**Поздравляем! Вы прошли тест!**
ДЛЯ УСПЕШНОГО ПРОХОЖДЕНИЯ 80% или выше

Продолжить обучение

ОЦЕНКА
84.61%

# Graded Quiz

ОЦЕНКА ПОСЛЕДНЕЙ РАБОТЫ
84.61%

**1. Which approach ensures continual exploration? (Select all that apply)** — 0 / 1 балл

- [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 $\epsilon$-soft policy

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

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

  You didn't select all the correct answers

**2. When can Monte Carlo methods, as defined in the course, be applied? (Select all that apply)** — 1 / 1 балл

- [ ] 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 балл

- [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 $\pi$ and the behavior policy $b$?** — 1 / 1 балл

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

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

**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 $v_\pi, q_\pi$ for the policy $\pi$? (Select all that apply)** 1 / 1 балл

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

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

**6. Monte Carlo methods in Reinforcement Learning work by...** 1 / 1 балл

- Averaging sample rewards
- (•) 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. Which of the following is a requirement for using Monte Carlo policy evaluation with a behavior policy $b$ for a target policy $\pi$?** 1 / 1 балл

- For each state $s$ and action $a$, if $b(a \mid s) > 0$ then $\pi(a \mid s) > 0$
- (•) For each state $s$ and action $a$, if $\pi(a \mid s) > 0$ then $b(a \mid s) > 0$
- All actions have non-zero probabilities under $\pi$

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

**8. 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 балл

- $3$
- $15$
- (•) $5$
- $3.5$

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

**9. When does Monte Carlo prediction perform its first update?** 1 / 1 балл

- 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.

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

- [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

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

- ( ) The number of states
- ( ) The number of possible actions in each state
- (•) The length of the episode

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

**12. Which approach can find an optimal deterministic policy? (select all that apply)** 0 / 1 балл

- [ ] Exploring Starts
- [ ] $\epsilon$-greedy exploration
- [x] Off-policy learning with an $\epsilon$-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.

You didn't select all the correct answers

**13. 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 балл

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

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