

## MDPs

TOTAL POINTS 16

1. The learner and decision maker is the \_\_\_\_.

1 / 1 point

- ☒ Agent
- ☐ Reward
- ☐ State
- ☐ Environment

✓ Correct

Correct!

2. At each time step the agent takes an \_\_\_\_.

1 / 1 point

- ☐ Environment
- ☐ State
- ☒ Action
- ☐ Reward

✓ Correct

Correct!

3. What equation(s) define  $q_{\pi}(S_t, A_t)$  in terms of subsequent rewards?

1 / 1 point

- ☒  $q_{\pi}(s, a) = \mathbb{E}_{\pi}[G_t | S_t = s, A_t = a]$   
where:  $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} \dots$

✓ Correct

Correct!

- ☐  $q_{\pi}(s, a) = \mathbb{E}_{\pi}[G_t]$   
where:  $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} \dots$

- ☐  $q_{\pi}(s, a) = \mathbb{E}_{\pi}[R_{t+1} | S_t = s, A_t = a]$

- ☒  $q_{\pi}(s, a) = \mathbb{E}_{\pi}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} \dots | S_t = s, A_t = a]$

✓ Correct

Correct!

- ☐  $q_{\pi}(s, a) = [G_t | S_t = s, A_t = a]$   
where:  $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} \dots$

4. 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!

5. If the reward is always +1 what is the sum of the discounted infinite return when  $\gamma < 1$

1 / 1 point

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

- ☐  $G_t = \frac{\gamma}{1-\gamma}$

- ☐  $G_t = 1 * \gamma^k$

- ☒  $G_t = \frac{1}{1-\gamma}$

- ☐ Infinity.

✓ Correct

Correct!

6. What is the difference between a small gamma (discount factor) and a large gamma?

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 size of the discount factor has no effect on the agent.

✓ Correct

Correct!

7. 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

- ☐ -3

- ☐ 12

- ☐ 8.24

- ☒ 6.2736

- ☐ 11.592

✓ Correct

Correct!

8. Suppose  $\gamma = 0.8$  and the reward sequence is  $R_1 = 5$  followed by an infinite sequence of 10s. What is  $G_0$ ?

1 / 1 point

- ☒ 45

- ☐ 15

- ☐ 55

✓ Correct

Correct!

$$G_2 = 10 / (1 - 0.8) = 50$$

$$G_1 = 10 + .8 * (50) = 50$$

$$G_0 = 5 + .8 * 50 = 45$$

9. Suppose reinforcement learning is being applied to determine moment-by-moment temperatures and stirring rates for a bioreactor (a large vat of nutrients and bacteria used to produce useful chemicals). The actions in such an application might be target temperatures and target stirring rates that are passed to lower-level control systems that, in turn, directly activate heating elements and motors to attain the targets. The states are likely to be thermocouple and other sensory readings, perhaps filtered and delayed, plus symbolic inputs representing the ingredients in the vat and the target chemical. The rewards might be moment-by-moment measures of the rate at which the useful chemical is produced by the bioreactor. Notice that here each state is a list, or vector, of sensor readings and symbolic inputs, and each action is a vector consisting of a target temperature and a stirring rate. Is this a valid MDP?

1 / 1 point

- ☒ Yes

- ☐ No

✓ Correct

Correct!

10. Consider using reinforcement learning to control the motion of a robot arm in a repetitive pick-and-place task. If we want to learn movements that are fast and smooth, the learning agent will have to control the motors directly and have low-latency information about the current positions and velocities of the mechanical linkages. 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!

11. 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? Suppose your camera was broken that day and you received no images at all, all day. Would you have access to the Markov state then?

1 / 1 point

- ☒ You have access to the Markov state before and after damage.

- ☐ You have access to the Markov state before damage, but you don't have access to the Markov state after damage.

- ☐ You don't have access to the Markov state before damage, but you do have access to the Markov state after damage.

- ☐ You don't have access to the Markov state before or after damage.

✓ Correct

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. The key in this case is that the future is impoverished. All the possible futures are the same (all blank), so nothing need be remembered in order to predict them.

12. What does MDP stand for?

1 / 1 point

- ☒ Markov Decision Process

- ☐ Meaningful Decision Process

- ☐ Markov Decision Protocol

- ☐ Markov Deterministic Policy

✓ Correct

Correct!

13. What is the reward hypothesis?

1 / 1 point

- ☐ Ignore rewards and find other signals.

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

- ☒ Goals and purposes can be thought of as the maximization of the expected value of the cumulative sum of rewards received.

- ☐ Goals and purposes can be thought of as the minimization of the expected value of the cumulative sum of rewards received.

✓ Correct

Correct!

14. 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.

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

✓ Correct

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.

- ☒ Give the agent -1 at each time step.

✓ Correct

Correct!

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

15. 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!

16. When may you want to formulate a problem as continuing?

1 / 1 point

- ☐ When the agent-environment interaction naturally breaks into sequences and each sequence begins independently of how the previous sequence 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!