CS 4789 PA 3 Discussion

Chad Yu

March 19, 2024

1 Question 1

With the default hyperparameters, for the Cartpole-v0 environment, the mean reward and mean episode length are both 200.0, while for the LunarLander-v2 environment, the mean reward and mean episode length are 60.965 and 350.2, respectively.

2 Question 2

2.1 2a

Given the default setup of the problem, with the given hyperparameters, and the location/frequency of the target update within the training process, training with the target network was not important. In fact, by simply removing the target network, I got significantly better performance in the LunarLander-v2 environment, with both a shorter average episode length at 307.8, and a higher average reward at 251.38, and I got the exact same average episode length and average reward for the Cartpole-v0 environment. I think that the improvement in not using the target network could come from how frequent we do the target update and the specific τ that we use in the default hyperparameters. In general, having a lower target network update frequency with respect to the learning frequency could roughly make the network and target network get further apart each episode that the network is updated but the target network is not. This can cause instability in the training and make it slower as well, as the network will go through several training steps, and then get worse loss as the target is updated and the target network is again initialized to something further away from the network we're training, so that it will take more iterations/episodes to reduce the loss a significant amount. Especially in the case of the LunarLander-v2 default hyperparameters, we have that $\tau = 0.01$ so that the update puts almost no weight on the training network in factoring into the new target network, so that there is more loss after the update than if the training network had more weight.

2.2 2b

For the LunarLander-v2 environment, putting the update targets call within the episode loop improves the average reward incurred greatly, up to 229.175, but the average episode length goes up to 404.4. However, I don't see this as much of an issue, as I could clearly see that this trained model was very accurate; it would wobble a little in the beginning, but the spaceship would always correct itself and push itself within the flags by the end for each trial. For the Cartpole-v0 environment, again, I am getting the same average episode length and average reward incurred as the previous two cases tested. Note that only the location of the target update is changed in these experiments, not any of the other hyperparameters.

2.3 2c

For the LunarLander-v2 environment, changing τ to 0.75 and the target update frequency to 1 improved the performance greatly, with even shorter average episode length and higher average reward than removing the target network, at 298.2, and 267.34, respectively. Also, the simulation itself performed very well, landing the spaceship within the flags consistently and efficiently, with little to no wobbling. On the other hand, for the Cartpole-v0 environment, although changing some hyperparameters renders a less wobbly visual result (e.g. $\epsilon = 0.7$), we have that the maximum average episode length and average reward incurred is always at 200.0. Some hyperparameter changes made these quantities lower, but visually rendered a simulation that didn't keep the cartpole upright well.