ASEN 5519-002 Decision Making under Uncertainty
Quiz 2: Reinforcement Learning and POMDPs

Show all work and justify and box answers.
You may consult any source, but you may NOT communicate with any person except the instructor.
If you run into a problem that you don't know how to solve, *skip it* and come back later to maximize your score.

Question 1. (20 pts) You enter a casino and there are two slot machines, A and B, that pay either \$0 or \$1 per play, each with potentially different winning probabilities. You have been playing for a few rounds and keeping track of the outcomes of each attempt and want to use UCB1 exploration with exploration parameter $c = \$1$ to select your actions.

a) If you have played A 10 times, receiving winnings of \$8, and B 4 times, winning \$3. Which machine should you play next according to UCB1 exploration? Justify your answer quantitatively.

b) If you have played A 10 times, winning \$9, and B 10 times, winning \$8. Which machine should you play next according to UCB1 exploration? Justify your answer quantitatively.

c) In which of the above situations did UCB1 exploration select the same action as a greedy policy? Justify your answer.

Question 2. (30 pts) Consider a 2-state, 2 action POMDP with $\mathcal{S} = \{1, 2\}$ and $\mathcal{A} = \{0, 1\}$. State 2 is terminal and the discount factor is $\gamma = 0.9$. Suppose that you are performing reinforcement learning, and you observe an episode that takes the following 3-step trajectory:

$$(s = 1, a = 0, r = 1, s' = 1)$$

$$(s = 1, a = 0, r = 1, s' = 1)$$

$$(s = 1, a = 1, r = 1, s' = 2)$$

a) (6 pts) Suppose you are using **maximum likelihood model-based reinforcement learning (MLMBRL)**. After observing the trajectory above, what are the maximum likelihood transition probabilities for action $a = 0$?

b) (18 pts) Suppose that you are using the **Q-learning** algorithm with learning rate $\alpha = 0.1$ and all Q values starting at 0 before the episode. What are the Q value estimates after the episode?

c) (6 pts) Suppose that you are using **policy gradient** with a policy parameterized with $\theta = [\theta_1, \theta_2]$ defined as

$$\pi_\theta(a = 1 | s = i) = [\theta_i]_0^1$$

where $[x]_a^b = \text{clamp}(x, a, b) = \min(b, \max(a, x))$. That is, θ_i is probability of taking action 1 in state i. If the parameter values are $\theta = [0.5, 0.5]$, what is the policy gradient estimate calculated from the trajectory above? Do not use baseline subtraction. You may find the following derivatives useful: $\frac{\partial}{\partial \theta_i} \log(\theta_i) = \frac{1}{\theta_i}$, $\frac{\partial}{\partial \theta_i} \log(1 - \theta_i) = -\frac{1}{1-\theta_i}$.

Question 3. (50 pts) You have been tasked with preventing poaching at a large national park. This problem can be formulated as a POMDP with two states: either there are poachers (P), or the park is clear (C). There are three actions, fly a UAV over the park (U), send in rangers in jeeps (J), or wait (W). At each step, you receive an observation of the state ($\mathcal{O} = \{P, C\}$). If the action is wait, both observations are equally likely. If the UAV or jeeps are employed, the observation is always accurate. Action J eliminates poachers immediately, otherwise their presence remains unchanged. Action J always results in a cost of 5 regardless of the state; waiting (W) has no cost if the state is C, and a cost of 10 if the state is P; the UAV (U) has a cost of 1 if the state is C and 11 if the state is P. In summary,

$$\mathcal{S} = \mathcal{O} = \{P, C\} \tag{1}$$

$$\mathcal{A} = \{U, J, W\} \tag{2}$$

$$\mathcal{R}(s, a) = \begin{cases} -5 \text{ if } a = J \\ -10 \text{ if } a = W \text{ and } s = P \\ -1 \text{ if } a = U \text{ and } s = C \\ -11 \text{ if } a = U \text{ and } s = P \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

$$\mathcal{T}(s' \mid s, a) = \begin{cases} 1 \text{ if } a \in \{U, W\} \text{ and } s' = s \\ 1 \text{ if } a = J \text{ and } s' = C \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

$$\mathcal{Z}(o \mid a, s') = \begin{cases} 1 \text{ if } a \in \{U, J\} \text{ and } o = s' \\ 0 \text{ if } a \in \{U, J\} \text{ and } o \neq s' \\ 0.5 \text{ if } a = W \end{cases} \tag{5}$$

$$\gamma = 1 \tag{6}$$

($\gamma = 1$ means this problem is ill-defined for an infinite horizon, but we only consider finite-horizon plans).

a) Calculate and write out **one step** alpha vectors for each action.

b) Draw and label the **one step** alpha vectors in the manner done in class.

c) According to the policy defined by the one-step alpha vectors above, under what circumstances would you take the U action? Why?

d) Suppose you use a certainty-equivalent approach that takes the best action for the most likely state. Which action would this CE approach *avoid* in this POMDP[1]? Why?

e) Draw diagrams[2] for the following **two step** conditional plans:

   (a) Always wait (W)

   (b) Always send in rangers on jeeps (J)

   (c) Fly the UAV (U), then wait (W) if the observation is clear (C) or send in jeeps (J) if poachers are detected (P)

f) Calculate and write out the alpha vectors for the **two step** conditional plans above.

g) Draw and label the **two step** alpha vectors for the conditional plans above in the manner done in class.

h) In a policy defined by the **two step** plans above, what action would be selected if the belief is uniform (i.e. $b(P) = 0.5$)?

[1] For this part only, you can assume $\gamma < 1$

[2] similar to Figure 20.1 in the book