schulman, John, Wolski, Filip, Dhariwal, Prafulla, Radford, Alec, and Klimov, Oleg. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.