← Back **MDPs**
Practice Assignment • 45 min

English

Your grade: 100%

Your latest: **100%** • Your highest: **100%** • To pass you need at least 80%. We keep your highest score.

Next item →

1. The learner and decision maker is the _______. **1 / 1 point**

● Agent

○ State

○ Environment

○ Reward

✓ **Correct**
Correct!

2. At each time step the agent takes an _______. **1 / 1 point**

○ Environment

○ State

● Action

○ Reward

✓ **Correct**
Correct!

3. Imagine the agent is learning in an episodic problem. Which of the following is true? **1 / 1 point**

● The number of steps in an episode is stochastic: each episode can have a different number of steps.

○ The number of steps in an episode is always the same.

○ The agent takes the same action at each step during an episode.

✓ **Correct**
Correct!

4. If the reward is always +1 what is the sum of the discounted infinite return when $\gamma < 1$ **1 / 1 point**

$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

○ Infinity.

● $G_t = \frac{1}{1-\gamma}$

○ $G_t = \frac{\gamma}{1-\gamma}$

○ $G_t = 1 * \gamma^k$

✓ **Correct**
Correct!

5. How does the magnitude of the discount factor (gamma/γ) affect learning? **1 / 1 point**

○ With a smaller discount factor the agent is more far-sighted and considers rewards farther into the future.

● With a larger discount factor the agent is more far-sighted and considers rewards farther into the future.

○ The magnitude of the discount factor has no effect on the agent.

✓ **Correct**
Correct!

6. Suppose $\gamma = 0.8$ and we observe the following sequence of rewards: $R_1 = -3, R_2 = 5, R_3 = 2,$ $R_4 = 7$, and $R_5 = 1$, with $T = 5$. What is G_0? Hint: Work Backwards and recall that $G_t = R_{t+1} + \gamma G_{t+1}$. **1 / 1 point**

● 6.2736

○ 12

○ 8.24

○ 11.592

○ -3

✓ **Correct**
Correct!

7. What does MDP stand for? **1 / 1 point**

○ Meaningful Decision Process

● Markov Decision Process

○ Markov Deterministic Policy

○ Markov Decision Protocol

✓ **Correct**
Correct!

8. Consider using reinforcement learning to control the motion of a robot arm to pick up objects and place them into new positions. The actions in this case might be the voltages applied to each motor at each joint, and the states might be the latest readings of joint angles and velocities. The reward might be +1 for each object successfully picked up and placed. To encourage smooth movements, on each time step a small, negative reward can be given as a function of the moment-to-moment "jerkiness" of the motion. Is this a valid MDP? **1 / 1 point**

● Yes

○ No

✓ **Correct**
Correct!

9. **Case 1**: Imagine that you are a vision system. When you are first turned on for the day, an image floods into your camera. You can see lots of things, but not all things. You can't see objects that are occluded, and of course you can't see objects that are behind you. After seeing that first scene, do you have access to the Markov state of the environment? **1 / 1 point**

Case 2: Imagine that the vision system never worked properly: it always returned the same static imagine, forever. Would you have access to the Markov state then? (Hint: Reason about $P(S_{t+1}|S_t, ..., S_0)$, where S_t = AllWhitePixels)

● You have access to the Markov state in both Case 1 and 2.

○ You have access to the Markov state in Case 1, but you don't have access to the Markov state in Case 2.

○ You don't have access to the Markov state in Case 1, but you do have access to the Markov state in Case 2.

○ You don't have access to the Markov state in both Case 1 and 2.

✓ **Correct**
Correct! Because there is no history before the first image, the first state has the Markov property. The Markov property does not mean that the state representation tells all that would be useful to know, only that it has not forgotten anything that would be useful to know.

The case when the camera is broken is different, but again we have the Markov property. All the possible futures are the same (all white), so nothing needs to be remembered in order to predict them.

10. What is the reward hypothesis? **1 / 1 point**

○ That all of what we mean by goals and purposes can be well thought of as the minimization of the expected value of the cumulative sum of a received scalar signal (called reward)

○ Ignore rewards and find other signals.

○ Always take the action that gives you the best reward at that point.

● That all of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward)

✓ **Correct**
Correct!

11. Imagine, an agent is in a maze-like gridworld. You would like the agent to find the goal, as quickly as possible. You give the agent a reward of +1 when it reaches the goal and the discount rate is 1.0, because this is an episodic task. When you run the agent its finds the goal, but does not seem to care how long it takes to complete each episode. How could you fix this? (**Select all that apply**) **1 / 1 point**

□ Give the agent a reward of 0 at every time step so it wants to leave.

□ Give the agent a reward of +1 at every time step.

☑ Give the agent -1 at each time step.

> ✓ **Correct**
> Correct! Giving the agent a negative reward on each time step, tells the agent to complete each episode as quickly as possible.

☑ Set a discount rate less than 1 and greater than 0, like 0.9.

> ✓ **Correct**
> Correct! From a given state, the sooner you get the +1 reward, the larger the return. The agent is incentivized to reach the goal faster to maximize expected return.

12. When may you want to formulate a problem as episodic? **1 / 1 point**

● When the agent-environment interaction naturally breaks into sequences. Each sequence begins independently of how the episode ended.

○ When the agent-environment interaction does not naturally break into sequences. Each new episode begins independently of how the previous episode ended.

> ✓ **Correct**
> Correct!