Assignment 3: Q-Learning and Actor-Critic Algorithms

Yulun Rayn Wu, 3034358565

October 20, 2020

1 Part 1: Q-Learning

- Question 1

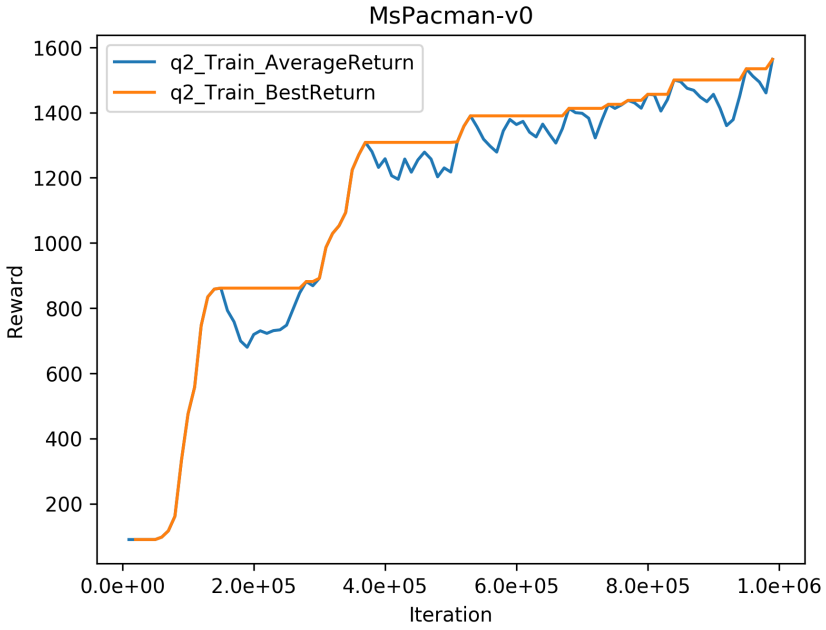


Figure 1: Settings: env_name: MsPacman-v0; ep_len: 200; batch_size: 32; eval_batch_size: 1000; num_agent_train_steps_per_iter: 1; num_critic_updates_per_agent_update: 1; seed: 0.

Configurations: Changed `num_timesteps` to $2e6$ and stopped training at $1e6$. It affected the performance because schedulers are scaled accordingly.

- Question 2

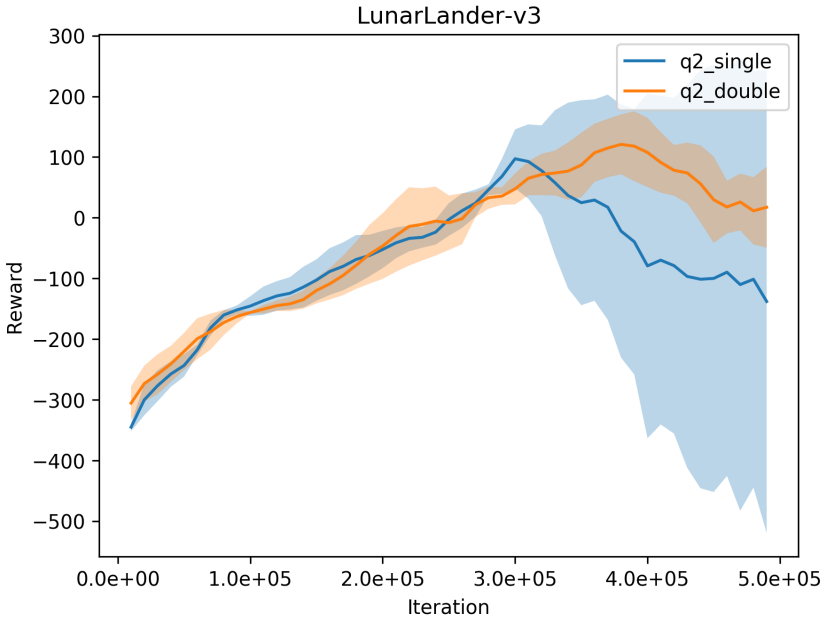


Figure 2: Action Selection: Target vs. Online. **Settings**: env_name: LunarLander-v3; ep_len: 200; batch_size: 32; eval_batch_size: 1000; num_agent_train_steps_per_iter: 1; num_critic_updates_per_agent_update: 1; seed: 0.

In the experiments, double DQN yields better rewards and stability than DQN with vanilla target network.

- Question 3

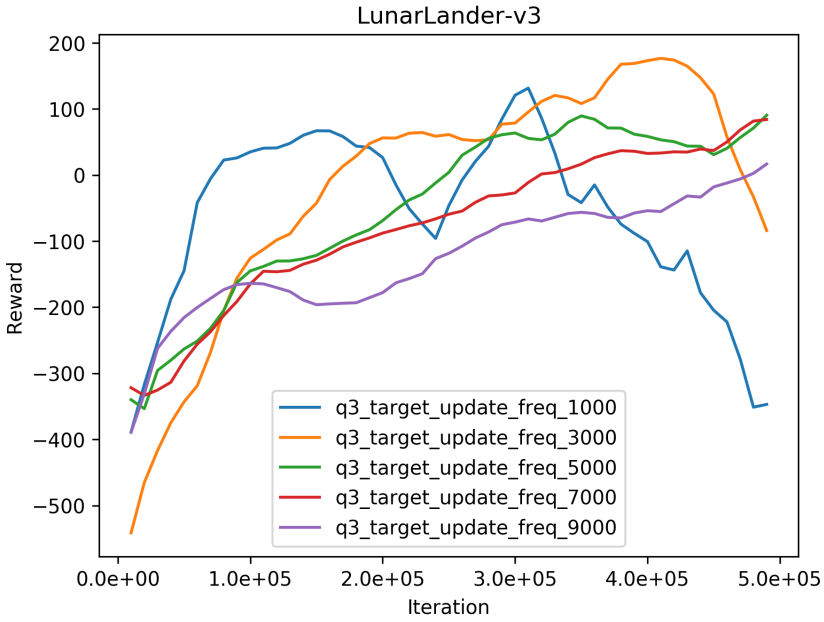


Figure 3: Target Update Frequency: once per 1e3, 3e3, 5e3, 7e3, 9e3 steps. **Settings**: env_name: LunarLander-v3; ep_len: 200; batch_size: 32; eval_batch_size: 1000; num_agent_train_steps_per_iter: 1; num_critic_updates_per_agent_update: 1; seed: 0.

Smaller update frequency for the targets results in better stability, but also makes training slower.

2 Part 2: Actor-Critic

- Question 4

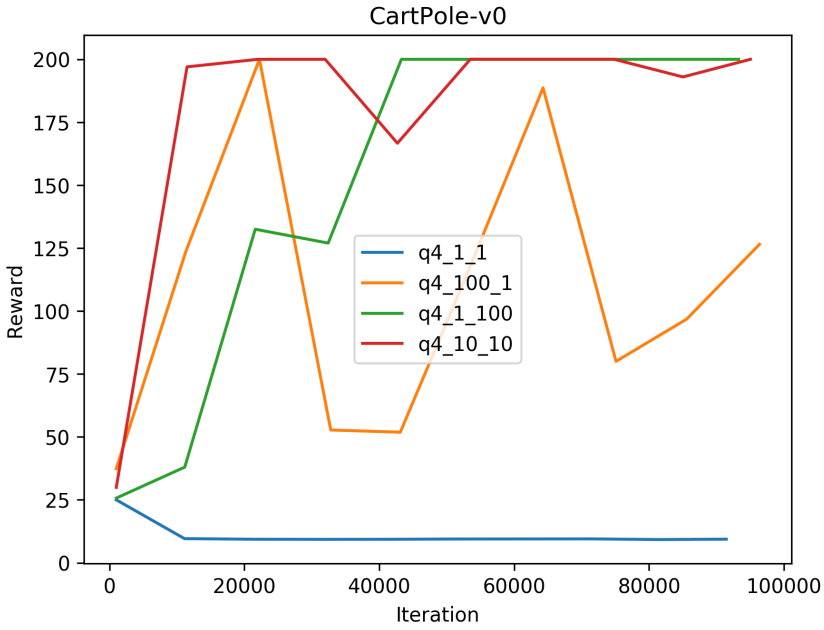
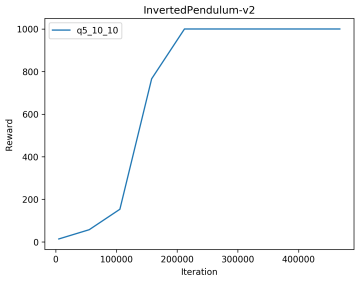


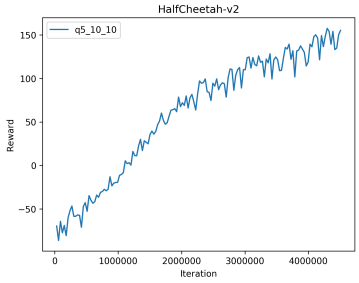
Figure 4: Search: num_target_updates, num_grad_steps_per_target_update. **Settings**: env_name: CartPole-v0; ep_len: 200; n_iter: 100; batch_size: 1000; eval_batch_size: 400; train_batch_size: 1000; discount: 1.0; learning_rate: 5e-3; n_layers: 2; size(hidden layer): 64; seed: 1.

It is hard to draw conclusions based on these 4 experiments alone. Ideally, one might want to use a fairly small frequency for target update for the sake of stability but not so small that it would affect training speed too much, while also having a larger number of gradient steps per update to match the low frequency.

- Question 5



(a) –ep_len 1000 –discount 0.95 –scalar_log_freq 10 -n 100 -s 64 -b 5000 -eb 400 -lr 0.01



(b) –ep_len 150 –discount 0.90 –scalar_log_freq 1 -n 150 -s 32 -b 30000 -eb 1500 -lr 0.02

Figure 5: Environment: InvertedPendulum-v2, HalfCheetah-v2. **Settings**: train_batch_size: 1000; n_layers: 2; seed: 1.