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

- [x] Exploring Starts
- [ ] ϵ-greedy exploration
- [x] Off-policy learning with an ϵ-soft behavior policy and a deterministic target 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
- [x] When the problem is **episodic** and given a batch of data containing sample episodes (sequences of states, actions, and rewards)
- [x] 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

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

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 point

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

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

- ( ) Performing **sweeps** through the state set
- ( ) **Planning** with a model of the environment
- (•) Averaging sample returns
- ( ) Averaging sample 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 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 number of possible actions in each state
- ◉ The length of the episode
- ○ The number of states

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
- ◉ $1 - \epsilon + \frac{\epsilon}{\mathcal{A}}$
- ○ $\frac{\epsilon}{\mathcal{A}}$