**Graded Quiz**
Graded Assignment • 30 min

English Due Jun 9, 11:59 PM +07 Exit

## Your grade: 90%

Your latest: **90%** • Your highest: **90%** • To pass you need at least 80%. We keep your highest score.

Next item →

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

- [x] Exploring starts

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 $\epsilon$-soft policy

Correct! $\epsilon$-soft policies assign non-zero probabilities to all state-action pairs.

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

Correct! $\epsilon$-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 $\epsilon$-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! 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 / 1 point

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

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

- [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).

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

- ( ) All actions have non-zero probabilities under $\pi$
- (•) 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 $\pi(a \mid s) > 0$ then $b(a \mid s) > 0$

Correct! Every action taken under $\pi$ must have a non-zero probability under $b$.

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

5. When is it possible to determine a policy that is greedy with respect to the value functions $v_\pi, q_\pi$ for the policy $\pi$? (Select all that apply)

1 point

- [x] When state values $v_\pi$ 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.

- [x] When state values $v_\pi$ are available but no model is available.

Incorrect, please review Lesson 2 (Video: Using Monte Carlo for estimating action-values)

- [x] When action values $q_\pi$ and a model are available

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

- [x] When action values $q_\pi$ 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.

- Performing **sweeps** through the state set
- 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

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

- $3$
- $15$
- $5$

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
- At the end of the first episode

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

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

1 / 1 point

Hint: look at the innermost loop of the algorithm

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

Correct! Monte Carlo Prediction updates the estimated value of each state visited during the episode.

**10.** In an $\epsilon$-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

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

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

- ○ $\frac{\epsilon}{\mathcal{A}}$

Like Dislike Report an issue