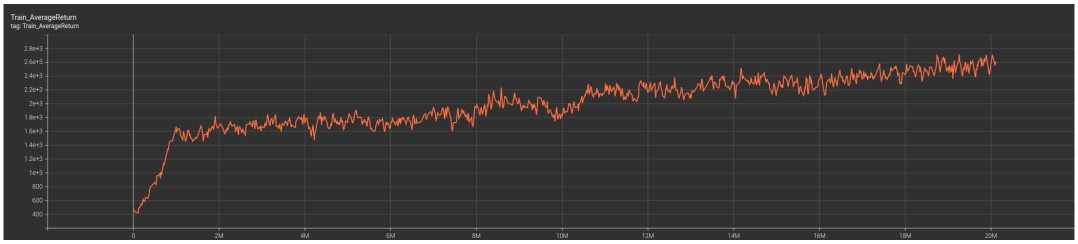
Assignment 3: Q-Learning and Actor-Critic Algorithms

huseyinabanox@gmail.com

January 2023

1 Part 1: Q-Learning

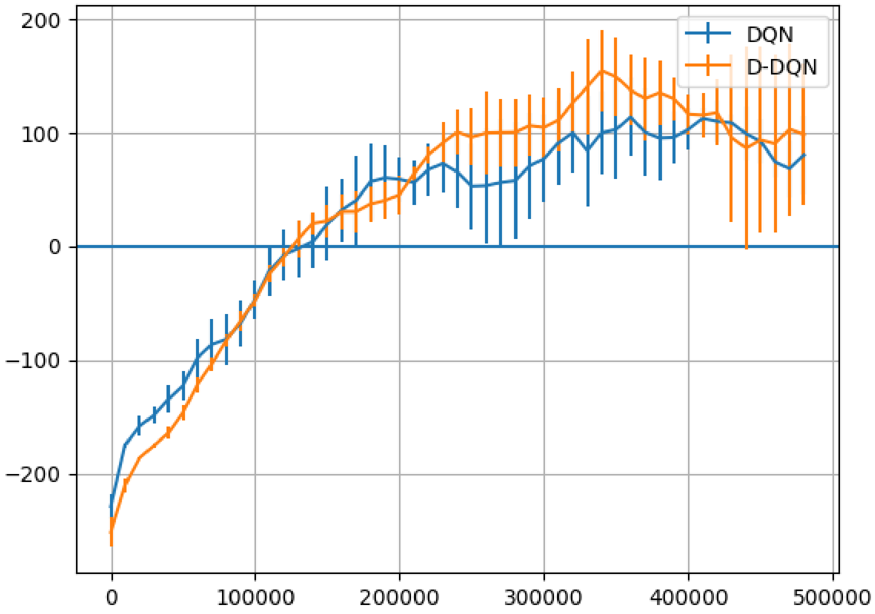
Question 1: basic Q-learning performance (DQN)



After 20M iterations the learning curve looks like above. It is taken from tensorboard. Relevant log files can be found under data folder.

Double Q network obtains better returns, as expected.

Question 2: double Q-learning (DDQN)



LunarLander-v3 environment is used for 3 seeds per configuration e.g. DQN vs D-DQN as stated in the question. Different seed results are averaged and the plot above is generated. combine_results_q2.py file is used to combine results.

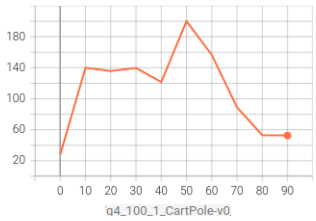
Question 3: experimenting with hyperparameters

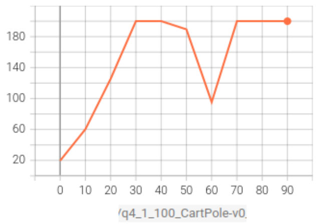
TODO: solve this question

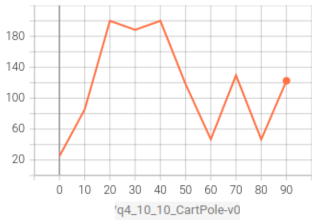
2 Part 2: Actor-Critic

Question 4: Sanity check with Cartpole

Actor critic algorithm is tested in cartpole environment. Different target update parameters and gradient steps parameters are tried.







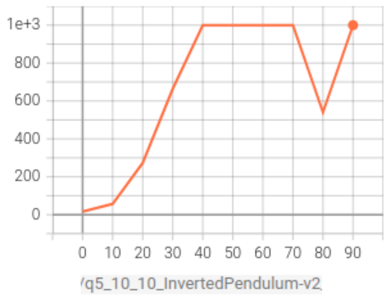
.

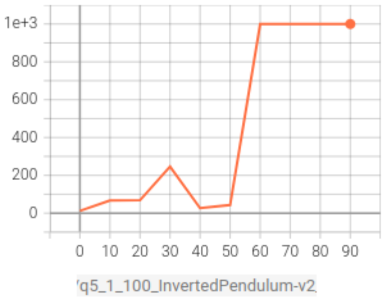
Subtitles show the used parameters.

Best performance is obtained when both parameters are set to 10. If target update is set to 1 and gradient steps are set to 100, learning seems to be more stable. This results shows the importance of gradient steps.

Question 5: Run actor-critic with more difficult tasks

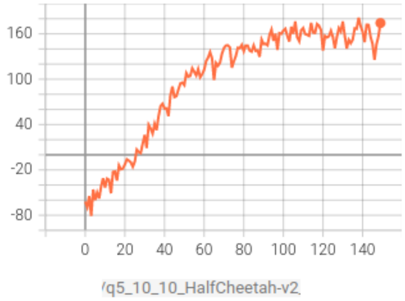
Best results are obtained when both parameters are set to 10.

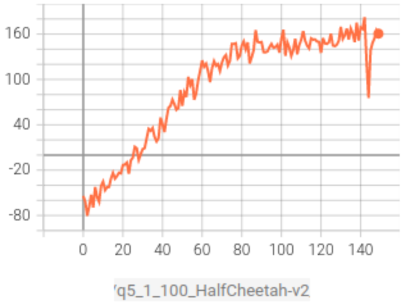




.

After 100 iterations, InvertedPendulum return is around 1000, as expected. After 20 iterations, InvertedPendulum return should is above 100, as expected.





.

After 150 iterations, HalfCheetah return is around 150, as expected. After 20 iterations, HalfCheetah return is above -40, as expected.

3 Part 3: Soft Actor-Critic

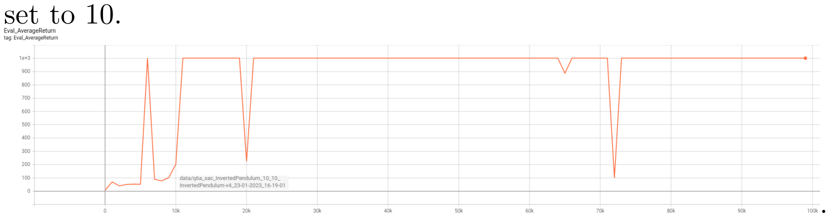
Question 6: Run soft actor-critic more difficult tasks

Actor updates and critic updates are set to 10.

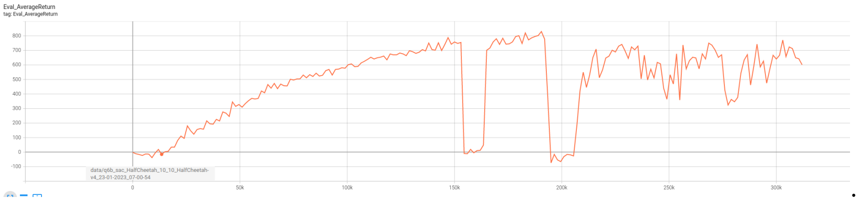
After 20000 steps, InvertedPendulum return is expected to reach 1000. Actually it is reached after 9k steps.

Question 6: Run soft actor-critic more difficult tasks

The same parameters are used as stated in the question. Additionally <num_critic_updates_pe_agent_update>and <num_acto_updates_per_agent_update>parameters are set to 10.



After 10000 steps, InvertedPendulum return is expected to be near or above 100. After 20000 steps, InvertedPendulum return is expected to reach 1000. Our implementation reaches 1000 before 10K steps and stabilizes around 1000 after 10K steps.



After 10000 steps, HalfCheetah return is expected to be above -40 (trending toward positive). After 50000 steps, HalfCheetah return should be around 200. Our implementation return is close to zero at 10K steps and is above 300 after 50K steps.