Reinforcement Learning

Q1: What are actor-critic methods?

Actor-critic methods combine policy gradients with value estimation. The actor chooses actions based on the current state, while the critic estimates how good that action was. This helps reduce the high variance seen in pure policy gradient methods like REINFORCE.

Q2: What is the actor?

The actor represents the policy $\pi_\theta(a|s)$. In practice, it is usually a neural network that outputs action probabilities (discrete case) or parameters of a distribution (continuous control). It is updated to increase expected reward using feedback from the critic.

Q3: What is the critic?

The critic estimates a value function like $V(s)$ or $Q(s, a)$ using TD learning. It provides a baseline or advantage estimate that guides the actor updates, making training more sample-efficient and stable.

Q4: How does actor-critic differ from REINFORCE with baseline?

REINFORCE with baseline uses full episode returns, so updates only happen after an episode ends. Actor-critic uses bootstrapped TD targets such as $r + \gamma V(s')$, allowing learning during the episode. This usually makes it faster and less noisy.

Q5: What is a major advantage of a TD-based critic?

TD methods have lower variance because they do not rely entirely on long returns. Even though TD adds some bias, the updates are faster and more stable, leading to better learning in practice.