Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 4

Solution by : Amirhossein Asadi

Q1

Which of the following statements is true about Monte-Carlo (MC) and Temporal Difference (TD) learning?

A) MC methods require complete episodes to update value estimates, while TD methods can update estimates after every step.
B) TD methods have higher variance in their updates compared to MC methods.
C) MC methods use bootstrapping, while TD methods do not.
D) TD methods cannot learn in continuing (non-terminating) environments.

Correct Answer: A

In MC methods, the value estimates are updated only after a complete episode has been observed, meaning that MC methods require episodes to terminate before the update can occur. On the other hand, TD methods can update value estimates after every step, using bootstrapping, which allows TD methods to update more frequently during the process.

Q2

Which of the following statements about bootstrapping and sampling in reinforcement learning is false?

A) Monte-Carlo methods do not use bootstrapping but do use sampling.
B) Temporal Difference methods use both bootstrapping and sampling.
C) Dynamic Programming methods use bootstrapping but do not use sampling.
D) Monte-Carlo methods use bootstrapping to estimate the value function.

Correct Answer: D

MC methods do not use bootstrapping, as they rely on complete episodes to estimate the value function. They use sampling to estimate returns directly from the observed episodes. In contrast, TD methods use bootstrapping, which involves updating value estimates based on other value estimates. Dynamic Programming methods also use bootstrapping, but they do not rely on sampling.

Q3

You are implementing an epsilon-greedy policy for a reinforcement learning agent. The agent has 4 possible actions in a given state, and the current Q-values for those actions are:

$$\text{Action 1: } Q = 5, \quad \text{Action 2: } Q = 3, \quad \text{Action 3: } Q = 7, \quad \text{Action 4: } Q = 2$$

If epsilon = 0.2, what is the probability of selecting Action 3 (the action with the highest Q-value)?

A) 0.8
B) 0.85
C) 0.9
D) 0.95

Correct Answer: A

In an epsilon-greedy policy, the agent selects the best action (highest Q-value) with probability $1-\epsilon$ and a random action with probability ϵ. Since $\epsilon = 0.2$, the probability of selecting the best action (Action 3 with the highest Q-value) is:

$$\text{Probability of selecting Action 3} = 1 - \epsilon = 1 - 0.2 = 0.8$$

Therefore, the probability of selecting Action 3 is **0.8**.

Q4

Consider a policy π that always selects the action with the highest Q-value in a given state. If you use policy iteration to improve this policy, what is the most likely outcome after one iteration of policy improvement?

A) The policy will remain unchanged because it is already optimal.
B) The policy will change to explore more actions, even if they have lower Q-values.
C) The policy will improve by selecting actions that lead to higher long-term rewards.
D) The policy will become stochastic, selecting actions randomly.

Correct Answer: C

In policy iteration, the policy is evaluated and then improved. If the initial policy already selects the action with the highest Q-value (greedy policy), policy improvement will refine the policy to ensure that it selects actions that lead to higher long-term rewards, rather than just immediate rewards. Therefore, after one iteration, the policy will improve by focusing on long-term benefits, not just the immediate highest Q-values.

Q5

You are training a reinforcement learning agent using an epsilon-greedy policy. Initially, the agent explores a lot epsilon, but over time, you reduce epsilon to encourage more exploitation. However, you notice that the agent's performance plateaus and does not improve further. What could be the reason for this?

A) The agent is not exploring enough to discover better actions.
B) The agent is exploiting too much and getting stuck in suboptimal policies.
C) The learning rate alpha is too high.
D) The discount factor gamma is too low.

Correct Answer: B

If the agent's performance plateaus, it could be because the agent is exploiting the current policy too much and getting stuck in suboptimal solutions. Reducing exploration too much can prevent the agent from discovering better actions that may lead to higher long-term rewards. To improve performance, the agent needs a balance between exploration and exploitation.

Q6

Imagine an agent exploring a grid world. It follows an epsilon-greedy policy and learns the value of different state-action pairs using Monte-Carlo control. If the agent has a low learning rate alpha and high exploration probability epsilon, how would this affect the learning process? What challenges might arise?

A low learning rate (α) causes slow updates, making learning inefficient. High exploration (ϵ) leads to excessive random actions, preventing the agent from exploiting its current knowledge. As a result, the agent may take longer to converge to an optimal policy, and learning may become inefficient due to the balance between exploration and slow updates.

Q7

One of the main problems with Monte-Carlo (MC) methods is the high variance in the estimates. If there is a dominant action (an action that is much more likely to be chosen than others), does this exacerbate the variance problem in MC?

Yes, having a dominant action can exacerbate the variance problem in Monte-Carlo methods. If one action is much more likely to be chosen, the agent will sample more from the episodes that include this action, which can result in high variance in the value estimates. This happens because the estimate of the value function will depend heavily on the returns from the dominant action, leading to larger fluctuations in the estimated values. This is a major source of the high variance problem in MC methods.