CS 285 HW 3

Yanlai Yang

October 21, 2020

1 Question 1 (DQN)

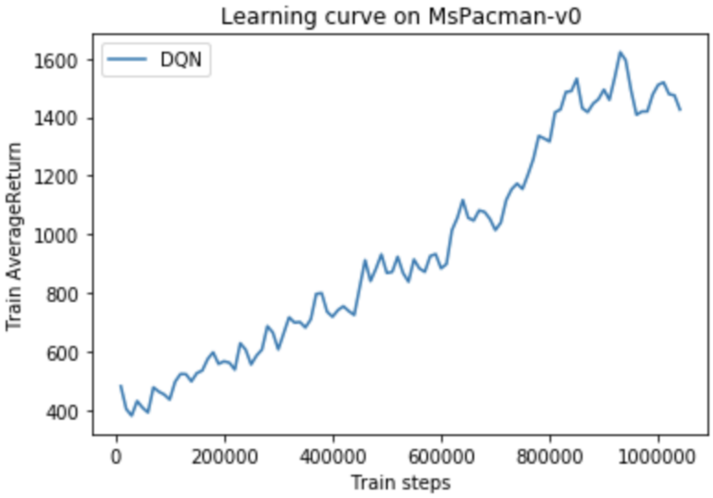

Figure 1: The learning curve plot of DQN implementation on Ms. Pac-Man.

2 Question 2 (DDQN)

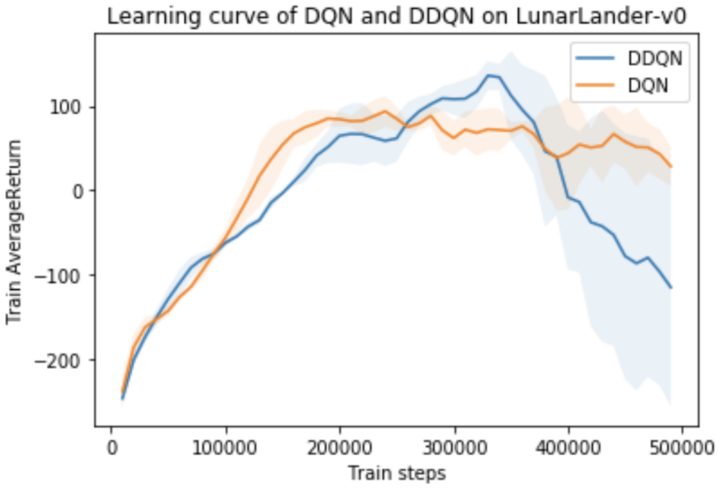

Figure 2: The learning curve plot of DDQN and vanilla DQN on LunarLander-v3. The results are averaged across three runs with different random seeds and the shaded area denotes the error bar (standard deviation of the return across different runs). We observe that DDQN indeed reaches a much higher best average return than vanilla DQN.

3 Question 3 (Experimenting with Hyperparameters)

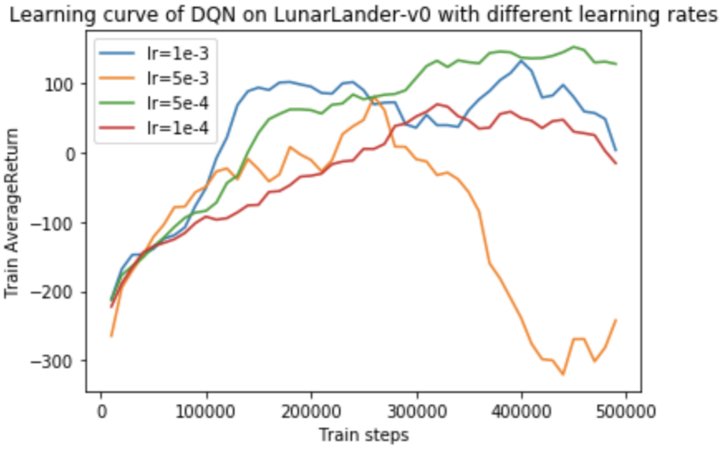

Figure 3: The learning curve plot of DQN with different learning rates. I chose to experiement with different learning rates because I would hope to know how sensitive DQN is to changes in learning rates. The outcome is similar to what is expected: when the learning rate is small (1e-4), the agent learns slowly but steadily; as the learning rate increases, the agent learns faster and reaches a higher best average return (it seems that 5e-4 is the optimal learning rate for this experiment); as the learning rate further increases (1e-3), it learns faster in the beginning but oscillates more often afterwards; as the learning rate becomes even larger (5e-3), the algorithm diverges very soon. The learning rates from 1e-4 to 5e-3 are all used often for optimizers such as SGD and Adam, while the outcomes are very different, as shown in the figure. The results show that the DQN algorithm is fairly sensitive to changes in the learning rate.

4 Question 4 (Actor Critic on Cartpole)

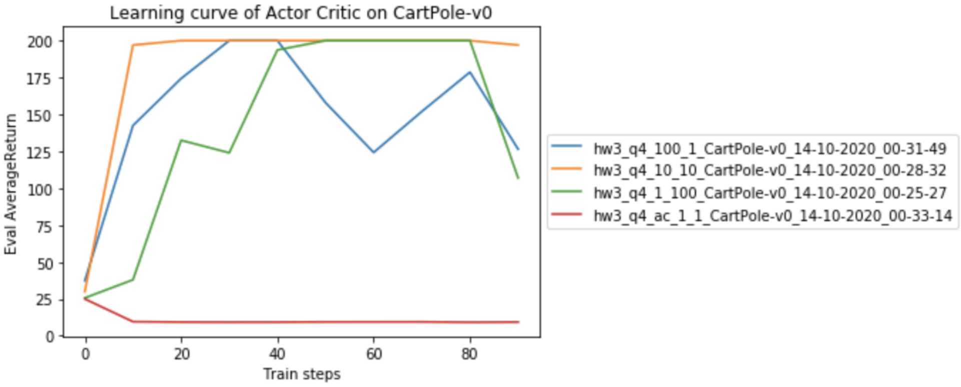

Figure 4: Alternating between taking one target update and one gradient update step (red curve) does not work at all. Updating the target too often (blue curve) also does not work very well, since we need a fixed target value to make the training more stable. It turned out that increasing both the number of target updates and number of gradient updates (orange curve) works best, since in which case the target value will be fixed for several iterations (we will not try to hit a moving target) and we will not take too many gradient steps towards a very inaccurate target in the beginning of training.

5 Question 5 (Actor Critic on More Difficult Tasks)

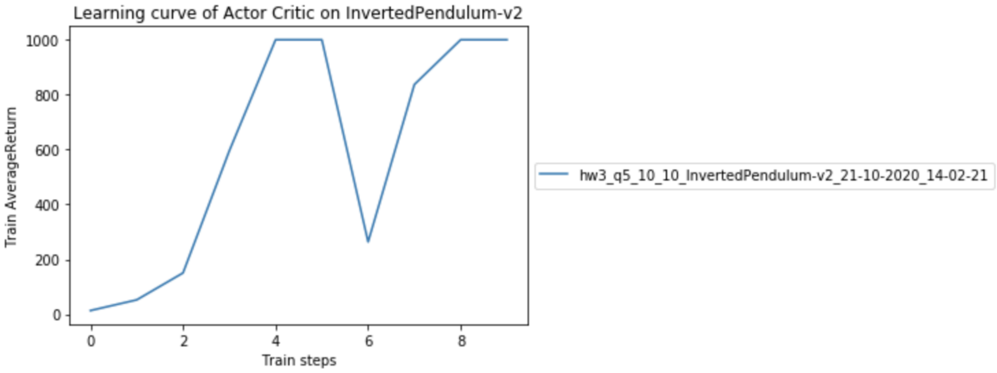

Figure 5: The learning curve plot of actor critic on InvertedPendulum-v2, with the best hyperparameter setting from the previous question. It learns very quickly and converges to reward 1,000.

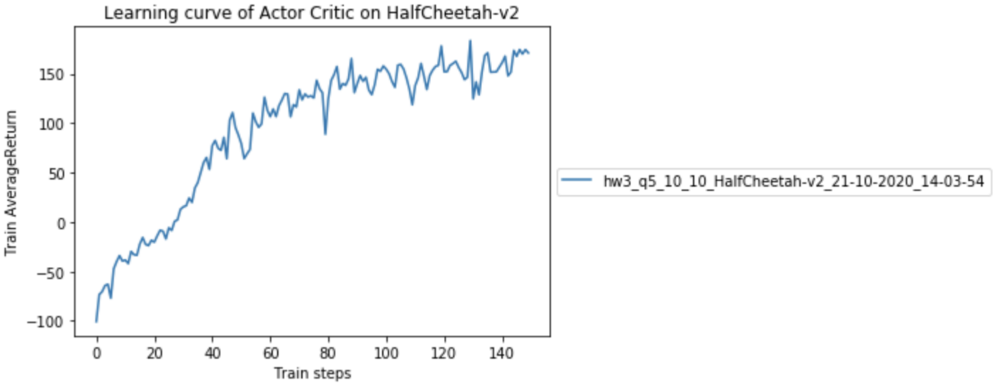

Figure 6: The learning curve plot of actor critic on HalfCheetah-v2, with the best hyperparameter setting from the previous question. It converges to around 150 reward within 150 iterations.