

Congratulations! You passed!

TO PASS 80% or higher

Keep Learning

GRADE

100%

Graded Quiz

LATEST SUBMISSION GRADE

100%

1. Which approach ensures continual exploration? (Select all that apply)

1 / 1 point

 Exploring starts**Correct**

Correct! Exploring starts guarantee that all state-action pairs are visited an infinite number of times in the limit of an infinite number of episodes.

 On-policy learning with a deterministic policy On-policy learning with an ϵ -soft policy**Correct**Correct! ϵ -soft policies assign non-zero probabilities to all state-action pairs. Off-Policy learning with an ϵ -soft behavior policy and a deterministic target policyCorrect! ϵ -soft policies have non-zero probabilities for all actions in all states. The behavior policy is used to generate samples and should be exploratory. Off-Policy learning with an ϵ -soft target policy and a deterministic behavior policy

2. When can Monte Carlo methods, as defined in the course, be applied? (Select all that apply)

1 / 1 point

 When the problem is continuing and there are sequences of states, actions, and rewards When the problem is continuing and there is a model that produces samples of the next state and reward When the problem is episodic and there are sequences of states, actions, and rewards**Correct**

Correct! Well-defined returns are available in episodic tasks.

 When the problem is episodic and there is a model that produces samples of the next state and reward**Correct**

Correct! Well-defined returns are available in episodic tasks.

3. Which of the following learning settings are examples of off-policy learning? (Select all that apply)

1 / 1 point

 Learning about multiple policies simultaneously while following a single behavior policy Learning the optimal policy while continuing to explore**Correct**

Correct! An off-policy method with an exploratory behavior policy can assure continual exploration.

 Learning from data generated by a human expert**Correct**

Correct! Applications of off-policy learning include learning from data generated by a non-learning agent or human expert. The policy that is being learned (the target policy) can be different from the human expert's policy (the behavior policy).

4. If a trajectory starts at time
- t
- and ends at time
- T
- , what is its relative probability under the target policy
- π
- and the behavior policy
- b
- ?

1 / 1 point

 $\prod_{k=t}^{T-1} \frac{\pi(A_k | S_k)}{b(A_k | S_k)}$ $\sum_{k=t}^{T-1} \frac{\pi(A_k | S_k)}{b(A_k | S_k)}$ $\frac{\pi(A_{T-1} | S_{T-1})}{b(A_{T-1} | S_{T-1})}$ $\frac{\pi(A_t | S_t)}{b(A_t | S_t)}$ **Correct**

Correct! Monte Carlo methods in Reinforcement Learning sample and average returns much like bandit methods sample and average rewards.

5. When is it possible to determine a policy that is greedy with respect to the value functions
- v_π, q_π
- for the policy
- π
- ? (Select all that apply)

1 / 1 point

 When state values v_π and a model are available**Correct**

Correct! With state values and a model, one can look ahead one step and see which action leads to the best combination of reward and next state.

 When state values v_π are available but no model is available. When action values q_π and a model are available**Correct**

Correct! Action values are sufficient for choosing the best action in each state.

 When action values q_π are available but no model is available.**Correct**

Correct! Action values are sufficient for choosing the best action in each state.

6. Monte Carlo methods in Reinforcement Learning work by...

1 / 1 point

 Averaging sample rewards Planning with a model of the environment Performing choose through the state set**Correct**

Correct! Monte Carlo methods in Reinforcement Learning sample and average returns much like bandit methods sample and average rewards.

7. Which of the following is a requirement for using Monte Carlo policy evaluation with a behavior policy
- b
- for a target policy
- π
- ?

1 / 1 point

 For each state s and action a , if $b(a | s) > 0$ then $\pi(a | s) > 0$ For each state s and action a , if $\pi(a | s) > 0$ then $b(a | s) > 0$ All actions have non-zero probabilities under π **Correct**Correct! Every action taken under π must have a non-zero probability under b .

8. Suppose the state
- s
- has been visited three times, with corresponding returns 8, 4, and 3. What is the current Monte Carlo estimate for the value of
- s
- ?

1 / 1 point

 3 15 5 3.5**Correct**

Correct! The Monte Carlo estimate for the state value is the average of sample returns observed from that state.

 After the first time step When every state is visited at least once At the end of the first episode**Correct**

Correct! Monte Carlo Prediction updates value estimates at the end of an episode.

10. In Monte Carlo prediction of state-values, memory requirements depend on (select all that apply)

1 / 1 point

 The number of states**Correct**

Correct! Monte Carlo Prediction needs to store the estimated value for each state.

 The number of possible actions in each state The length of episodes**Correct**

Correct! Monte Carlo Prediction needs to store the sequence of states and rewards during an episode

11. For Monte Carlo Prediction of state-values, the number of updates at the end of an episode depends on

1 / 1 point

 The length of the episode

Correct! Monte Carlo Prediction updates the estimated value of each state visited during the episode.

12. Which approach can find an optimal deterministic policy? (select all that apply)

1 / 1 point

 Exploring Starts**Correct**

Correct! Exploring starts ensure that every state-action pair is visited even if the policy is deterministic.

 ϵ -greedy exploration Off-policy learning with an ϵ -soft behavior policy and a deterministic target policy**Correct**

Correct! In this case, the behavior policy can maintain exploration while the target policy is deterministic.

13. In an
- ϵ
- greedy policy over
- A
- actions, what is the probability of the highest valued action if there are no other actions with the same value?

1 / 1 point

 $1 - \epsilon$ ϵ $1 - \epsilon + \frac{\epsilon}{A}$ $\frac{\epsilon}{A}$

Correct! Monte Carlo Prediction updates the estimated value of each state visited during the episode.