1. Which approach can find an optimal deterministic policy? (**Select all that apply**) 0.7 / 1 point

- [x] Off-policy learning with an ϵ-soft behavior policy and a deterministic target policy

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

- [x] ϵ-greedy exploration

Incorrect, with ϵ-greedy exploration the agent will find an ϵ-soft policy, which is stochastic. Please review Lesson 3 (Video: Epsilon-Soft Policies)

- [x] Exploring Starts

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

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 given a batch of data containing 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
- [x] When the problem is **episodic** and given a batch of data containing sample episodes (sequences of states, actions, and rewards)

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

- [x] When the problem is **episodic** and there is a model that produces samples of the next state and reward

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 point**

- [ ] Learning the optimal policy while continuing to explore
- [x] Learning from data generated by a human expert

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).

You didn't select all the correct answers

4. Which of the following is a requirement *on the behaviour policy* b for using **off-policy** Monte Carlo policy evaluation? This is called the *assumption of coverage*. **1 / 1 point**

- (●) For each state s and action a, if $\pi(a \mid s) > 0$ then $b(a \mid s) > 0$

Correct! Every action taken under π must have a non-zero probability under b.

- ( ) For each state s and action a, if $b(a \mid s) > 0$ then $\pi(a \mid s) > 0$
- ( ) All actions have non-zero probabilities under π

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**

- [x] When state values v_π and a model are available

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.

- [x] When action values q_π and a model are available

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

- [x] When action values q_π are available but no model is available.

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**

Hint: recall we used the term *sweep* in dynamic programming to discuss updating all the states systematically. This is **not** the same as visiting a state.

(●) Averaging sample returns

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

( ) Averaging sample rewards

( ) **Planning** with a model of the environment

( ) Performing **sweeps** through the state set

7. 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
- (x) 5

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

- ( ) 3.5

8. When does Monte Carlo prediction perform its first update?

1 / 1 point

- ( ) After the first time step
- ( ) After every state is visited at least once
- (x) At the end of the first episode

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

9. In Monte Carlo prediction of state-values, **memory** requirements depend on (Select all that apply).

1 / 1 point

Hint: think of the two data structures used in the algorithm

- [x] The number of states

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

- [ ] The number of possible actions in each state

- [x] The length of episodes

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

10. 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
- (•) $1-\epsilon+\frac{\epsilon}{\mathrm{A}}$

Correct! The highest valued action still has a chance of being selected as an exploratory action.

- ( ) $\frac{\epsilon}{\mathrm{A}}$

Like Dislike Report an issue