1. Which approach ensures continual (never-ending) exploration? (**Select all that apply**) **1 / 1 point**

- [x] 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
- [x] On-policy learning with an ϵ-soft policy

  Correct
  Correct! ϵ-soft policies assign non-zero probabilities to all state-action pairs.

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

  Correct
  Correct! 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 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

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 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**

Hint: pay attention to the time subscripts of A and S in the answers below.

Hint: Sums and products are not the same things!

(●) $$\prod_{k=t}^{T-1} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)}$$

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

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

  Correct
  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
  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 rewards

- [x] Averaging sample returns
- [ ] Performing **sweeps** through the state set
- [ ] **Planning** with a model of the environment

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
- [x] 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
- [ ] After every state is visited at least once
- [x] 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**

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

- [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
- (x) $1 - \epsilon + \frac{\epsilon}{\mathcal{A}}$
- ( ) $\frac{\epsilon}{\mathcal{A}}$

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