Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 3

Solution by : Amirhossein Asadi

Q1

How does policy improvement work in Policy Iteration?

A) By randomly selecting actions to explore the environment
B) By computing the temporal difference error for each state-action pair
C) By updating the transition probabilities to match the current policy
D) By selecting actions that maximize the Q-value for each state

Correct Answer: D

In Policy Iteration, policy improvement works by selecting actions that maximize the Q-value for each state. This ensures that the policy is improved by selecting the best possible actions, according to the current value function. In each iteration, the policy is updated to choose the actions that lead to the highest expected return, leading to a better policy over time. This process continues until convergence, when the policy no longer improves.

Q2

Why does Policy Iteration generally converge faster than Value Iteration?

because it alternates between policy evaluation and improvement, which are more efficient than the repeated value updates in Value Iteration. In each iteration, Policy Iteration makes more significant progress by updating the policy in one step, while Value Iteration requires many updates to the value function, making it slower.

Q3

If the discount factor γ were set to 1, how would the computed values of $V^*(s)$ be affected?

A) They would remain the same
B) They would increase since future rewards are not discounted
C) They would decrease as earlier rewards become more important
D) The values would become negative

Correct Answer: B

If the discount factor γ is set to 1, future rewards will not be discounted, meaning the agent will place equal importance on both immediate and future rewards. As a result, the computed values of $V^*(s)$ will increase because the agent will now consider the full, unmodified sum of future rewards when calculating the value of a state.

Q4

What is the difference between $\pi_\theta(a_t|o_t)$ and $\pi_\theta(a_t|s_t)$?

- $\pi_\theta(a_t|o_t)$ represents the policy based on the observation o_t, which typically refers to the partial or noisy information the agent has about the environment.
- $\pi_\theta(a_t|s_t)$, on the other hand, represents the policy based on the full state s_t, which is the complete information available about the environment at time t.
In general, $\pi_\theta(a_t|s_t)$ is used when the agent has full access to the environment's state, whereas $\pi_\theta(a_t|o_t)$ is used in partially observable environments.

Q5

A higher discount factor γ (closer to 1) makes the agent focus more on short-term rewards rather than long-term rewards.

A) True
B) False

Correct Answer: B

Q6

What is the primary goal of policy evaluation in reinforcement learning?

A) To find the optimal policy that maximizes long-term rewards
B) To estimate the value function for a given policy
C) To update the policy using a greedy improvement step
D) To explore the environment and collect new data

Correct Answer: B

The primary goal of policy evaluation in RL is to estimate the value function for a given policy. This value function helps evaluate how good a policy is, by estimating the expected cumulative rewards from each state under that policy. Policy evaluation is typically performed iteratively to approximate the value function, which is then used to improve the policy.