Started on	Friday, 31 January 2025, 1:10 PM
State	Finished
Completed on	Friday, 31 January 2025, 1:22 PM
Time taken	12 mins 24 secs
Grade	**100.00** out of 100.00

Question **1**

Correct

Mark 4.00 out of 4.00

Why are Upper Confidence Bound (UCB) methods effective for action selection?

a. They eliminate the need for exploration.

b. They prioritize actions with high uncertainty. ✔

c. Using stochastic rewards for initialization

d. Setting probabilities equal for all actions.

e. Avoiding suboptimal actions entirely.

Question **2**

Correct

Mark 4.00 out of 4.00

In Monte Carlo Reinforcement Learning, which strategy can be used to ensure that all state-action pairs are sampled (visited) eventually?

a. Using purely deterministic policies

b. Restricting the agent to a subset of actions

c. Using an infinitely high discount factor

d. Relying on bootstrapping from immediate rewards

e. Exploring starts, where each episode begins from every possible state-action pair with nonzero probability ✔

Question **3**

Correct

Mark 4.00 out of 4.00

What is the main difference between Q-Learning and SARSA?

a. Q-Learning does not use a discount factor, while SARSA does.

b. Q-Learning is an on-policy method, while SARSA is off-policy.

c. Q-Learning's update uses the greedy action $(\max_{a'} Q)$ for the next state, whereas SARSA uses the action actually taken by the current policy. ✔

d. SARSA requires knowledge of the transition probabilities.

e. SARSA converges faster than Q-Learning in all cases

Question **4**

Correct

Mark 4.00 out of 4.00

Which of the following is true regarding Monte Carlo methods and infinite episodes?

a. they assume episodes are infinite and never update the value function.

b. They cannot handle infinite-horizon problems at all.

c. They ignore discounting entirely in infinite episodes.

d. They require episodes to terminate eventually, or use a concept like continuing tasks with average reward methods. ✔

e. They rely on partial returns mid-episode for updates.

Question **5**

Correct

Mark 4.00 out of 4.00

Why does DQN use a separate target network?

a. To generate randomized actions for exploration

b. To eliminate the need for discounting future rewards

c. To stabilize Q-value updates by keeping target estimates fixed for a while ✔

d. To independently learn a model of the transition probabilities

e. To convert a continuous action space into a discrete one

Question **6**

Correct

Mark 4.00 out of 4.00

Why is Q-Learning considered an off-policy method?

a. It never explores and only exploits the best action.

b. It requires running a secondary "behavior policy" to gather experience but updates the values as if it followed a greedy policy. ✔

c. It abandons temporal difference learning.

d. It does not require a replay buffer or any historical data.

e. It updates its estimates based on Monte Carlo returns only.

Question **7**

Correct

Mark 4.00 out of 4.00

Which technique is commonly used in DQN to stabilize learning?

a. Dynamically modifying the environment's reward signals in every step.

b. Completely ignoring experience replay so as not to overfit.

c. Maintaining a target network that is updated slowly compared to the main Q-network. ✔

d. Using purely on-policy updates with SARSA.

e. Removing the discount factor to avoid infinite returns.

Question **8**

Correct

Mark 4.00 out of 4.00

In off-policy methods, what is the main purpose of importance sampling?

a. To convert on-policy data into a model-based approach

b. To correct for the mismatch between the behavior policy (used to generate data) and the target policy (used to evaluate/improve) ✔

c. To ensure that the discount factor can be changed dynamically

d. To reduce the variance of Monte Carlo estimates by ignoring certain transitions

e. To enable purely deterministic updates in a continuous action space

Question **9**

Correct

Mark 4.00 out of 4.00

TD methods differ from Dynamic Programming (DP) primarily because TD methods

a. Always converge faster than DP methods

b. Do not use the concept of value functions

c. Require a perfect model of the environment's transitions

d. Can learn directly from raw experience without knowing transition probabilities ✔

e. Only work in deterministic environments

Question **10**

Correct

Mark 4.00 out of 4.00

What does the term “Markov property” signify in Markov Decision Processes (MDPs)?

a. The policy is deterministic for every state

b. The environment is stationary

c. The transition probabilities are stochastic

d. The future depends only on the present state ✔

e. The rewards are discounted exponentially

Question **11**

Correct

Mark 4.00 out of 4.00

Q-Learning and SARSA both estimate Q-values, but Q-Learning is considered off-policy because

a. The actions used in the bootstrapped target are always taken from the same policy that generates behavior

b. It never uses \epsilon-greedy exploration

c. It ignores the discount factor in updates

d. It updates using a greedy action for the next state, not necessarily the one followed by the agent during data collection ✔

e. It uses the same policy for both exploration and evaluation

Question **12**

Correct

Mark 4.00 out of 4.00

In Monte Carlo methods, which of the following is a typical requirement for estimating value functions?

a. Episodes can be truncated at any time and the incomplete return is used directly.

b. The method only works with deterministic environments.

c. Each episode must be guaranteed to be infinite.

d. No environment interaction is needed; it relies on purely analytical solutions.

e. The method uses the average of the returns observed for each state (or state-action pair) across many episodes. ✔

Question **13**

Correct

Mark 4.00 out of 4.00

In DQN, what is the key purpose of the Experience Replay Buffer?

a. It amplifies the most recent transition repeatedly to speed up learning

b. It stores past experiences and samples them randomly to break correlation in sequential data ✔

c. It ensures that all experiences are used exactly once to avoid correlation

d. It only stores states without actions or rewards

e. It replaces the need for a target network

Question **14**

Correct

Mark 4.00 out of 4.00

In SARSA (a TD control method), the update rule for the state-action value function Q typically includes which of the following terms?

a. \max_{a' Q(s',a')

b. The target action chosen by an off-policy method

c. A direct model of state transitions

d. A value function that depends on no discount factor

e. The next action actually taken by the current policy ✔

Question **15**

Correct

Mark 4.00 out of 4.00

How does the n-step TD approach differ from TD(0)?

a. TD(0) is only used for deterministic policies, while n-step TD is for stochastic policies.

b. TD(0) is an on-policy method while n-step TD is off-policy.

c. n-step TD randomly selects how many steps to wait before an update.

d. n-step TD updates only at the end of the episode, just like Monte Carlo.

e. n-step TD uses longer traces of rewards and states before performing a single update, rather than a one-step lookahead. ✔

Question **16**

Correct

Mark 4.00 out of 4.00

In a standard DQN, the neural network typically

a. Outputs a single Q-value for each possible action in the environment ✔

b. Stores data in a tabular format without hidden layers

c. Receives the next action as part of the input

d. Automatically splits the environment into separate tasks

e. Directly predicts the best action without any Q-value

Question **17**

Correct

Mark 4.00 out of 4.00

What is one main reason why a plain DQN might struggle in very high-dimensional continuous action spaces?

a. DQN is not a function approximator

b. Experience Replay is impossible to maintain for large observation spaces

c. Continuous spaces do not allow for discounting of future rewards

d. DQN outputs Q-values for discrete actions, and enumerating a continuous space is infeasible ✔

e. The \epsilon-greedy strategy can only handle continuous actions

Question **18**

Correct

Mark 4.00 out of 4.00

Which key feature distinguishes Temporal Difference (TD) learning from Monte Carlo methods?

a. TD uses bootstrapping from current estimates rather than waiting for the final outcome. ✔

b. TD requires access to the full model of the environment's transition probabilities.

c. TD is only applicable to deterministic policies.

d. TD waits until the end of an episode to update value estimates.

e. TD cannot update its estimates online.

Question **19**

Correct

Mark 4.00 out of 4.00

In the context of TD methods, "bootstrapping" refers to which concept?

a. Taking random actions in the environment to initialize the replay buffer

b. Bypassing the need for an explicit Q-value function

c. Updating value estimates based partly on other learned estimates rather than exclusively on actual returns ✔

d. Learning only from complete returns collected at the end of each episode

e. Combining multiple policies simultaneously

Question **20**

Correct

Mark 4.00 out of 4.00

In DQN for discrete actions, how does the agent select the best action after the network outputs Q-values?

a. It chooses the action with the highest Q-value. ✔

b. The network outputs a probability distribution, from which the action is sampled.

c. It always picks actions in a round-robin manner.

d. It uses an actor network to select continuous actions.

e. It picks the action with the smallest Q-value to minimize cost.

Question **21**

Correct

Mark 4.00 out of 4.00

In a standard DQN architecture, which statement is true about how the neural network is used?

a. The network directly outputs the optimal action without any Q-value estimation.

b. The network only estimates the next state's reward, ignoring future states.

c. The network takes the state as input and outputs a single Q-value, forcing you to run it multiple times.

d. The network takes the state as input and outputs Q-values for all possible discrete actions. ✔

e. The neural network outputs the policy probabilities for each action.

Question **22**

Correct

Mark 4.00 out of 4.00

In SARSA, the next action used in the update is

a. The greedy action that maximizes the Q-value in the next state

b. Irrelevant, because SARSA does not require a next action

c. Provided by a known model of the environment

d. Determined by a different policy than the one being evaluated

e. The action actually chosen by the current (often \epsilon-greedy) policy in the next time step ✔

Question **23**

Correct

Mark 4.00 out of 4.00

Which statement best describes the Monte Carlo approach in Reinforcement Learning?

a. It updates value estimates based on a single step lookahead.

b. It uses complete episodes to estimate returns and updates the value function only at the end of each episode. ✔

c. It is fundamentally incompatible with policy evaluation.

d. It estimates value functions purely by bootstrapping from existing estimates.

e. It performs value updates after every state transition without waiting for episode completion.

Question **24**

Correct

Mark 4.00 out of 4.00

Double Q-Learning was introduced primarily to address which issue?

a. Handling continuous actions without an actor-critic method

b. Overestimation bias in Q-value updates due to using \max over the same Q function ✔

c. The instability caused by batch updates in Q-Learning
E. The inability of Q-Learning to handle function approximation

d. The inability of Q-Learning to handle function approximation

e. Lack of exploration in Q-Learning

Question **25**

Correct

Mark 4.00 out of 4.00

What does the Bellman Equation represent in Reinforcement Learning?

a. The probability of state transitions

b. The best policy derivation for multi-agent systems

c. The expected long-term reward for a policy

d. The immediate reward plus discounted future rewards ✔

e. The difference between predicted and actual rewards

◄ Lecture 7

Jump to...

Lecture 8 ►