Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 7

Solution by : Amirhossein Asadi

Q1

In Actor-Critic methods, which option correctly describes the role of the Critic?

A) Updating the policy using gradients provided by the Actor.
B) Selecting actions based on the current policy.
C) Estimating the value function to evaluate the policy.
D) Calculating future rewards for policy updates.

Correct Answer: C

The Critic in Actor-Critic methods estimates the value function, which is used to evaluate the performance of the current policy. It provides feedback to the Actor by estimating the expected return (value) of the current state or state-action pair, allowing the Actor to adjust the policy accordingly.

Q2

Explain why the discount factor γ is necessary in RL. What happens if $\gamma = 1$ or $\gamma = 0$? In the game of chess, why would a high discount factor be beneficial?

The discount factor γ is necessary in reinforcement learning (RL) because it determines the importance of future rewards compared to immediate rewards. A lower discount factor (γ) makes the agent focus more on short-term rewards, while a higher discount factor (γ) places more importance on long-term rewards.
- If $\gamma = 0$, the agent will only care about immediate rewards, essentially behaving myopically.
- If $\gamma = 1$, the agent considers future rewards as important as immediate rewards, leading it to plan for long-term outcomes.
In the game of chess, a high discount factor is beneficial because the game involves many moves before the final outcome (win or loss) is realized. Therefore, it encourages the agent to plan ahead, considering the long-term consequences of its moves rather than just focusing on immediate gains.

Q3

What is the main difference between Batch Actor-Critic and Online Actor-Critic?

A) Batch Actor-Critic uses a batch of trajectories, while Online Actor-Critic updates after each step.
B) Batch Actor-Critic uses a batch of trajectories, while Online Actor-Critic uses a fixed Baseline.
C) Batch Actor-Critic uses a fixed Baseline, while Online Actor-Critic uses a batch of trajectories.
D) Batch Actor-Critic uses a fixed Baseline, while Online Actor-Critic updates after each step.

Correct Answer: A

The main difference between Batch Actor-Critic and Online Actor-Critic is how the updates are made. In Batch Actor-Critic, a batch of trajectories is collected and then used to perform updates, while in Online Actor-Critic, updates are made after each step of the process, continuously refining the policy.

Q4

Which of the following statements accurately describes the nature of Actor-Critic methods?

A) They are inherently off-policy, learning the value of a target policy while following a different behavior policy.
B) They are inherently on-policy, learning the value of the policy that is currently being executed.

Correct Answer: B

Actor-Critic methods are inherently on-policy. They learn the value of the policy that is currently being executed by the agent. In Actor-Critic methods, the Actor makes decisions based on the current policy, and the Critic estimates the value of these decisions. Unlike off-policy methods, which use data from a different policy, Actor-Critic methods update the policy based on the actions the agent actually takes during training.

Q5

If a baseline in a policy gradient method is a function of the policy parameters (τ), what is the likely consequence?

A) The gradient estimation will have reduced variance without introducing bias.
B) The learning algorithm will converge faster.
C) The gradient estimation may become biased.
D) There will be no effect on the gradient estimation.

Correct Answer: C

When the baseline is a function of the policy parameters, it can introduce bias into the gradient estimation. While the baseline helps reduce variance, using the policy parameters in the baseline might lead to a biased estimate. This happens because the baseline itself may depend on the same parameters that we are trying to optimize, potentially distorting the gradient.

Q6

What potential issue arises if trajectories are defined using only states without corresponding actions?

A) The agent cannot accurately estimate the state transition probabilities.
B) The agent cannot properly evaluate the expected cumulative reward.
C) The agent cannot determine the optimal policy.
D) The agent cannot effectively learn the state-action value function.

Correct Answer: D

In RL, the state-action value function $Q(s, a)$ represents the expected cumulative reward given a state-action pair. If trajectories are only defined using states without corresponding actions, the agent cannot properly associate actions with rewards, which makes it impossible to learn the state-action value function effectively. As a result, the agent will not be able to accurately estimate the value of taking specific actions in particular states, which is crucial for determining an optimal policy.

Q7

Explain how PPO addresses data instability caused by improper learning rates in earlier RL methods.

PPO addresses data instability, which is common in earlier RL methods due to improper learning rates, by introducing a clipping mechanism in its objective function. This clipping helps prevent overly large updates to the policy, which can destabilize the learning process.
In traditional RL methods, large policy updates could cause large deviations from the optimal policy, particularly if the learning rate is too high. This issue could lead to overfitting or underfitting and result in poor performance. PPO mitigates this by clipping the probability ratio between the new and old policies, ensuring that the updates remain within a safe region. By controlling the magnitude of policy updates, PPO stabilizes the training process, ensuring more consistent and reliable improvements in the agent's policy.
Additionally, PPO uses an adaptive learning rate and maintains a balance between exploration and exploitation through its clipped objective. This mechanism ensures that the policy is updated in a controlled manner, promoting stability in the learning process while still making progress toward the optimal policy.