National Taiwan University
Department of Electrical Engineering
Name: Chen-Che Huang
ID: r10946005

112-1 (Fall 2023)
Reinforcement Learning

Assignment 2 Report

Q1. Discuss and plot learning curves under ϵ values of $(0.1, 0.2, 0.3, 0.4)$ on MC, SARSA, and Q-Learning

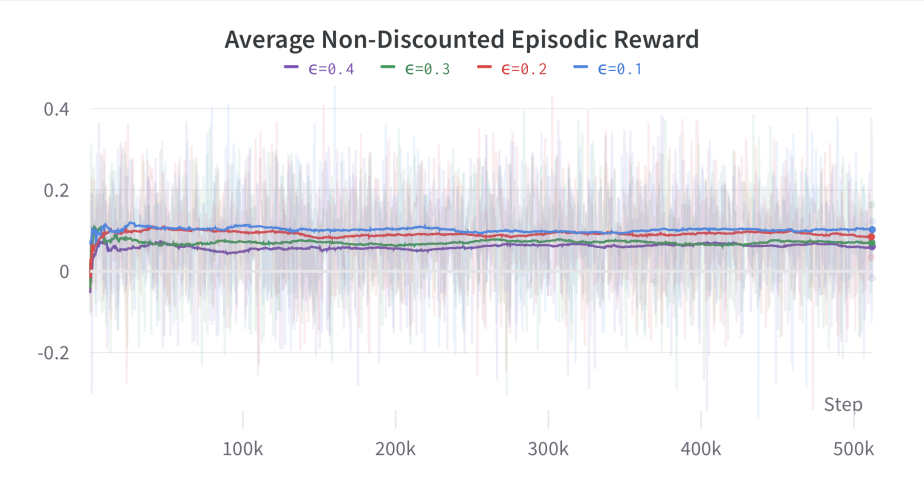

Figure 1: Learning curves of MC

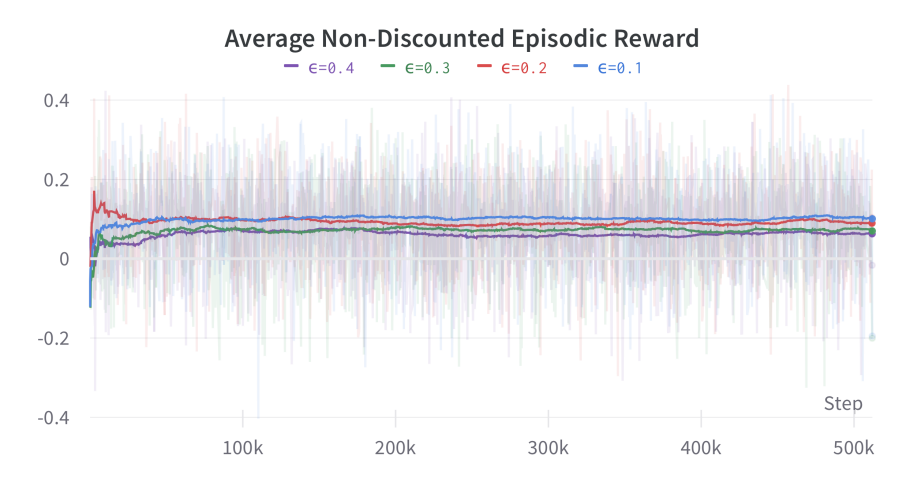

Figure 2: Learning curves of SARSA

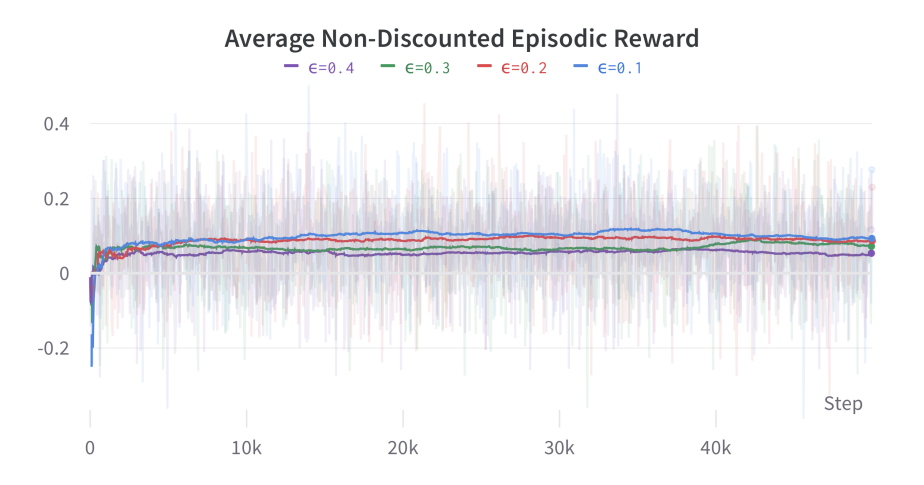

Figure 3: Learning curves of Q learning

We can see that in the beginning of the training, the average non-discounted episodic rewards change dramatically because the Q values and the policies are not accurate. As time goes by, whatever the method is used, the ranks of the rewards under different ϵ values become stable. The experiment with $\epsilon = 0.1$ can obtain the largest episodic reward. The experiments with $\epsilon = 0.2$, $\epsilon = 0.3$, $\epsilon = 0.4$ get the second, third, and forth rank, respectively. The reason is quite obvious and simple. The agent with lower ϵ can chooses the best actions more often to acquire large rewards.

Q2. Discuss and plot loss curves under ϵ values of $(0.1, 0.2, 0.3, 0.4)$ on MC, SARSA, and Q-Learning

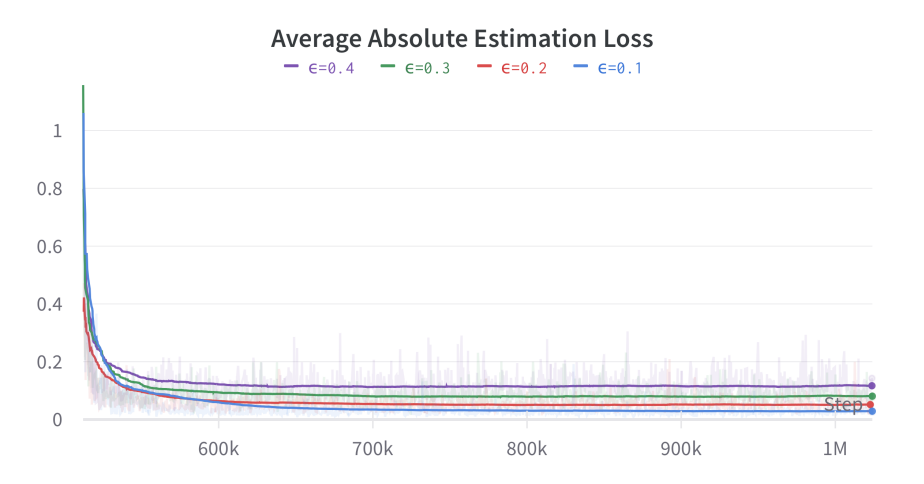

Figure 4: Loss curves of MC

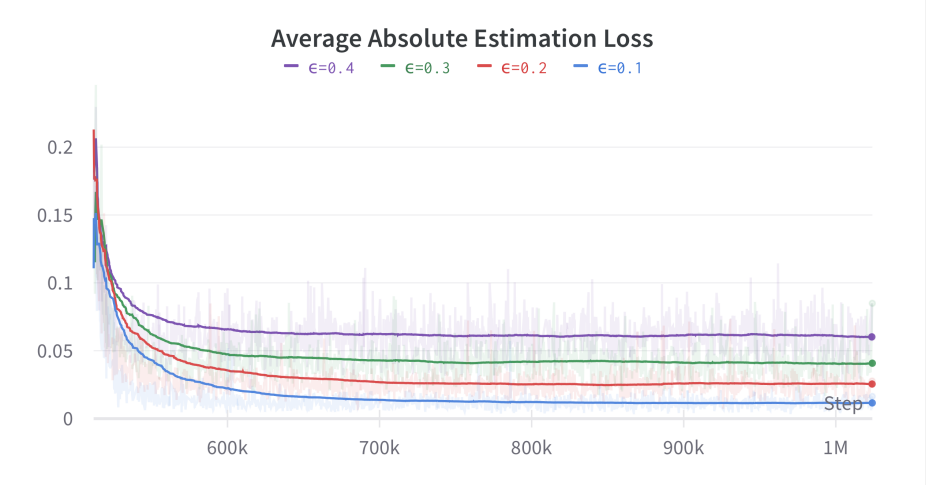

Figure 5: Loss curves of SARSA

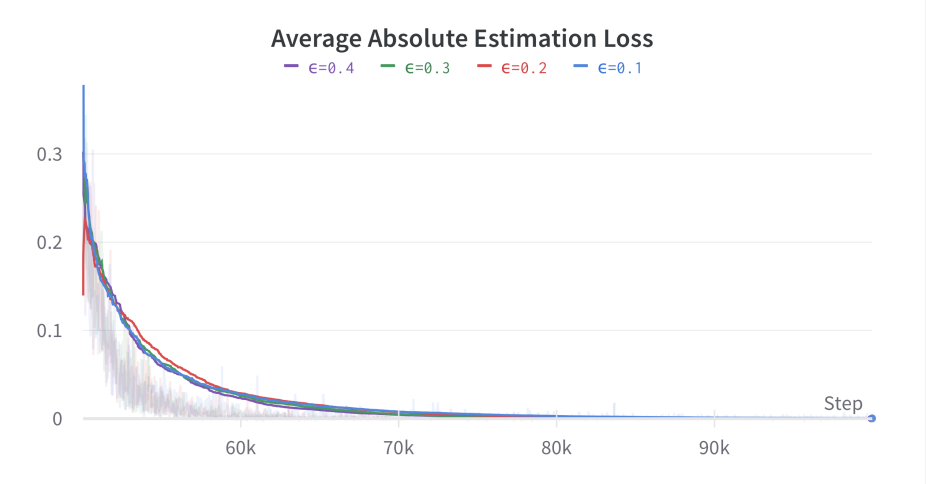

Figure 6: Loss curves of Q learning

In the loss curves, whatever the method is applied, the ranking order is completely opposite to that in the reward curves, as desired.

Note that the loss of MC and SARSA can not converge to 0, while Q learning can reduce loss to 0 almost. The possible reason is that MC and SARSA are on-policy while Q learning is off-policy. Specifically, the loss of MC is defined as:

$$\delta = G - Q_\pi(S, A). \tag{1}$$

Due to non-zero ϵ, the policy π can not reach to the real optimal policy, and it also introduces randomness in G, causing G and $q_\pi(s, a)$ to be unable to get close. Similarly, the loss of SARSA is defined as:

$$\delta = R + \gamma Q_\pi(S', A') - Q_\pi(S, A). \tag{2}$$

Again, the target $R + \gamma Q(S', A')$ and $Q(S, A)$ can not be close to each other as there is irreducible randomness in the policy (specifically, the next action A'). However, in Q learning, the loss is defined as:

$$\delta = R + \gamma \max_{A'} Q(S', A') - Q_\pi(S, A). \tag{3}$$

There is no randomness in the target policy as the max operator is deterministic. Once the Q values converge, then for any action A, we have $\max_{A'} Q(S', A') = Q_\pi(S, A)$. Therefore, the loss of Q learning can be reduce to near 0.

Q3. Discuss and plot reward and loss curves of Q learning under different discount factors, learning rates, update frequencies, and sample batch sizes

- Different Discount Factors

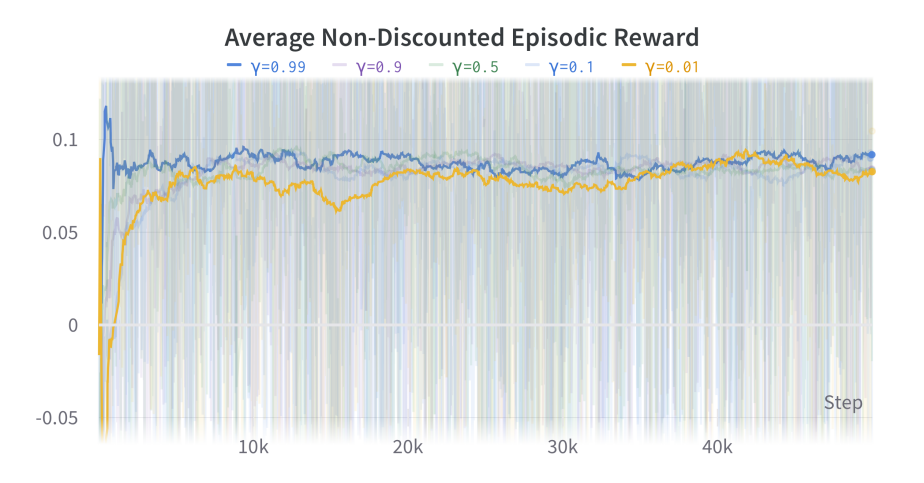

Figure 7: Reward curves of Q learning under different discount factors

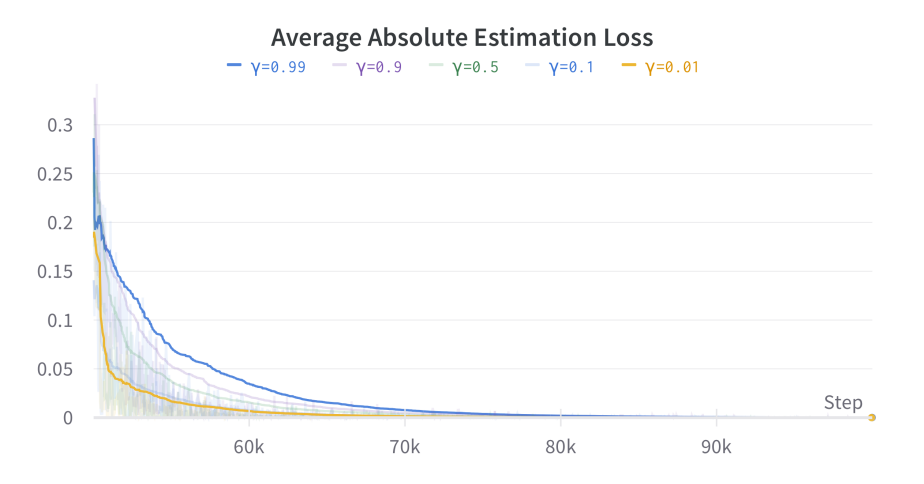

Figure 8: Loss curves of Q learning under different discount factors

Although in Fig. 8, it seems that small discount factor can make Q values converge quickly. However, because discount factor directly change the magnitude of each Q value, we can not make the previous conclusion directly.

In Fig 7, Q learning is not very sensitive to the value of the discount factor. But by comparing the two curves with deeper color; that is, the blue curve with $\gamma = 0.99$ and the yellow curve with $\gamma = 0.01$, we can still conclude that too small discount factor may make training harder as the agent becomes more short-sighted.

- Different Learning Rates

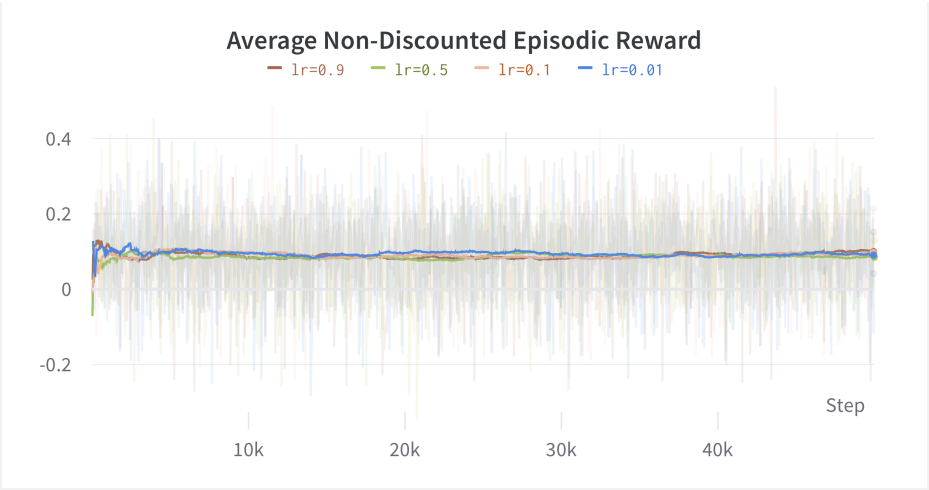

Figure 9: Reward curves of Q learning under different learning rates

Figure 10: Loss curves of Q learning under different learning rates

The effect of the value of learning rate can be observed in Fig. 10: the larger learning rate, the faster the Q values converges.

- Different Update Frequencies

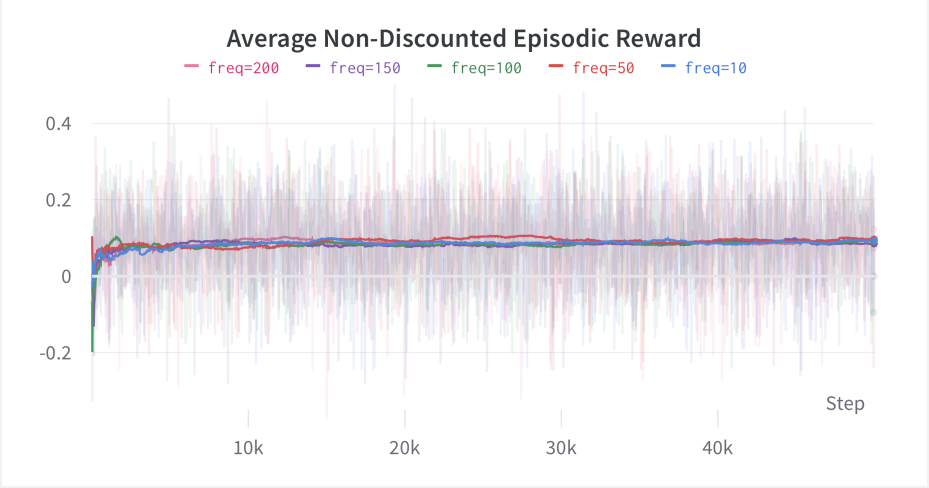

Figure 11: Reward curves of Q learning under different update frequencies

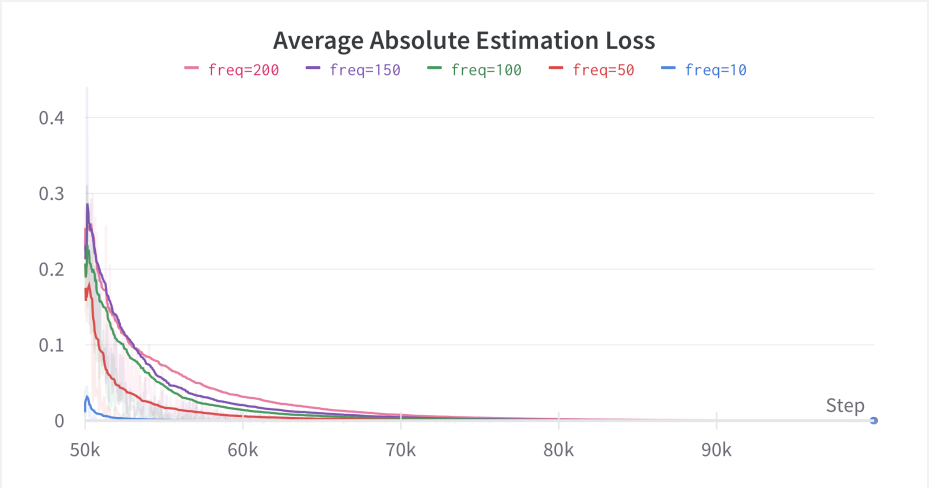

Figure 12: Loss curves of Q learning under different update frequencies

Small update frequency means that the Q values can be updated more immediate and more often, so it can boost the convergence. The loss curves prove this hypothesis.

- Different Sample Batch Sizes

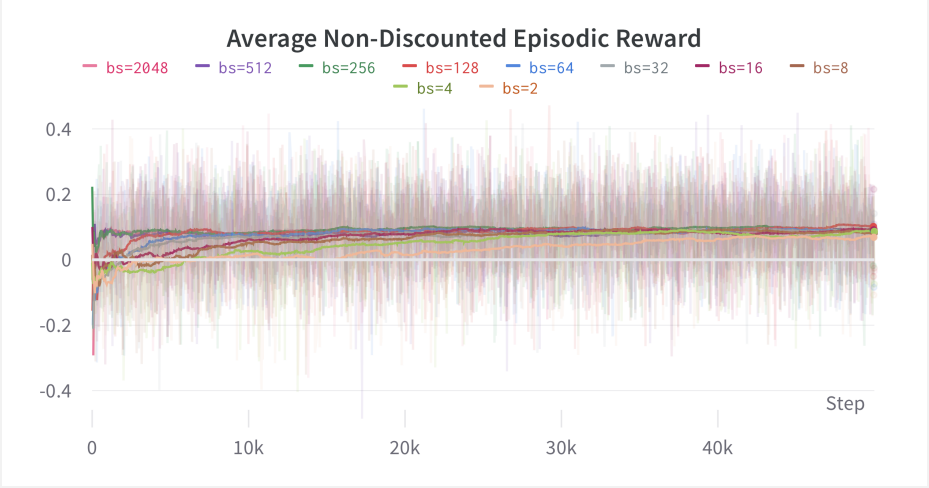

Figure 13: Reward curves of Q learning under different sample batch sizes

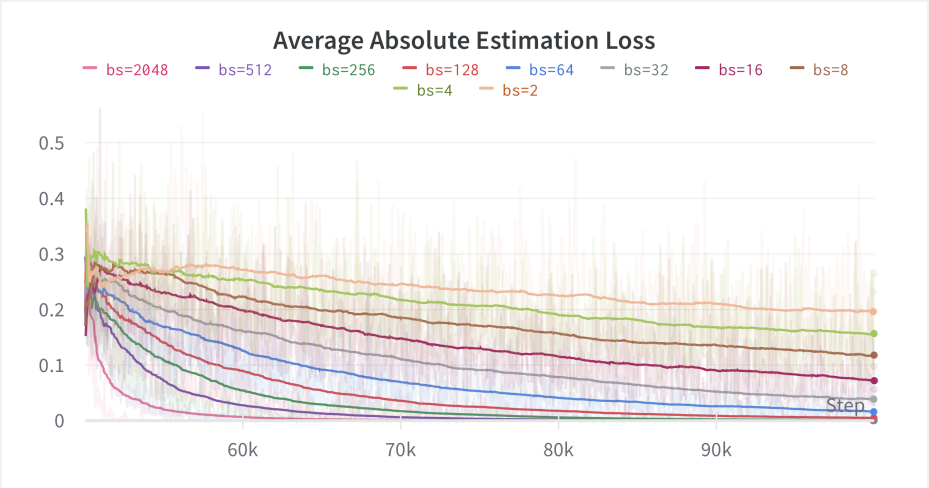

Figure 14: Loss curves of Q learning under different sample batch sizes

In the loss curves, it is quite obvious that large sample batch size can speed up convergence, as Q values can be updated more times in a single batch.