CS285 HW 2 by Tomas Santos

Question 1: basic Q-learning performance.

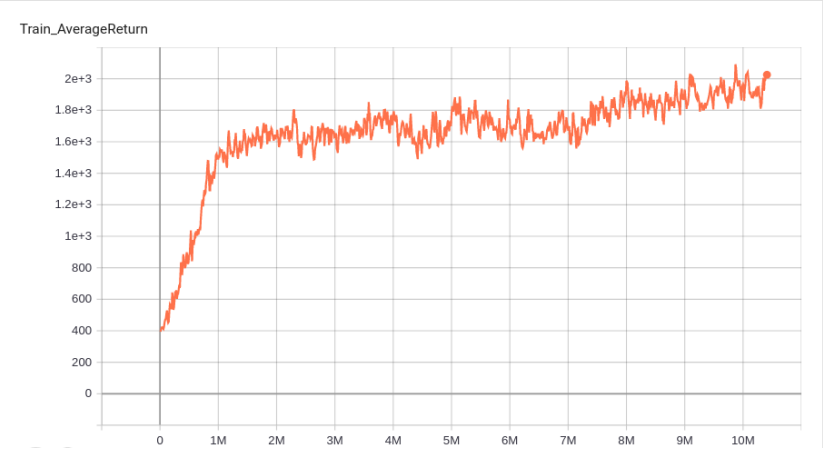

Figure 1: MsPacman-v0

Question 2: double Q-learning (DDQN)

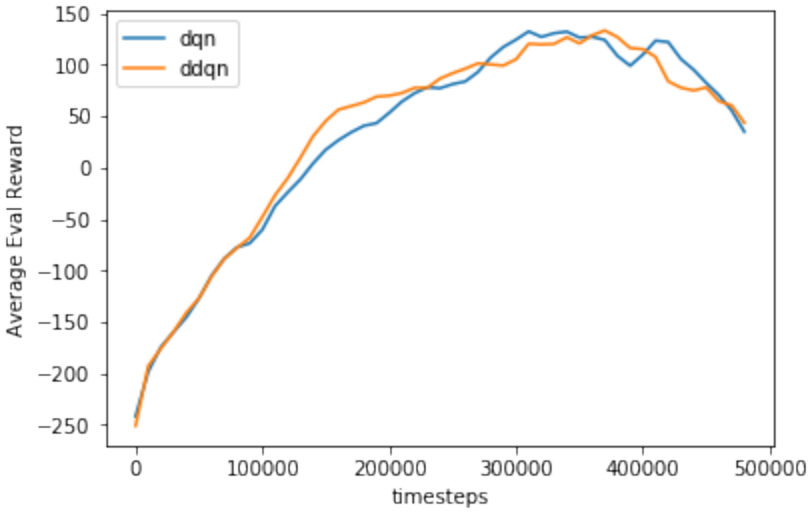

Figure 2: Average DQN vs DDQN

It's pretty hard to tell the difference between the ddqn and the dqn. I ended up doing an average over 10 or so runs which led to the above graphs. It seems as though ddqn has a slight edge over the regular dqn, but perhaps this experiement is too simple for the differences to shine through. Unfortunately I don't have time to test it with a longer experiement like MsPacman.

Question 3: experimenting with hyperparameters

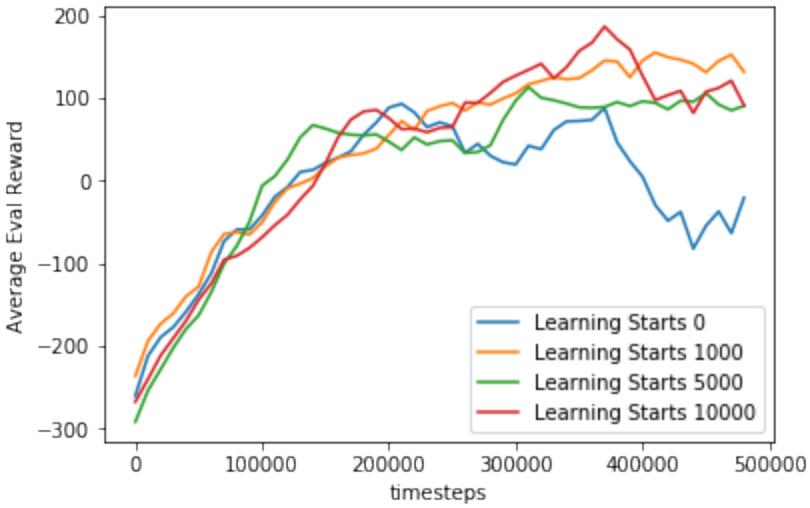

Figure 3: Modifying Batch Size

I'm very pleased with this plot because it demonstrates exactly what we would expect to see when we vary how early the learning starts. If we immediately try to use "learned" values, then they're really just going to be randomly choosen but the algorithm is not goign to explore as much as it should. As a result the learner ends with much lower retrun than either of the other two experiments. The difference between the green and the orange plot are more subtle, but it's clear that the green at one point had the highest return and we can theorize this is because it has the most knowledge of the state space to exploit because it explored the most in the beginning. Another interesting observation is that for the first 10000 or so steps, the learners are all very close to each other. It doesn't even seem like the learner with the most delays "learning_start" parameter was significantly handicamped in the beginning and it clearly outshines the others in the end.

Question 4: Sanity check with Cartpole

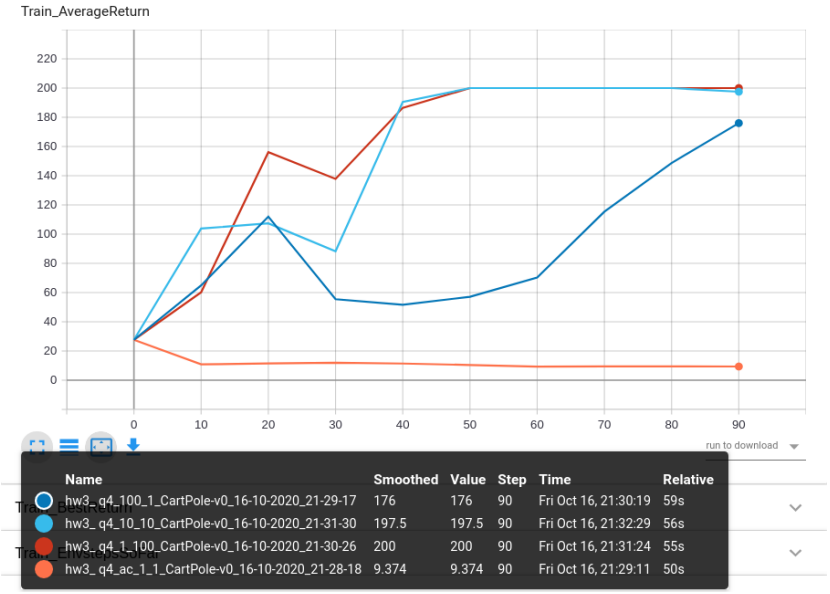

Figure 4: Different Actor-Critic update step numbers

Clearly, the worst performing was an actor ciritc with only one update step for both actor and critic. The two highest were 1) when we had 100 update steps for the value function and only one step for updating the target function, and 2) when we updated the target 10 times and the value function 10 times. I presume the later is better even through it's hard to tell from the graph because with only one change to the target function it's possible that we're "aiming at the wrong target" and but we're aiming at it really well. I think that's probably why the light blue seems to be about to fall off from the 200 mark and the red one converges solidly.

Question 5: Run actor-critic with more difficult tasks

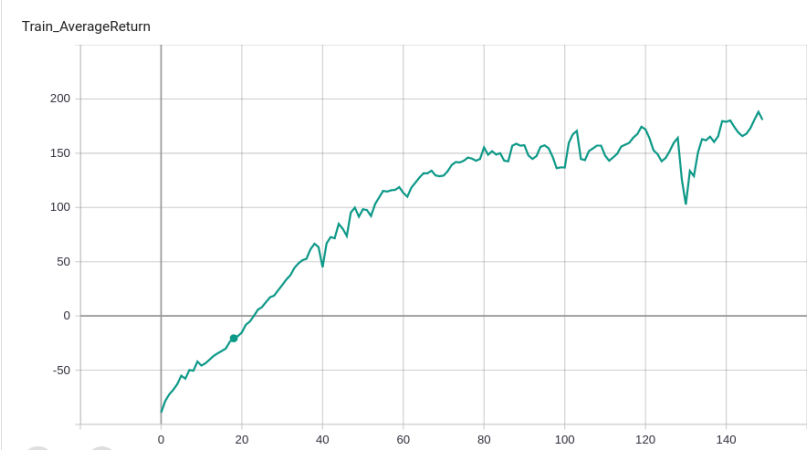

Figure 5: Cheeta

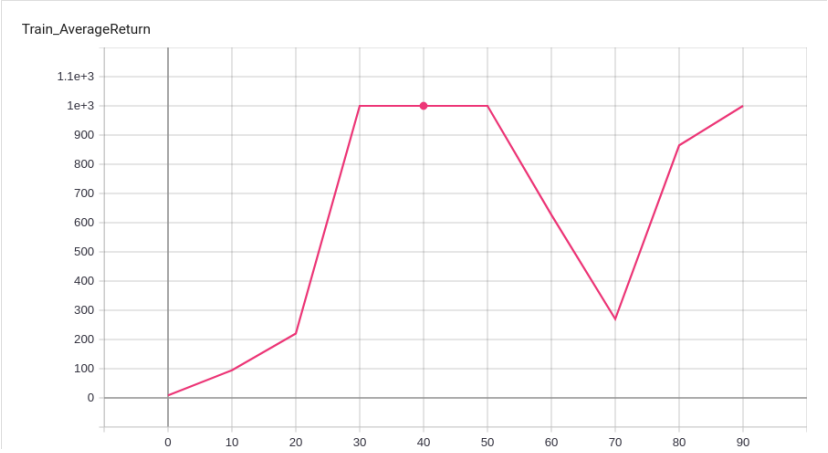

Figure 6: Inverted Pendulum