ASEN 5264 Decision Making under Uncertainty
Exam 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. (25 pts) A space probe is trying to communicate with a ground station and has three antennas that it can use to communicate with earth on. Each antenna has a probability of successful message transmission that is unknown to the probe. This situation can be modeled as a Bernoulli multi-armed bandit where each frequency band is an arm, and each successful message transmission is a win.

a) Suppose the number of wins and losses for the arms are $w = [1, 8, 4]$ and $l = [1, 2, 4]$. Write down the *probabilities* of choosing each arm at the next time step under the following strategies:

- Explore-then-commit with $k = 30$.
- ϵ greedy with $\epsilon = 0.1$.
- UCB with $c = 2$.

b) Suppose that after the numbers of wins and losses above, a fault makes antenna 2 inoperable. Which antenna should be used at the next time step under a UCB strategy with $c = 2$?

Question 2. (15 pts) Consider a 1 state, 2 action infinite horizon MDP defined as follows:

$$\mathcal{S} = \{1\} \quad \mathcal{A} = \{1, 2\} \tag{1}$$
$$\mathcal{R}(1, 1) = 1 \quad \mathcal{R}(1, 2) = 0 \tag{2}$$
$$\mathcal{T}(1 \mid 1, a) = 1 \; \forall \, a \quad \gamma = 0.9 \tag{3}$$

Suppose that the current Q-value estimates are

$$Q(1, 1) = 8 \tag{4}$$
$$Q(1, 2) = 6. \tag{5}$$

What will the estimated Q values be after a single update of each of the following algorithms with learning rate $\alpha = 0.1$:

a) Q-learning with $a = 1$

b) SARSA with $a = 1$ and $a' = 2$

Question 3. (50 pts) Suppose that you are trying to decide whether to invest in Germanium Hills Bank. The bank might be well-managed (W) or poorly-managed (P), and this will not change as you are deciding whether to invest. At each time step, you can either invest (I), decline (D), or research the financial position of the bank (R). If you invest in a well-managed bank, you will receive a +10 reward; if you invest in a poorly managed bank, you receive a -10 reward. Researching has a reward of -1, but after researching, you receive an observation of the management state of the bank that is accurate 90% of the time. In summary, you have a POMDP defined as follows:

$$\mathcal{S} = \mathcal{O} = \{W, P\} \tag{6}$$

$$\mathcal{A} = \{D, I, R\} \tag{7}$$

$$\mathcal{R}(s,a) = \begin{cases} 0 \text{ if } a = D \\ -1 \text{ if } a = R \\ +10 \text{ if } a = I \text{ and } s = W \\ -10 \text{ if } a = I \text{ and } s = P \end{cases} \tag{8}$$

$$\mathcal{T}(s' \mid, s, a) = \begin{cases} 1 & \text{if } s = s' \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$\mathcal{Z}(o \mid a, s') = \begin{cases} 0.9 \text{ if } a = R \text{ and } o = s' \\ 0.1 \text{ if } a = R \text{ and } o \neq s' \\ 0.5 \text{ if } a \in \{D, I\} \end{cases} \tag{10}$$

$$\gamma = 0.9 \tag{11}$$

(Note that this formulation does not include terminal states, but we will only consider up to two-step 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 R action? Why?

d) Draw diagrams (similar to Figure 20.1 in the book) for the following **two step** conditional plans:

   (a) Always decline (D).

   (b) First research (R) and then decline (D) if the observation is poor management (P) or invest (I) if the observation is well-managed (W).

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

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

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

Question 4. (5 pts) Suppose that you are given a large batch of pre-collected trajectories from an environment and cannot collect any additional data. Which model-free reinforcement learning algorithm would be easier to use: Q learning or Policy Gradient? Briefly justify your answer.

Question 5. (5 pts) Consider the following function approximator:

$$f(x) = \sigma(a_2) + M_2(a_1 + M_1 x)$$

where M_i and a_i are matrix and vector parameters and σ is the sigmoid function. Does this function approximator have more or less expressive power than a neural network? In other words, can this function approximate a wider or narrower range of functions than a neural network? Briefly justify your answer.