Reinforcement Learning with Deep Energy-Based Policies

Manan Tomar : ED14B023

Abstract

This work (Haarnoja et al., 2017) proposes a soft Q Learning method which uses energy based policies for optimizing multimodal stochastic policies. Relation of this with entropy regularized reinforcement learning is also studied. The experimental results are shown on two Mujoco tasks, the swimmer and the quadrupedal robot.

1. Approach

The authors begin by exploring the need for having stochastic optimal solutions that try to learn all the ways of peforming a task instead of learning the best one. Moreover, the main motivation comes from requiring stochastic policies that are not restrictive in nature, such as using gaussian or linear policies. This is alleviated by borrowing theory from Energy Based Models (EBM) through which a stochastic policy can be written in the following form :

$$\pi(a_t|s_t) \propto \exp(-\mathcal{E}(s_t, a_t)) \tag{1}$$

where $\mathcal{E}$ denotes an arbitrary energy function. The authors set $\mathcal{E}$ as $\frac{1}{\alpha}Q_{soft}(s_t, a_t)$ so as to connect such energy based models with the entropy regularized reinforcement learning objective. Following this, a soft Q iteration scheme is proposed for tabular settings. A continuous state extension to this problem is also presented, using the following objective :

$$J_Q(\theta) = \mathbb{E}_{s_t \sim q_{st}, a_t \sim q_{at}} \left[\frac{1}{2} (\hat{Q}^{\bar{\theta}}_{\text{soft}}(s_t, a_t) - Q^{\theta}_{\text{soft}}(s_t, a_t))^2 \right] \tag{2}$$

with $\bar{\theta}$ being the parameters of a target network and $\hat{Q}^{\bar{\theta}}_{soft}(s_t, a_t) = r_t + \gamma \, \mathbb{E}_{s_{t+1} \sim \rho_s} \left[V^{\bar{\theta}}_{\text{soft}}(s_{t+1}) \right]$. $V^{\bar{\theta}}_{\text{soft}}(s_{t+1})$ is in turn given by $\alpha \log \mathbb{E}_{q_{a'}} \left[\frac{\exp(\frac{1}{\alpha} Q^{\theta}_{\text{soft}}(s_t, a'))}{q_{a'(a')}} \right]$. This algorithm still requires a tractable way to sample actions to execute them and to estimate the soft value function. This is done by using a stochastic sampling network based on Stein variational gradient descent (Liu & Wang, 2016). Such a network allows sampling actions conditioned on the state and given noise as input. The network is optimized by minimizing the KL divergence between samples collected from the network and using the earlier definition of the policy based on the soft Q and V values. Overall, the algorithm collects data using the sampling network and performs optimization of equation 2 in an off policy manner. Each gradient step is accompanied by an optimization step of updating the policy as well.

2. Experiments

The authors show results on Swimmer and Quadrupedal robot simulations. They compare the performance of the proposed method with DDPG (Deep Deterministic Policy Gradients). The focus here is on learning multimodal policies that help in exploration and thus eventually in solving the overall task. The other tasks involve analyzing performance in cases of fine tuning a previously learnt policy for a given specific task. In such cases, the authors show that a multi modal policy explores well again and thus fine tunes faster and to better asymptotic performance.

3. Critique

Although the method is motivated by the need for generalized stochastic policies, the experiments mainly highlight only one of the discussed advantages, i.e. exploration in the case where multimodal policies are useful. Other benefits such as robust nature of such policies under uncertain dynamics and the usage in imitation learning are not touched upon. It will be interesting to see how well this method performs in such settings. Moreover, the method is only compared to DDPG and not to other works such as stochastic neural networks which use inherently stochastic policy algorithms like TRPO and PPO.

References

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

Liu, Qiang and Wang, Dilin. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pp. 2378–2386, 2016.