# Mini-Course 1, Module 2
# Markov Decision Processes
# Worksheet Class

CMPUT 397
Fall 2020

# Reminders: Sept 16, 2019

- Today, we will do worksheet questions
  - This one: https://marthawhite.github.io/rlcourse/docs/w-c1m2.pdf, link on the github schedule and pasted in Zoom
- I can go over a few questions in the practice quiz
- A few organizational polls

# Practice Quiz Review

- I can discuss Questions 9, 10, 11 and 14
- Feel free to ask questions about any of the others

# PQ 9

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?

   - ◯ Yes
   - ◯ No

# PQ 10

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?

- ◯ Yes
- ◯ No

# PQ 11

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?

- ◯ 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.

# PQ 14

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)

- [ ] Give the agent a reward of +1 at every time step.
- [ ] Give the agent -1 at each time step.
- [ ] Set a discount rate less than 1 and greater than 0, like 0.9.
- [ ] Give the agent a reward of 0 at every time step so it wants to leave.

# Worksheet Questions

- Worksheet: https://marthawhite.github.io/rlcourse/docs/w-c1m2.pdf
- Organization:
  - I will stay in the main room, and anyone who just wants to work on the worksheet quietly and type a clarifying question can stay there with me
  - Everyone still gets randomly assigned to a room, but can come back to the main one or leave to their own private one
  - If you stay in a breakout room, it can be more interactive (lead by the TA)

# Worksheet Question 1

Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2, R_2 = -2, R_3 = 0$ followed by an infinite sequence of 7s. What are $G_1$ and $G_0$?

# Worksheet Question 2

Assume you have a bandit problem with 4 actions, where the agent can see rewards from the set $\mathcal{R} = \{-3.0, -0.1, 0, 4.2\}$. Assume you have the probabilities for rewards for each action: $p(r|a)$ for $a \in \{1, 2, 3, 4\}$ and $r \in \{-3.0, -0.1, 0, 4.2\}$. How can you write this problem as an MDP? Remember that an MDP consists of $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P, \gamma)$.

**More abstractly**, recall that a Bandit problem consists of a given action space $\mathcal{A} = \{1, ..., k\}$ (the $k$ arms) and the distribution over rewards $p(r|a)$ for each action $a \in \mathcal{A}$. Specify an MDP that corresponds to this Bandit problem.

# Worksheet Question 3

Prove that the discounted sum of rewards is always finite, if the rewards are bounded: $|R_{t+1}| \leq R_{\max}$ for all $t$ for some finite $R_{\max} > 0$.

$$\left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| < \infty \qquad \text{for } \gamma \in [0, 1)$$

Hint: Recall that $|a + b| < |a| + |b|$.

# Worksheet Question for Bandits

(Exercise 2.2 from S&B 2nd edition) Consider a $k$-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using $\epsilon$-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all $a$. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = 1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = 2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the $\epsilon$ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?