INSTRUCTIONS

- **Due: Monday, April 1, 2024 at 10:00 PM EDT.** Remember that you may use up to 2 slip days for the written homework making the last day to submit **Wednesday, April 3, 2024 at 10:00 PM EDT**.
- **Format:** Write your answers in the `yoursolution.tex` file and compile a pdf (preferred) or you can type directly on the blank pdf. Make sure that your answers are within the dedicated regions for each question/part. If you do not follow this format, we may deduct points. Handwritten solutions are not acceptable and may lead to lost points.
- **How to submit:** Submit a pdf with your answers on Gradescope. Log in and click on our class 15-281, click on the HW8 assignment, and upload your pdf containing your answers. **Misaligned submissions will have at least 5% taken off their score.**
- **Policy:** See the course website for homework policies and academic integrity.

Name	
Andrew ID	
Hours to complete?	◯ (0, 2] hours ◯ (2, 3] hours ◯ (3, 4] hours ◯ (4, 5] hours ◯ (5, 6] hours ◯ (6, 7] hours ◯ (7, 8] hours ◯ > 8 hours

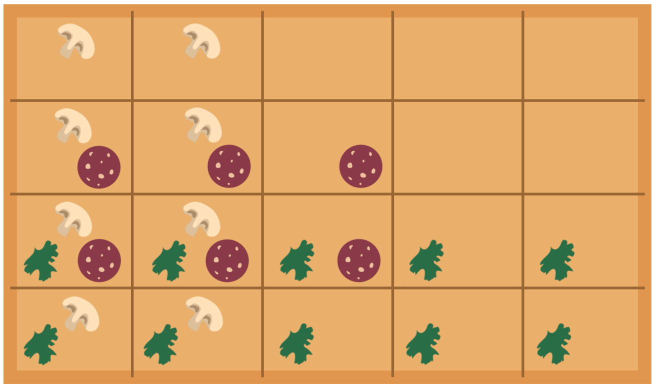
Q1. [11 pts] Probability: Product Rule and Chain Rule

Omega Pizzeria

As we step through the probability concepts of product rule and chain rule, you may find it helpful to check your understanding by referring back to our Omega Pizzeria example.

Let's define the following binary random variables and outcomes:

- X_1: Spinach random variable
  - $-x_1$: no spinach outcome
  - $+x_1$: spinach outcome
- X_2: Mushroom random variable
  - $-x_2$: no mushrooms outcome
  - $+x_2$: mushrooms outcome
- X_3: Pepperoni random variable
  - $-x_3$: no pepperoni outcome
  - $+x_3$: pepperoni outcome



Product Rule

Sometimes you have conditional and marginal distributions, but you actually want to compute the full joint distribution. We know the basic product rule:

$$P(X_1, X_2) = P(X_1 \mid X_2)P(X_2) \tag{1}$$

and

$$P(X_1, X_2) = P(X_2 \mid X_1)P(X_1) \tag{2}$$

We can generalize the product rule a bit more when there are more than two variables. The following are two valid ways to break up the joint distribution of three variables using a more general application of product rule:

$$P(X_1, X_2, X_3) = P(X_1, X_2 \mid X_3)P(X_3) \tag{3}$$

and

$$P(X_1, X_2, X_3) = P(X_1 \mid X_2, X_3)P(X_2, X_3) \tag{4}$$

You can check the math with the pizzeria to make sure that all three of these sentences are equal:

- The probability of getting a slice with spinach, mushrooms, and pepperoni.
- (The probability of getting a slice with spinach and mushrooms, given that we asked for a slice with pepperoni) times (the probability of getting a slice with pepperoni)
- (The probability of getting a slice with spinach, given that we asked for a slice with mushrooms and pepperoni) times (the probability of getting a slice with mushrooms and pepperoni)

Chain Rule

We can further break down equation 4 by applying the product rule again, $P(X_2, X_3) = P(X_2 \mid X_3)P(X_3)$:

$$\begin{aligned} P(X_1, X_2, X_3) &= P(X_1 \mid X_2, X_3)P(X_2, X_3) && (4) \\ &= P(X_1 \mid X_2, X_3)P(X_2 \mid X_3)P(X_3) && (5) \end{aligned}$$

This brings us to the general chain rule for N random variables, $X_1, X_2, \ldots, X_N$:

$$P(X_1, X_2, \ldots, X_N) = \prod_{n=1}^{N} P(X_n \mid X_1, \ldots, X_{n-1}) \tag{6}$$

Mutton, Lettuce, and Tomato

"True love is the greatest thing in the world...except a nice MLT: mutton, lettuce, and tomato sandwich, where the mutton is nice and lean..." –*Miracle Max*

You are given the following probability tables for binary random variables M, L, T:

T	$P(T)$
$+t$	0.4
$-t$	0.6

T	L	$P(L \mid T)$
$+t$	$+l$	0.8
$+t$	$-l$	0.2
$-t$	$+l$	0.25
$-t$	$-l$	0.75

L	T	M	$P(M \mid L, T)$
$+l$	$+t$	$+m$	0.95
$+l$	$+t$	$-m$	0.05
$+l$	$-t$	$+m$	0.75
$+l$	$-t$	$-m$	0.25
$-l$	$+t$	$+m$	0.40
$-l$	$+t$	$-m$	0.60
$-l$	$-t$	$+m$	0.10
$-l$	$-t$	$-m$	0.90

(a) [8 pts] Calculate $P(L, T, M)$ from the tables given.

L	T	M	$P(L, T, M)$
$+l$	$+t$	$+m$	
$+l$	$+t$	$-m$	
$+l$	$-t$	$+m$	
$+l$	$-t$	$-m$	
$-l$	$+t$	$+m$	
$-l$	$+t$	$-m$	
$-l$	$-t$	$+m$	
$-l$	$-t$	$-m$	

(b) [3 pts] Which of the following are valid decompositions of the joint probability distribution of M, L, and T, given no assumptions about the relationship between these random variables? Select all that apply.

- ☐ i) $P(M)P(L)P(T)$
- ☐ ii) $P(M)P(M, L)P(M, L, T)$
- ☐ iii) $P(M, L \mid T)P(T)$
- ☐ iv) $P(M \mid L, T)P(L)$
- ☐ v) $P(M \mid L, T)P(T)$
- ☐ vi) None of the above

Q2. [16 pts] Probability: Chain Rule, Joint Distributions, and Marginalization

Recall *marginalization*, which means summing out unwanted variables from a joint distribution. Consider three binary random variables A, B, and C with domains $\{+a, -a\}$, $\{+b, -b\}$, and $\{+c, -c\}$, respectively. Remember that $P(A)$ refers to the table of probabilities of all the elements of A's domain.

(a) [4 pts] Express $P(A)$ in terms of the joint distribution $P(A, B, C)$. Your answer should contain summation notation.

$P(A) =$

(b) [4 pts] Express $P(A)$ in terms of $P(C)$, $P(B \mid C)$ and $P(A \mid B, C)$. Your answer should contain summation notation.

$P(A) =$

(c) [8 pts] Expand the sums from part (b) to express the two elements of $P(A)$ ($P(+a)$ and $P(-a)$) in terms of the individual probabilities (e.g. $P(+b \mid +c)$, $P(-c)$ instead of tables $P(B \mid C)$ or $P(C)$). Your answer should *NOT* contain summation notation.

$P(+a) =$

$P(-a) =$

Q3. [12 pts] More Probability

Pacman is trying to find a ghost in the Pacman grid. The ghost is not visible to Pacman, but Pacman can take sensor readings at different grid locations to help learn more about the ghost's location.

- The sensor lights up one of five different colors that give a noisy indication of how far away the ghost is.
- The grid has 5 x 6 = 30 different locations.

We are given probability tables for:

- $P(G)$: The probability of a ghost being at a location
- $P(C_{1,1} \mid G)$: The probability of the color of the sensor in grid location (1,1) given the location of the ghost.

We want to help Pacman compute:

- $P(G \mid C_{1,1})$: The probability of the ghost being at a location given the color of the sensor in (1,1).

(a) [6 pts] How many entries are in the tables and what value do the entries sum to? Write a '?' if there is not enough information.

(i) [2 pts] $P(G)$

Num:	**Sum:**

(ii) [2 pts] $P(C_{1,1} \mid G)$

Num:	**Sum:**

(iii) [2 pts] $P(G \mid C_{1,1})$

Num:	**Sum:**

(b) [3 pts] Given that the sensor in location (1,1) is *orange*, write an equation for the probability that the ghost is in location (1,1), referencing the probability tables that we have been given.

$P(G = ghost_{1,1} \mid C_{1,1} = orange) =$

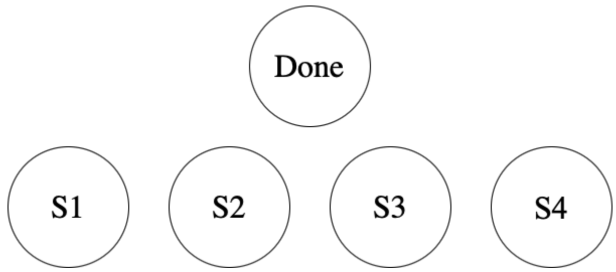
(c) [3 pts] Given that the sensor in location (1,1) is *orange*, write an equation that computes the most likely value of the variable G, referencing the probability tables that we have been given.

$\hat{G} =$

Q4. [11 pts] Reinforcement Learning

Trivia Game Show

Consider the simplified MDP from the Trivia Game Show lecture activity shown below. Assume that we no longer know the transition or reward functions (a.k.a we do not know the arrows shown between the states).



We want to practice Q-learning using the states $\{S1, S2, S3, S4, Done\}$ and actions $\{Play, Stop\}$ given a certain sequence of samples. Each sample is of the form (s, a, s', r) where (s, a) is the state-action pair, s' is the resulting state, and r is the associated outcome reward. For each sample (s, a, s', r) in our episode we update exactly one Q-value according to the following formula:

$$Q_{i+1}(s,a) = (1-\alpha)Q_i(s,a) + \alpha[r + \gamma \text{max}_{a'} Q_i(s', a')]$$

Note that the specific Q-value we update in any iteration is dependent on the specific state-action pair (s, a) we see in our sample. For our calculations, we will fix $\gamma = 0.9$. However our value of α will change over time, generally decreasing with each sample we look at.

(a) [1 pt] All Q-values are initialized to zero as shown below. The first sample is $(S1, Play, S1, -10)$ and $\alpha = 1$. Show your calculation for the correct updated Q-value. **Mark the circle corresponding to this Q-value.**

iteration 0	$S1$	$S2$	$S3$	$S4$	$Done$
$Q(s, Play)$	0	0	0	0	0
$Q(s, Stop)$	0	0	0	0	0
iteration 1	$S1$	$S2$	$S3$	$S4$	$Done$
$Q(s, Play)$	○	○	○	○	0
$Q(s, Stop)$	○	○	○	○	0

$Q_1(s, a)$**:**

(b) [1 pt] Below are the Q-values after i iterations. The next sample is $(S2, Play, S3, 0)$ and $\alpha = 0.75$. Show your calculation for the correct updated Q-value. **Mark the circle corresponding to this Q-value.**

i	$S1$	$S2$	$S3$	$S4$	$Done$
$Q(s, Play)$	-1	-3	-5	0	0
$Q(s, Stop)$	0	100	300	0	0
$i+1$	$S1$	$S2$	$S3$	$S4$	$Done$
$Q(s, Play)$	○	○	○	○	0
$Q(s, Stop)$	○	○	○	○	0

$Q_{i+1}(s, a)$**:**

(c) [2 pts] Below are the Q-values after i' iterations. The next two samples are $(S3, Play, S1, -10)$ and $(S1, Stop, Done, 0)$. Show your calculations for the correct updated Q-values. **Mark the circles corresponding to these Q-values.** Assume $\alpha = 0.5$ for both calculations.

i'	$S1$	$S2$	$S3$	$S4$	$Done$
$Q(s, Play)$	80	186	220	0	0
$Q(s, Stop)$	0	100	300	500	0
$i'+1$	$S1$	$S2$	$S3$	$S4$	$Done$
$Q(s, Play)$	○	○	○	○	0
$Q(s, Stop)$	○	○	○	○	0
$i'+2$	$S1$	$S2$	$S3$	$S4$	$Done$
$Q(s, Play)$	○	○	○	○	0
$Q(s, Stop)$	○	○	○	○	0

$Q_{i'+1}(s, a)$:

$Q_{i'+2}(s, a)$:

Conceptual Section

(d) [7 pts]

(i) [1 pt] Temporal difference learning is an online learning method.

○ True ○ False

(ii) [1 pt] Q-learning: Using an optimal exploration function leads to no regret while learning the optimal policy.

○ True ○ False

(iii) [1 pt] In a deterministic MDP (i.e. one in which each state / action leads to a single deterministic next state), the Q-learning update with. a learning rate of α=1 will correctly learn the optimal q-values (assume that all state/action pairs are visited sufficiently often).

○ True ○ False

(iv) [1 pt] A small discount (close to 0) encourages greedy behavior.

○ True ○ False

(v) [1 pt] A large, negative living reward ($<< 0$) encourages greedy behavior.

○ True ○ False

(vi) [1 pt] A negative living reward can always be expressed using a discount < 1.

○ True ○ False

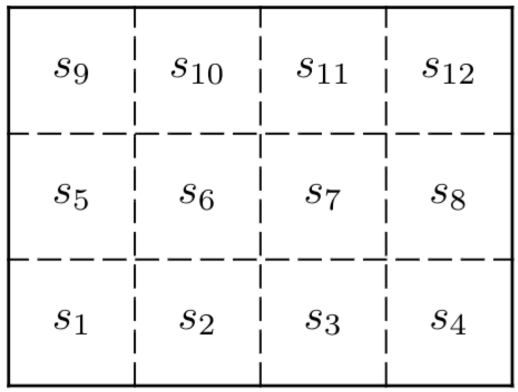
(vii) [1 pt] A discount < 1 can always be expressed as a negative living reward.

○ True ○ False

Q5. [16 pts] Dark Room Q-learning

A robot mysteriously finds itself in a dark room with 12 states as shown below. The robot can take four actions at each state (North, East, South, and West). We do not know where the interior walls are, except for the walls surrounding the room (represented by solid lines).

An action against a wall leaves the robot in the same state. Otherwise, the outcome of an action deterministically moves the robot in the direction corresponding to the action.



The robot learns the Q-values below. Note that only the maximum values for each state are shown, as the other values (represented with '-') do not affect the final policy.

	N	E	S	W
s_1	80	-	-	-
s_2	-	-	-	68
s_3	-	-	-	57
s_4	49	-	-	49
s_5	94	-	-	-
s_6	-	-	57	-
s_7	-	49	-	49
s_8	57	-	-	-
s_9	110	-	-	110
s_{10}	-	-		94
s_{11}	-	-		80
s_{12}	-	-		68

(a) [4 pts]

Using the learned Q-values, what are the first six actions the robot takes if it starts in state s_6? If there is more than one best action available, choose one randomly. For now, ignore any potential interior walls.

(b) [6 pts]

We are told the discount factor used during Q-learning was $\gamma = 0.85$.
We are also told there exists a single state, s^* such that $R(s^*, a, s') > 0 \ \forall a, s'$, and for all other states s, $R(s, a, s') = 0 \ \forall a, s'$. Also assume that $R(s^*, a, s')$ are equivalent for all a, s'.

(i) [2 pts] What is s^*, and what is the reward?

s^***:**

$R(s^*, a, s')$**:**

(ii) [2 pts] Explain how you found s^* in one sentence.

Answer:

(iii) [2 pts] Show how you solved for $R(s^*, a, s')$.

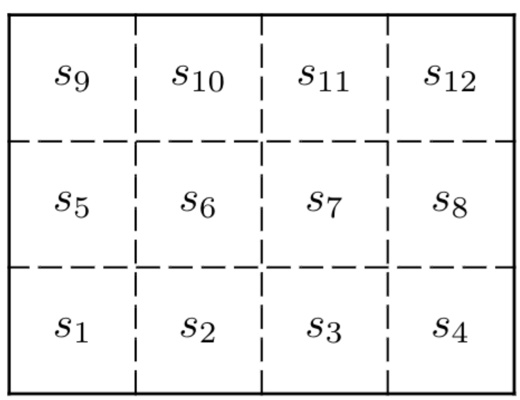
Answer:

(c) [6 pts]

Now the robot wants to locate the interior walls within the room.

(i) [4 pts] Where are these walls? Draw them in by filling the corresponding dashed lines below. For your submission to this problem, you may do one of the following:

- Draw/annotate on top of the existing images in the pdf.
- Edit the `figures/q5_states.png` image file to add markings.



(ii) [2 pts] In 1-2 sentences, explain how you know where the walls are.

Q6. [26 pts] Approximate Q-learning

A robot is trying to get to its office hours, occurring on floors 3, 4, or 5 in GHC. It is running a bit late and there are a lot of students waiting for it. There are three ways it can travel between floors in Gates: the stairs, the elevator, and the helix.

The **state** of the robot is the floor that it is currently on (either 3, 4, or 5).

The **actions** that the robot can take are *stairs*, *elevator*, or *helix*.

In this problem, we are using a linear, feature based approximation of the Q-values:

$$Q_w(s,a) = \sum_{i=0}^{3} f_i(s,a) w_i$$

We define the feature functions as follows:

Features	Initial Weights
$f_0(s,a) = 1$ (this is a bias feature that is always 1)	$w_0 = 1$
$f_1 = f_{\text{speed}}(s,a) = (\lvert s-4\rvert + 1)t$, where $t = \begin{cases} 46, a = \text{elevator} \\ 20, a = \text{stairs} \\ 8, a = \text{helix} \end{cases}$	$w_1 = 0.5$
$f_2 = f_{\text{accessibility}}(s,a) = \begin{cases} 2, a = \text{stairs} \\ 5, a = \text{helix} \\ 8, a = \text{elevator} \end{cases}$	$w_2 = 3$
$f_3 = f_{\text{emptiness}}(s,a) = \begin{cases} 60, a = \text{elevator}, s = 3 \\ 80, a = \text{elevator}, s = 4 \\ 60, a = \text{elevator}, s = 5 \\ 0, \text{otherwise} \end{cases}$	$w_3 = 0.2$

Furthermore, the weights will be updated as follows:

$$w_i \leftarrow w_i + \alpha[r + \gamma \max_{a'} Q_w(s',a') - Q_w(s,a)] \frac{\delta}{\delta w_i} Q_w(s,a)$$

(a) [6 pts] Calculate the following initial Q values given the initial weights above.

(i) [2 pts] Q_w**(4, elevator):**

(ii) [2 pts] Q_w**(4, stairs):**

(iii) [2 pts] Q_w**(4, helix):**

(b) [2 pts] For this question, suppose that the initial Q-values for state 3 happen to be equal to the corresponding initial Q-values for state 5.

(i) [1 pt] In this problem, as you update the weights, will these values remain equal? I.e., will $Q_w(3, a) = Q_w(5, a)$ given any action a and vector w?

◯ Yes ◯ No

(ii) [1 pt] Why or why not?

Answer:

(c) [6 pts] Given the Q-values for state 4 calculated in part (a), what are the probabilities that each of the following actions could be chosen when using ϵ-greedy exploration from state 4 (assume random movements are chosen uniformly from all actions)? Let $\epsilon = \frac{1}{2}$ for this problem.

(i) [2 pts] Elevator

$P(elevator) =$

(ii) [2 pts] Stairs

$P(stairs) =$

(iii) [2 pts] Helix

$P(helix) =$

(d) [8 pts] Given a sample with start state 3, action = stairs, successor state = 4, and reward = -2, update each of the weights using learning rate $\alpha = 0.15$ and discount factor $\gamma = 0.6$.

(i) [2 pts] w_0

$w_0 =$

(ii) [2 pts] w_1

$w_1 =$

(iii) [2 pts] w_2

$w_2 =$

(iv) [2 pts] w_3

$w_3 =$

(e) [4 pts]

(i) [2 pts] What is an advantage of using approximate Q-learning instead of the standard Q-learning?

Advantage:

(ii) [2 pts] What is a disadvantage of using approximate Q-learning instead of the standard Q-learning?

Disadvantage:

Q7. [8 pts] Ethics

Consider an application of reinforcement learning in which it is possible to train an AI chatbot to provide "helpful" feedback by rating the responses it gives positively +1 (helpful) or negatively -1 (unhelpful).

(a) [3 pts] Suppose a person is giving the feedback. People may not all agree on the same ratings for the same scenarios. How does this inconsistent feedback impact reinforcement learning in terms of the equations?

Answer:

(b) [3 pts] How might you incorporate the feedback from multiple people so that it is consistent?

Answer:

Now consider an agent that has been trained to be helpful. You'd like to test it thoroughly before putting it on the web for other people to use. See the following tweet for an (entertaining) example of how assistants can provide helpful yet harmful information for people online: `https://twitter.com/AnthropicAI/status/1562828015502954498`.

(c) [2 pts] In your opinion, would it be ethical to deploy this agent to the public? Why or why not?

Answer:

For more information on training a helpful/harmless agent and then red-teaming (testing) the agent, see the following paper:
`https://arxiv.org/pdf/2204.05862.pdf`

Note, these articles are rather technical and not part of the course material. However, an interesting observation from the articles is that by over-optimizing for harmlessness to avoid malicious situations like the one above, their new agent initially learned the extreme policy of suggesting professional help or therapy even in rather mild situations.