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

- [ ] Exploring starts
- [ ] On-policy learning with a **deterministic** policy
- [ ] On-policy learning with an ϵ-soft policy
- [ ] Off-Policy learning with an ϵ-soft behavior policy and a **deterministic** target policy
- [ ] 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 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
- [ ] When the problem is **episodic** and given a batch of data containing sample episodes (sequences of states, actions, and rewards)
- [ ] When the problem is **episodic** and there is a model that produces samples of the next state and reward

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
- [ ] Learning from data generated by a human expert

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 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!

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

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 point**

- ☐ When state values v_π and a model are available
- ☐ When state values v_π are available but no model is available.
- ☐ When action values q_π and a model are available
- ☐ When action values q_π are available but no model is available.

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

- ○ **Planning** with a model of the environment
- ○ Averaging sample rewards
- ○ Averaging sample returns
- ○ 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 point**

- ◯ 3
- ◯ 15
- ◯ 5
- ◯ 3.5

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

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

9. For Monte Carlo Prediction of state-values, the number of **updates** at the end of an episode depends on **1 point**

Hint: look at the innermost loop of the algorithm

- ◯ The length of the episode
- ◯ The number of states
- ◯ The number of possible actions in each state

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 point**

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