1. Which approach can find an optimal deterministic policy? (select all that apply) **0.6666666666666666 / 1 point**

- [x] ϵ-greedy exploration

  ✗ **This should not be selected**

  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] 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.

- [x] Exploring Starts

  ✓ **Correct**

  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 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

- [x] When the problem is episodic and there are sequences of states, actions, and rewards

Correct
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
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**

- [x] Learning about multiple policies simultaneously while following a single behavior policy

Correct
Correct! Off-policy learning enables learning about multiple target policies simultaneously using a single behavior policy.

- [x] Learning the optimal policy while continuing to explore

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

- [x] 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**

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

Correct

Correct! This is the importance sampling ratio and is used to weight returns in off-policy Monte-Carlo Policy Evaluation.

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

✓ 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**

○ Performing sweeps through the state set

○ Planning with a model of the environment

◉ Averaging sample returns

○ Averaging sample rewards

✓ **Correct**

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

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
- ◉ 5
- ○ 3.5

✓ **Correct**

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

8. When does Monte Carlo prediction perform its first update? **1 / 1 point**

- ○ 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.

9. In Monte Carlo prediction of state-values, **memory** requirements depend on (select all that apply) **1 / 1 point**

- [x] 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
- [x] The length of episodes

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

10. In an ϵ-greedy policy over $\mathcal{A}$ 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
- ◉ $1 - \epsilon + \frac{\epsilon}{\mathcal{A}}$
- ○ $\frac{\epsilon}{\mathcal{A}}$

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