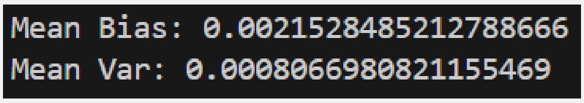
National Taiwan University 113-1 (Fall 2024)
Department of Electrical Engineering Reinforcement Learning
Name: Tzu-Tsen, Hsieh
ID: r11921072

Assignment 2 Report

Q1. Run Monte-Carlo prediction and TD(0) prediction for 50 seeds. Compare the resulting values with the GT values. Discuss the variance and bias.

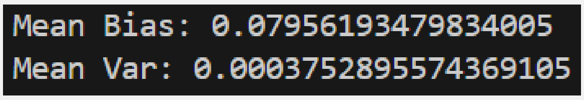
1. Comparison of Mean Bias and Variance Between MC and TD(0)

- **Result in Monte-Carlo (MC) and TD(0) Prediction :**



```
Mean Bias: 0.0021528485212788666
Mean Var: 0.0008066980821155469
```

Figure 1: MC prediction Mean Bias and Mean Variance



```
Mean Bias: 0.07956193479834005
Mean Var: 0.0003752895574369105
```

Figure 2: TD(0) prediction Mean Bias and Mean Variance

- **Monte-Carlo (MC) Prediction:**
  - **Mean Bias:** The mean bias of MC is very small (0.00215), indicating that MC prediction provides highly accurate estimates when compared to the ground truth (GT) values. This low bias is due to MC calculating the full return based on complete episodes, which leads to a more accurate estimate.
  - **Mean Variance:** MC's variance (0.00081) is higher than TD(0)'s. This is because MC relies on sampling full episodes, which introduces more variability depending on the outcomes and lengths of the episodes. Although MC has no bias, its high variance can lead to unstable predictions across different episodes, which is less stable for different episodes.
- **TD(0) Prediction:**
  - **Mean Bias:** TD(0) shows a larger bias (0.07956) compared to MC, indicating systematic deviation from the GT. This is expected because TD(0) uses bootstrapping, where it updates estimates based on other estimates rather than waiting for the full return. As a result, it introduces a bias, especially in early stages of learning when the estimates are not fully accurate.
  - **Mean Variance:** TD(0) has a lower variance (0.00038) compared to MC. Since TD(0) updates incrementally based on partial returns from each step, it stabilizes the estimates faster, leading to less variability in predictions. However, this results in the cost of higher bias.

2. Comparison of Bias and Variance Between MC and TD(0) in Each State

- **Compare the bias of each state in MC and TD(0)**

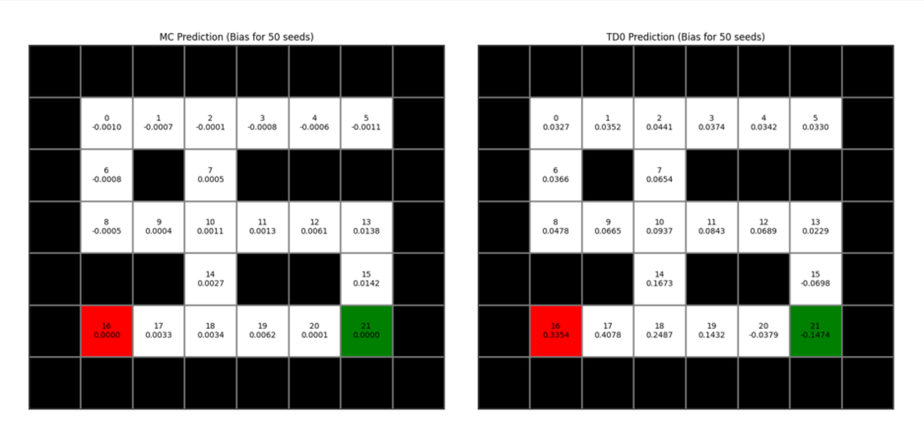


Figure 3: MC and TD(0) prediction Mean Bias for all states

When comparing the bias across different states, the bias of MC is significantly lower than that of TD(0), especially at the goal and trap positions where the bias is nearly zero. Additionally, the states further away from the goal and trap also exhibit smaller biases. This is because MC updates its estimates only after an episode ends, which can lead to larger deviations for states that are farther away. In contrast, although TD(0) shows larger biases across all states, these biases do not display any significant trends.

- **Compare the variance of each state in MC and TD(0)**

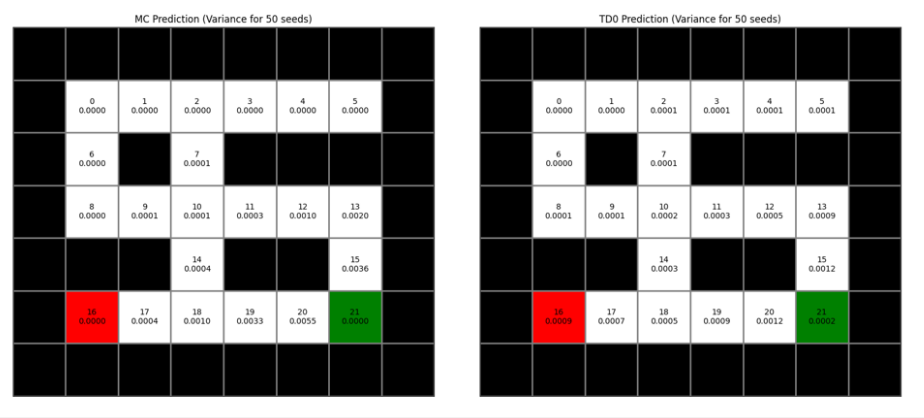


Figure 4: MC and TD(0) prediction Mean Variance for all states

From the variance, it is clear that as states move further away from state 0, their variance tends to increase. This is influenced by MC, rewards further away have lower weight, making them harder to estimate and leading to greater variance due to the higher uncertainty in future reward predictions. Since the overall return is accumulated from a series of rewards, states that are far from the maximum reward will see their variance increase with the uncertainty of the returns.

On the other hand, the TD(0) method uses immediate rewards and the predicted value of the next state to update the current state's value. This means that even when starting from a state far from the maximum reward, TD(0) can rely on the current immediate reward, thereby reducing its dependence on future rewards.

3. Conclusion:

Monte-Carlo (MC) prediction offers unbiased estimates but suffers from higher variance due to the use of full episode sampling. On the other hand, TD(0) prediction trades off lower variance for higher bias, as it relies on bootstrapping to update its estimates. The choice between the two methods depends on whether minimizing bias or variance is more critical for the specific application.

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

1. Monte Carlo (MC)

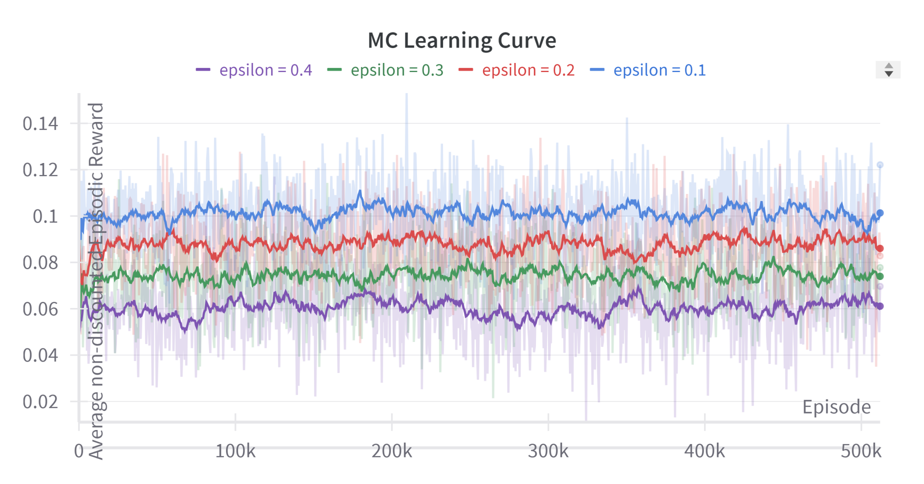


Figure 5: MC Learning Curve

- As the ϵ value increases, the reward per episode improves, but the average reward remains relatively flat, with no significant upward trend. This is because MC only updates the policy at the end of each episode. When ϵ increases, the probability of randomly selecting actions also increases, which benefits the agent's exploration of new states but leads to greater fluctuation in the learning curve.
- When ϵ is smaller, the agent's action choices are more greedy, meaning there are fewer opportunities to explore new information, resulting in a more stable learning curve. As the agent approaches the optimal policy, it converges to a higher reward.
- Since MC only updates the policy at the end of each episode, the learning curve is more stable within each episode.

2. SARSA

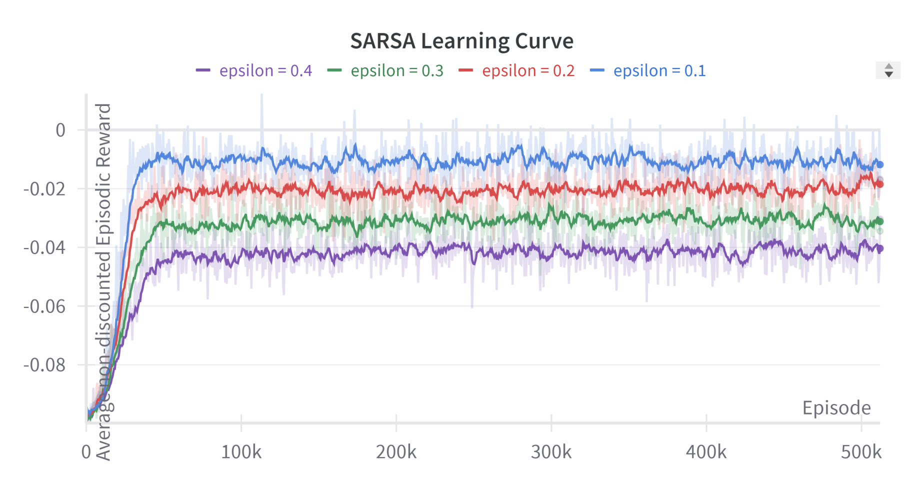


Figure 6: SARSA Learning Curve

- SARSA's learning curve rises rapidly in the early stages, and the smaller the ϵ value, the higher the average reward that can be achieved after convergence.
- When ϵ is smaller, SARSA relies more heavily on the Q-value to select actions, reducing the probability of random action selection. This helps the agent converge to a higher reward under more stable conditions. When ϵ is larger, even when the policy is close to optimal, random exploration may still affect the policy updates, making it harder to reach higher reward levels.
- Since SARSA updates the policy at each step, it can achieve better convergence compared to MC.

3. Q-Learning

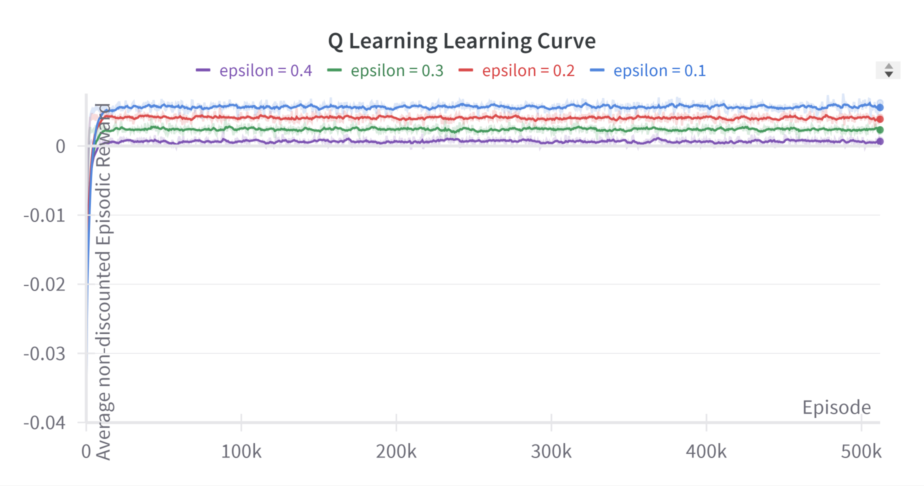


Figure 7: Q-Learning Learning Curve

- Q-Learning achieves convergence in fewer episodes and shows more stability after convergence.

- As ϵ increases, the converged reward tends to decrease, as Q-Learning is an off-policy method that tries to maximize future rewards regardless of the current policy.
- When ϵ is higher, increased exploration can hinder finding the optimal action. Conversely, when ϵ is lower, the agent focuses more on maximizing the current reward, resulting in higher convergence values.

Summary

As the epsilon value decreases, the agent achieves a higher level of convergence. This indicates the following:

- **Convergence Effectiveness Ranking**: Q-Learning > SARSA > MC.
- Q-Learning is often the preferred reinforcement learning algorithm due to its ability to learn stable and optimal policies in a shorter time. SARSA and MC may be more suitable for environments where stability is less critical.

Q3. Discuss and plot loss curves under ϵ values of $(0.1, 0.2, 0.3, 0.4)$ on MC, SARSA, and Q-Learning

1. Monte Carlo (MC)

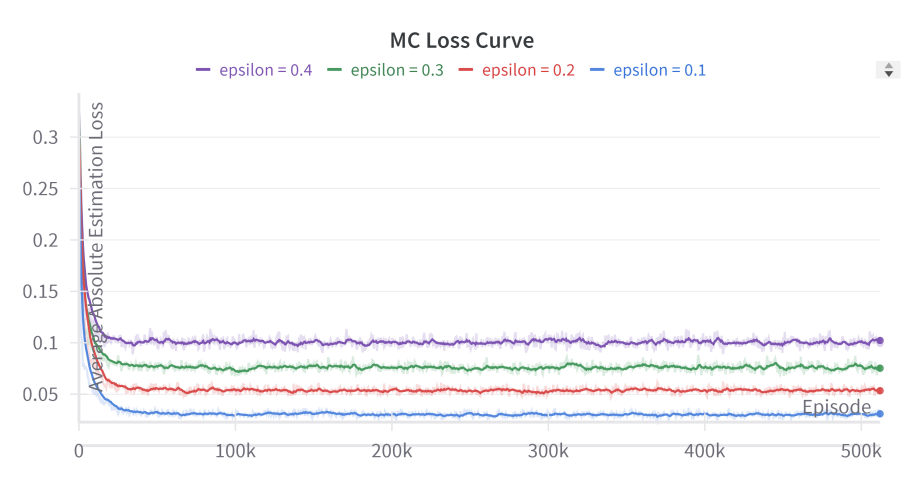


Figure 8: MC Loss Curve

The loss curve of MC shows that as ϵ decreases, the achievable convergence value also decreases. A higher ϵ allows the agent to explore more randomly, which can lead to larger prediction errors. In contrast, a smaller ϵn enables the loss to decrease more quickly and converge to a lower value, indicating that the agent is utilizing more known information to make action selections and reduce errors.

2. SARSA

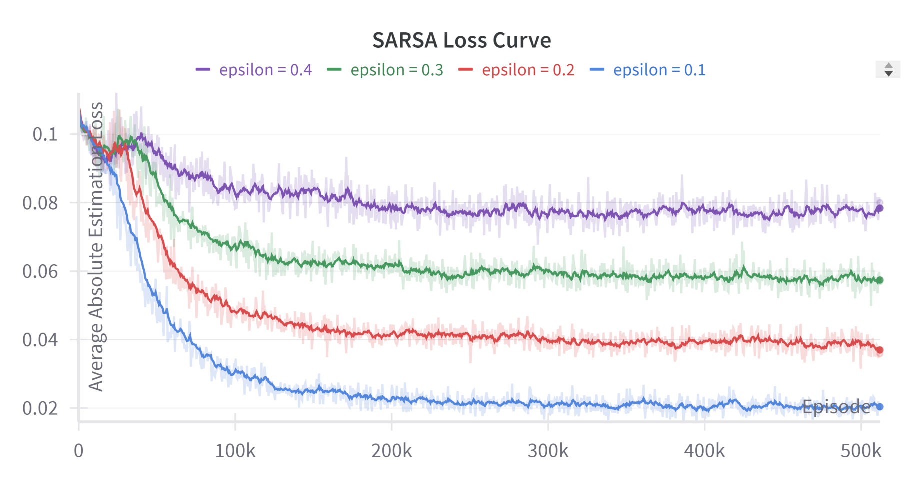


Figure 9: SARSA Loss Curve

During training, SARSA's loss is significantly lower than that of MC from the outset. This is because SARSA is based on an on-policy approach, which continually updates the policy. Consequently, it can update the value function more rapidly based on current experiences. Similar to MC, a higher ϵ leads to a higher loss convergence value, while a lower ϵ allows the loss to converge to a smaller value.

3. Q-Learning

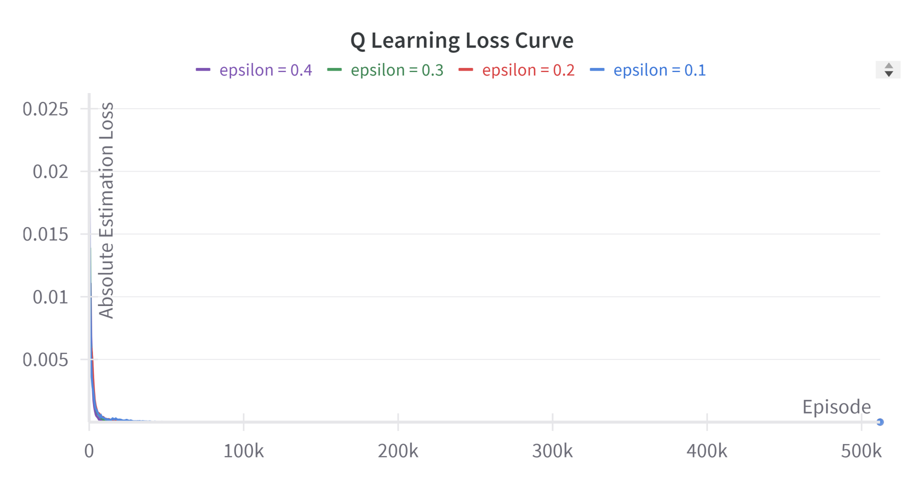


Figure 10: Q-Learning Loss Curve

Q-Learning achieves fast initial convergence, quickly approaching a loss near zero. The curves for different ϵ values are nearly overlapping, indicating very low convergence values. As an off-policy method, Q-Learning is less affected by the value of ϵ, enabling it to rapidly reach the optimal solution.

Summary

Overall, a smaller ϵ can lead to higher convergence values. In terms of convergence effectiveness and the ability to achieve lower loss values, the order is as follows:

$$\text{Q-Learning} > \text{SARSA} > \text{Monte Carlo (MC)} \tag{1}$$