Homework 5

ELEN E6885: Reinforcement Learning

November 25, 2019

Problem 1 (**DQN Algorithm, 20 credits**): Explain why we use "experience replay" and "fixed Q-targets" in DQN. In particular, explain why using "experience replay" and "fixed Q-targets" can help stabilize DQN algorithm when the correlations present in the sequence of observations (e.g., Atari games)?

Problem 2 (**DDPG,10 Credits**) Recall that DQN is able to stabilize the process of using deep learning for value function approximation, which was believed to be unstable. When it comes to the continuous action space, it is not straightforward to apply DQN directly. One naive approach is to discretize the continuous action space. However, this may not help in practice, why? How does DDPG handle this problem?

Problem 3 (**A3C, 15 Credits**) What is the benefit of using multiple agents in an asynchronous manner in A3C?

Problem 4 (**Gaussian Policy, 25 Credits**)
Assume Gaussian policy is used in the policy gradient reinforcement learning. The Gaussian mean is a linear combination of state features

$$\mu(s) = \phi(s)^T w = \sum_i \phi_i(s) w_i,$$

where $\phi(\cdot)$ represents the vector of feature functions and w represents the weight vector. Also, assume a fixed variance σ^2. Show that that the score function $\nabla_w \log \pi_w(a|s, w)$ of Gaussian policy is given by

$$\nabla_w \log \pi_w(a|s, w) = \frac{(a - \mu(s))\phi(s)}{\sigma^2}.$$

Problem 5 (**MCTS, 30 Credits**)
In this problem, you are asked to review AlphaGo paper "Mastering the game of Go with deep neural networks and tree search" and AlphaGo Zero paper "Mastering the game of Go without human knowledge". After the paper review, you need to

(1) Summarize the RL algorithms used in these papers. Especially, point out the differences between AlphaGo and AlphaGo Zero search algorithm.

(2) AlphaGo Zero is self-trained without the human domain knowledge and no pre-training with human games. Explain in detail how AlphaGo Zero achieves self-training.