CSE 6369 Homework 2

Krishna Khadka

1001751624

Deliverables

Part I: Completed on the code files

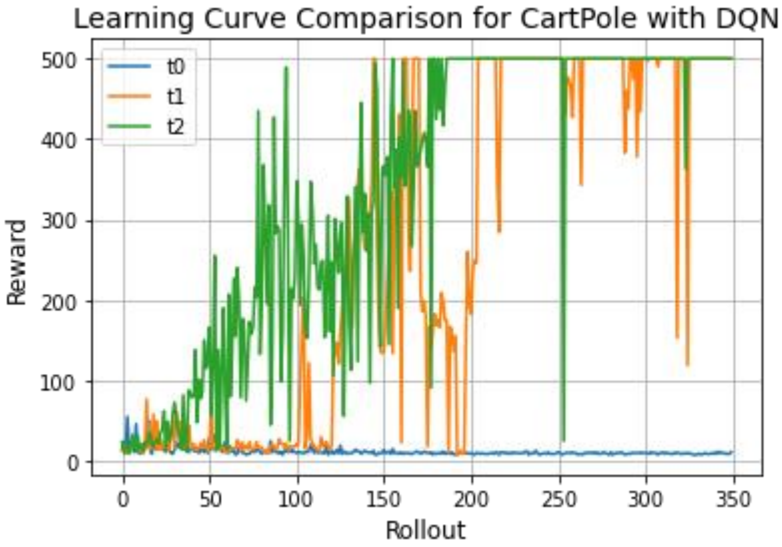
Part II: Experiment 1

1. Graph that compares learning curve from the three trials above.

   t_0 = CartPole_v1_t0

   t_1 = CartPole_v1_t1

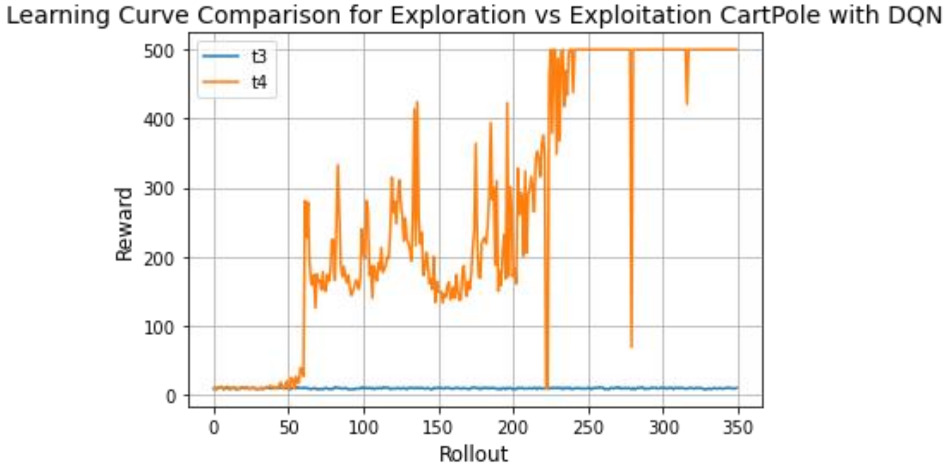
   t_2 = CartPole_v1_t2



Here, the **optimal value for tau is 0.005.**

2. Graph for Exploration vs Exploitation:

   t_3 = CartPole_v1_t3
   t_4 = CartPole_v1_t4



Learning Curve Comparison for Exploration vs Exploitation CartPole with DQN

Here, the optimal experiment is t_4 where the optimal values are 0.1 and 0.05. The tau optimal used was 0.005.

3.

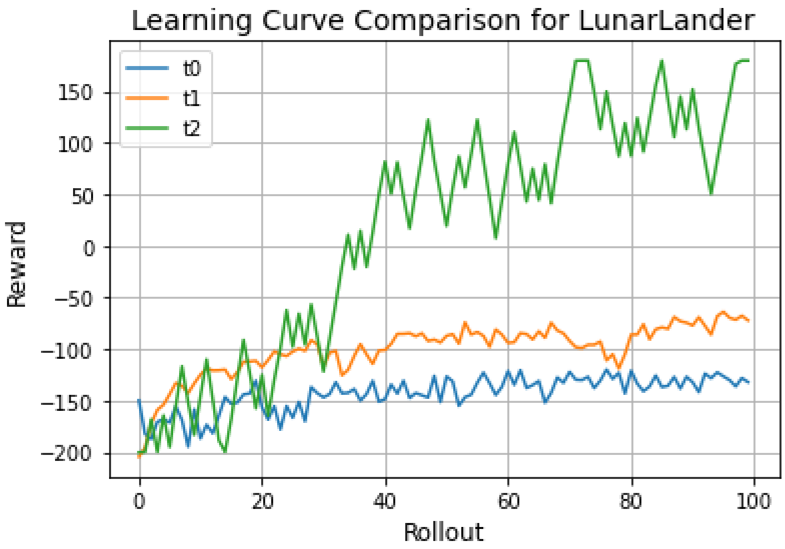
a. Changing the target network update rate significantly affects the learning curve. Lower rates, like in T2 (0.005), result in faster convergence due to more stable learning. Higher rates, as in T0 (0.5) and T1 (0.05), can lead to instability and slower convergence. T2's lowest rate led to the best performance and earliest convergence. Optimal update rates depend on the environment and algorithm used.

b.

Changing the range for ε (exploration parameter) significantly affects the learning curve. T4, with a wider range (0.1-0.05), outperformed T3 (0.0-0.00). T4 showed a dramatic rise in reward after 50 rollouts, reaching a maximum of 500, while T3 remained capped at 10-15 reward. A larger range for ε allows more exploration, aiding in the discovery of better policies and improving performance. Restricted exploration in T3 hindered learning, resulting in lower and stagnant rewards. Increasing ε range positively impacts the learning curve, facilitating faster convergence and better performance.

Part 3 – Experiment II (Lunar Lander Actor Critic)

1. Graph that compares the learning curve from the Actor-Critic Network.
   t0 = LunarLander_v2_t0
   t1 = LunarLander_v2_t1
   t2 = LunarLander_v2_t2



2. Changing the critic network update parameters, specifically the number of iterations and epochs, significantly affects learning performance in the actor-critic algorithm. Increasing the parameters, as seen in T1 (10 iterations, 10 epochs) and T2 (20 iterations, 20 epochs), led to improved performance. T2 performed the best, nearing a reward of 180. By increasing iterations and epochs, the critic network undergoes more extensive training and refinement, enabling better policy evaluation and action selection. This allows the network to capture and model the environment dynamics more accurately, resulting in more informed decision-making. The relationship suggests that more iterations and epochs enhance learning performance in the actor-critic algorithm, improving its ability to learn and achieve higher rewards.