ECE 276 Assignment 4 Report : Deep Deterministic Policy Gradients

Jingpei Lu
Jacob School of Engineering
University of California, San Diego

Date: November 15, 2019

1 Deep Deterministic Policy Gradients

For this implementation of the Deep Deterministic Policy Gradients (DDPG) algorithm, the parameters of the Actor and Critic network follow implementation of the original DDPG paper (Timothy et al. 2016). We use Adam as our optimizer and the learning rate for both actor and critic is 1e-3. The discount factor is $\gamma = 0.99$, the soft target updates rate is $\tau = 0.001$, and the initial start position is fixed. The size of our experience replay buffer is 10000. Before the training, we generate 1000 samples with state-transitions taken from random actions and store in the buffer. When sampling the actions, we add a zero mean Gaussian noise with 0.1 variance to the output of the actor network. We train the DDPG using 200000 steps and the networks are updated once for every 10 steps. We evaluate the average returns of DDPG after each 100 steps by executing the policy on the environment without any exploration noise for 5 episodes.

The following plots show the learning curves of DDPG for three different random seeds. The y axis indicates the average undiscounted returns over 5 episode for a given policy. The x axis indicates the number of the steps (the actual step numbers should be the numbers multiplied by 100 since we only evaluate the policy after each 100 steps).

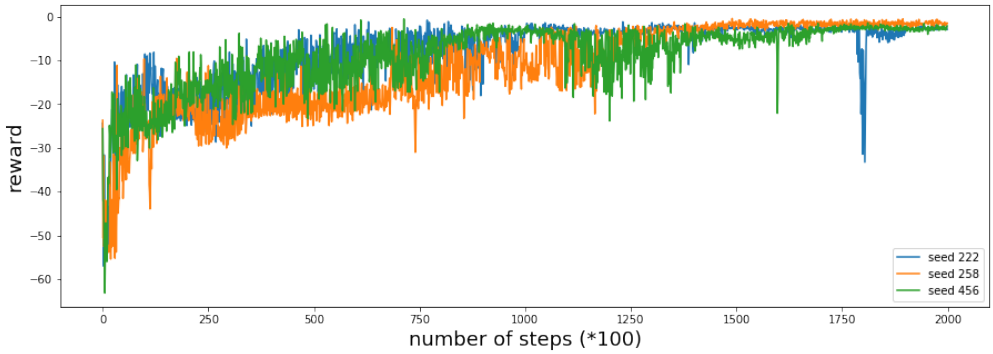

Figure 1: Average returns of the learned policy after each 100 steps with different seeds.

From the above plots, we observe that the DDGP takes around 150000 steps (15000 updates) to converge, which is much faster than the policy gradient as we implemented in homework 3 (policy gradient takes around 1000000 steps to converge).

2 TD3

Our implementation of the Actor and Critic network for the TD3 algorithm is adapted from the open sourced implementation of the original paper (Scott et al. 2018). The learning rate, soft target updates rate, discount factor, replay buffer size and training strategies are the same as the previous experiment of DDPG. The plots below shows the learning curve of the TD3. From my observation, TD3 converges faster than DDPG.

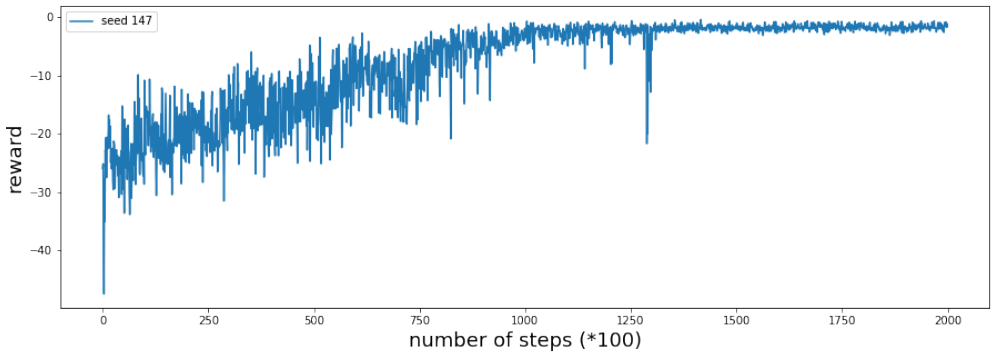

Figure 2: Average returns of the learned policy after each 100 steps.