

**CS3018 Reinforcement Learning**fast-logo

**Quiz 6**

Date: 8th May, 2025

1. Which of the following best represents the loss function used in Deep Q-Learning?

A) (r + γ max\_a Q(s', a') - Q(s, a))^2

B) - log π(a|s) R

C) 0.5 \* (V(s) - R)^2

D) γ (Q(s, a) - V(s))

1. In Gradient Monte Carlo methods, what is the main learning signal?

A) Bootstrapped estimates

B) Sampled cumulative return from episodes

C) Action probabilities from a critic

D) Target network estimates

1. Which of the following is a limitation of Gradient Monte Carlo?

A) Requires function approximation

B) Only works in continuous spaces

C) Has high variance due to full returns

D) Cannot optimize stochastic policies

1. What is the key idea behind policy gradient methods like REINFORCE?

A) Update Q-values directly

B) Use greedy action selection

C) Optimize policy parameters by maximizing expected return

D) Estimate the reward-to-go using TD(0)

1. In semi-gradient TD(0), why is the term 'semi-gradient' used?

A) It ignores the gradient of the target r + γ v̂(s')

B) It does not use function approximation

C) It computes a full return for the value

D) It calculates the exact gradient of the true value function

1. Which of the following is true for semi-gradient TD(0)?

A) It uses Monte Carlo sampling

B) It bootstraps using the value of the next state

C) It waits for episode termination to update

D) It requires the full return G\_t

1. Which of the following is true about the difference between Deep Q-Learning and Policy Gradient methods?

A) Deep Q-Learning directly optimizes the policy  
 B) Policy Gradient methods require discrete action spaces  
 C) Deep Q-Learning is value-based; Policy Gradient is policy-based  
 D) Policy Gradient always converges faster than Deep Q-Learning

1. In terms of action selection, how do Deep Q-Learning and Policy Gradient differ?

A) Both use epsilon-greedy exclusively  
 B) Deep Q-Learning outputs action probabilities; Policy Gradient outputs Q-values  
 C) Deep Q-Learning selects actions using Q-values; Policy Gradient samples from a probability distribution  
 D) Both methods select actions greedily