CS 188 Spring 2019 | Introduction to Artificial Intelligence

Practice Final

To earn the extra credit, one of the following has to hold true. Please circle and sign.

A I spent 4 or more hours on the practice final.

B I spent fewer than 4 hours on the practice final, but I believe I have solved all the questions.

Signature: ____

Q1. [11 pts] Approximate Q-Learning

Consider the following MDP: We have infinitely many states $s \in \mathbb{Z}$ and actions $a \in \mathbb{Z}$, each represented as an integer. Taking action a from state s deterministically leads to new state $s' = s + a$ and reward $r = s - a$. For example, taking action 3 at state 1 results in new state $s' = 1 + 3 = 4$ and reward $r = 1 - 3 = -2$.

We perform approximate Q-Learning, with features and initialized weights defined below.

Feature	Initial Weight
$f_1(s, a) = s$	$w_1 = 1$
$f_2(s, a) = -a^2$	$w_2 = 2$

(a) [3 pts] Write down $Q(s, a)$ in terms of w_1, w_2, s, and a.

____

(b) [2 pts] Calculate $Q(1, 1)$.

____

(c) [4 pts] We observe a sample (s, a, r, s') of $(1, 1, 0, 2)$. Assuming a learning rate of $\alpha = 0.5$ and discount factor of $\gamma = 0.5$, compute new weights after a single update of approximate Q-Learning.

w_1: ____________________________

w_2: ____________________________

(d) [2 pts] Compute the new value for Q(1,1).

____

Q2. [12 pts] Who Spoke When

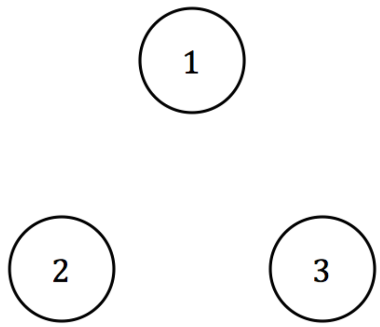
We are given a single audio recording (divided into equal and short time slots) and wish to infer when each person speaks. At every time step exactly one of N people is talking. This problem can be modeled using an HMM. Hidden variable $X_t \in \{1, 2, ..., N\}$ represents which person is talking at time step t.

(a) For this part, assume that at each time step:

- with probability p, the current speaker will continue to talk in the next time step.
- with probability $1 - p$, the current speaker will be interrupted by another person. Each other person is equally likely to be the interrupter.

Assume that $N = 3$.

(i) [2 pts] Complete the Markov Chain below and write down the probabilities on each transition.



____

(ii) [2 pts] What is the stationary probability distribution of this Markov chain? (Again, assume $N = 3$).

$P(X_{\text{inf}} = 1) =$ ________________________

$P(X_{\text{inf}} = 2) =$ ________________________

$P(X_{\text{inf}} = 3) =$ ________________________

____

(b) [2 pts] What is the **number of parameters** (or degrees of freedom) needed to model the transition probability $P(X_t|X_{t-1})$? Assume N people in the meeting and arbitrary transition probabilities.

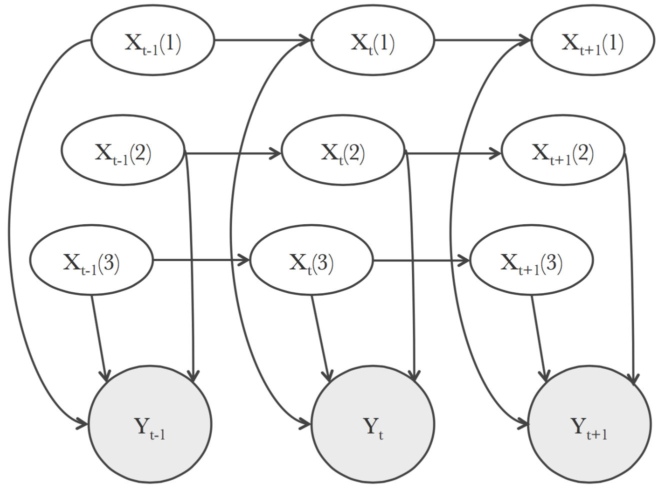
____

(c) [2 pts] Let's remove the assumption that people are not allowed to talk simultaneously. Now, hidden state $X_t \in \{0, 1\}^N$ will be a binary vector of length N. Each element of the vector corresponds to a person, and whether they are speaking.

Now, what is the **number of parameters** (or degrees of freedom) needed for modeling the transition probability $P(X_t|X_{t-1})$?

____

One way to decrease the parameter count is to assume independence. Assume that the transition probability between people is independent. The figure below represents this assumption for $N = 3$, where $X_t = [X_t(1), X_t(2), X_t(3)]$.



(d) [4 pts] Write the following in terms of conditional probabilities given from the Bayes Net. Assume N people in the meeting.

Transition Probability $P(X_t|X_{t-1})$

____

Emission Probability $P(Y_t|X_t)$

____

(e) [2 pts] What is the **number of parameters** (or degrees of freedom) needed for modeling transition probability $P(X_t|X_{t-1})$? Assume N people in the meeting.

____

Q3. [10 pts] Graph Search

You are trying to plan a road trip from city A to city B. You are given an **undirected graph** of roads of the entire country, together with the distance along each road between any city X and any city Y: $length(X, Y)$ (For the rest of this question, "shortest path" is always in terms of *length*, not number of edges). You would like to run a search algorithm to find the shortest way to get from A to B (assume no ties).

Suppose C is the capital, and thus you know the shortest paths from city C to every other city, and you would like to be able to use this information.

Let $path_{opt}(X \to Y)$ denote the shortest path from X to Y and let $cost(X, Y)$ denote the cost of the shortest path between cities X and Y. Let $[path(X \to Y), path(Y \to Z)]$ denote the concatenation.

(a) [5 pts] Suppose the distance along any edge is 1. You decide to initialize the queue with A, plus a list of all cities X, with $path(A \to X) = [path_{opt}(A \to C), path_{opt}(C \to X)]$. You run **BFS** with this initial queue (sorted in order of path length). Which of the following is correct? (Select all that apply)

- ☐ You always expand the exact same nodes as you would have if you ran standard BFS.
- ☐ You might expand a different set of nodes, but still find the shortest path.
- ☐ You might expand a different set of nodes, and find the sub-optimal path.

(b) [5 pts] You decide to initialize priority queue with A, plus a list of all cities X, with $path(A \to X) = [path_{opt}(A \to C), path_{opt}(C \to X)]$, and $cost(A, X) = cost(A, C) + cost(C, X)$. You run **UCS** with this initial priority queue. Which of the following is correct? (Select all that apply)

- ☐ You always expand the exact same nodes as you would have if you ran standard UCS.
- ☐ You might expand a different set of nodes, but still find the shortest path.
- ☐ You might expand a different set of nodes, and find the sub-optimal path.

Q4. [10 pts] Bayes Net Inference

A plate representation is useful for capturing replication in Bayes Nets. For example, Figure 1(a) is an equivalent representation of Figure 1(b). The N in the lower right corner of the plate stands for the number of replica.

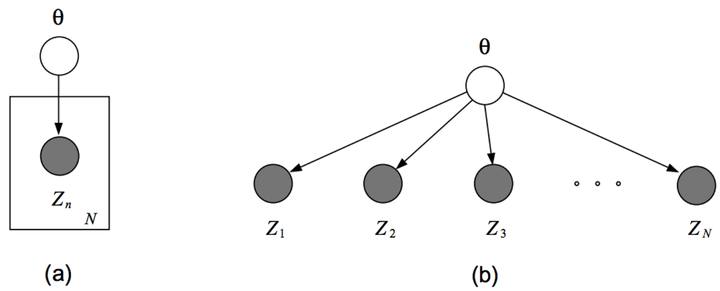


Figure 1

Now consider the Bayes Net in Figure 2. We use $X_{1:N}$ as shorthand for $(X_1, \cdots, X_N)$. We would like to compute the query $P(X_{1:N}|Y_{1:N} = y_{1:N})$. Assume all variables are binary.

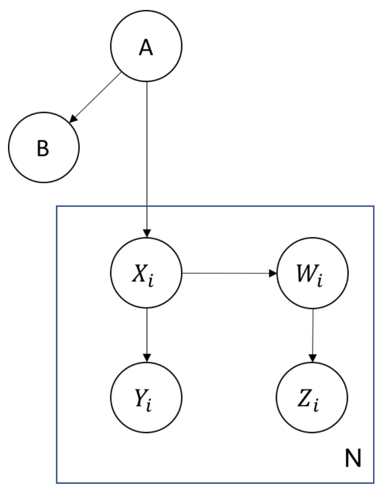


Figure 2

(a) [2 pts] What is the number of rows in the largest factor generated by *inference by enumeration*, for this query?

○ 2^{2N} ○ 2^{3N} ○ 2^{2N+2} ○ 2^{3N+2}

(b) [4 pts] Mark all of the following variable elimination orderings that are optimal for calculating the answer for the query $P(X_{1:N}|Y_{1:N} = y_{1:N})$. (A variable elimination ordering is optimal if the **largest** factors generated is smallest among all possible elimination orderings).

☐ $Z_1, \cdots, Z_N, W_1, \cdots, W_N, B, A$
☐ $W_1, \cdots, W_N, Z_1, \cdots, Z_N, B, A$
☐ $A, B, W_1, \cdots, W_N, Z_1, \cdots, Z_N$
☐ $A, B, Z_1, \cdots, Z_N, W_1, \cdots, W_N$

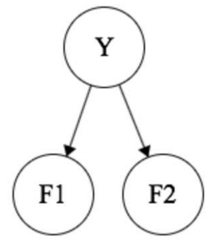
(c) [4 pts] Which of the following variables can be deleted before running variable elimination, without affecting the inference result? Deleting a variable means not putting its CPT in our initial set of factors when starting the algorithm.

☐ W_1 ☐ Z_1 ☐ A ☐ B ☐ None

Q5. [18 pts] Naive Bayes

(a) We use a Naive Bayes classifier to differentiate between Pacmen and Ghosts, trained on the following:

F_1	F_2	Y
0	1	Ghost
1	0	Ghost
0	0	Pac
1	1	Pac



Assume that the distributions generated from these samples perfectly estimate the CPTs. Given features f_1, f_2, we predict $\hat{Y} \in \{Ghost, Pacman\}$ using the Naive Bayes decision rule. If $P(Y = Ghost|F_1 = f_1, F_2 = f_2) = 0.5$, assign $\hat{Y}$ based on flipping a fair coin.

(i) [3 pts] Compute the table $P(\hat{Y}|Y)$.
Value $P(\hat{Y} = Ghost|Y = Ghost)$ is the probability of correctly classifying a *Ghost*, while $P(\hat{Y} = Pacman|Y = Ghost)$ is the probability of confusing a *Ghost* for a *Pacman*.

$P(\hat{Y}\|Y)$	$\hat{Y} = Ghost$	$\hat{Y} = Pacman$
$Y = Ghost$		
$Y = Pacman$		

For each modification below, recompute table $P(\hat{Y}|Y)$. **The modifications for each part are separate, and do not accumulate.**

(ii) [3 pts] Add extra feature $F_3 = F_1 + F_2$, and modify the Naive Bayes classifier appropriately.

$P(\hat{Y}\|Y)$	$\hat{Y} = Ghost$	$\hat{Y} = Pacman$
$Y = Ghost$		
$Y = Pacman$		

(iii) [3 pts] Add extra feature $F_3 = F_1 \times F_2$, and modify the Naive Bayes classifier appropriately.

$P(\hat{Y}\|Y)$	$\hat{Y} = Ghost$	$\hat{Y} = Pacman$
$Y = Ghost$		
$Y = Pacman$		

(iv) [3 pts] Add extra feature $F_3 = F_1 - F_2$, and modify the Naive Bayes classifier appropriately.

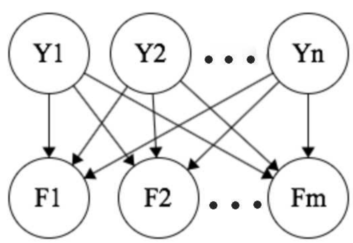
$P(\hat{Y}\|Y)$	$\hat{Y} = Ghost$	$\hat{Y} = Pacman$
$Y = Ghost$		
$Y = Pacman$		

(v) [3 pts] Perform Laplace Smoothing with $k = 1$.

$P(\hat{Y}\|Y)$	$\hat{Y} = Ghost$	$\hat{Y} = Pacman$
$Y = Ghost$		
$Y = Pacman$		

(b) [3 pts] Now, we reformulate the Naive Bayes classifier so that it can choose more than one class. For example, if we are choosing which genre a book is, we want the ability to say that a romantic comedy is *both* a romance and a comedy.

To do this, we have multiple label nodes $Y = \{Y_1 ... Y_n\}$ which all point to all features $F = \{F_1 ... F_m\}$.



Select all of the following expressions which are valid Naive Bayes classification rules, i.e., equivalent to $\arg\max_{Y_1...Y_n} P(Y_1, Y_2, ..., Y_n | F_1, F_2, ..., F_m)$:

☐ $\arg\max_{Y_1...Y_n} \prod_i^n \left[P(Y_i) \prod_j^m P(F_j|Y_i) \right]$

☐ $\arg\max_{Y_1...Y_n} \prod_i^n \left[P(Y_i) \prod_j^m P(F_j|Y_1...Y_n) \right]$

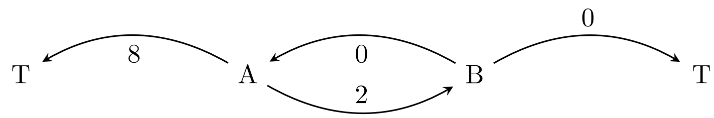
☐ $\arg\max_{Y_1...Y_n} \prod_i^n [P(Y_i)] \prod_j^m [P(F_j|Y_1...Y_n)]$

☐ $\prod_i^n \left[\arg\max_{Y_i} \left\{ P(Y_i) \prod_j^m P(F_j|Y_i) \right\} \right]$

☐ $\prod_i^n \left[\arg\max_{Y_i} \left\{ P(Y_i) \prod_j^m P(F_j|Y_1...Y_n) \right\} \right]$

Q6. [16 pts] MDP: Left or Right

Consider the following MDP:



The state space $\mathcal{S}$ and action space $\mathcal{A}$ are

$$\mathcal{S} = \{A, B, T\}$$

$$\mathcal{A} = \{\texttt{left}, \texttt{right}\}$$

where T denotes the terminal state (both T states are the same). When in a terminal state, the agent has no more action and gets no more reward. In non-terminal states, the agent can only go left or right, but their action only succeeds (goes in the intended direction) with probability p. If their action fails, then they go the opposite direction. The numbers on the arrows denote the reward associated with going from one state to another.

For example, at state A taking action `left`:

- with probability p, the next state will be T and the agent will get a reward of 8. The episode is then terminated.
- with probability $1 - p$, the next state will be B and the reward will be 2.

For this problem, the discount factor γ is 1. Let π_p^* be the optimal policy, which may or may not depend on the value of p. Let $Q^{\pi_p^*}$ and $V^{\pi_p^*}$ be the corresponding Q and V functions of π_p^*.

(a) [1 pt] If $p = 1$, what is π_p^*? (Select one)

- ◯ $\pi_p^*(A) = \texttt{left}$, $\pi_p^*(B) = \texttt{left}$
- ◯ $\pi_p^*(A) = \texttt{left}$, $\pi_p^*(B) = \texttt{right}$
- ◯ $\pi_p^*(A) = \texttt{right}$, $\pi_p^*(B) = \texttt{left}$
- ◯ $\pi_p^*(A) = \texttt{right}$, $\pi_p^*(B) = \texttt{right}$

(b) [1 pt] If $p = 0$, what is $\pi_p^*(A)$? (Select one)

- ◯ $\pi_p^*(A) = \texttt{left}$, $\pi_p^*(B) = \texttt{left}$
- ◯ $\pi_p^*(A) = \texttt{left}$, $\pi_p^*(B) = \texttt{right}$
- ◯ $\pi_p^*(A) = \texttt{right}$, $\pi_p^*(B) = \texttt{left}$
- ◯ $\pi_p^*(A) = \texttt{right}$, $\pi_p^*(B) = \texttt{right}$

(c) [5 pts] Suppose $\pi_p^*(A) = \texttt{left}$. Which of the following statements **must** be true? (Select all that apply)

Hint: Don't forget that if $x = y$, then $x \geq y$ and $x \leq y$.

- ☐ $Q^{\pi_p^*}(A, \texttt{left}) \leq Q^{\pi_p^*}(A, \texttt{right})$
- ☐ $Q^{\pi_p^*}(A, \texttt{left}) \geq Q^{\pi_p^*}(A, \texttt{right})$
- ☐ $Q^{\pi_p^*}(A, \texttt{left}) = Q^{\pi_p^*}(A, \texttt{right})$
- ☐ $V^{\pi_p^*}(A) \leq V^{\pi_p^*}(B)$
- ☐ $V^{\pi_p^*}(A) \geq V^{\pi_p^*}(B)$
- ☐ $V^{\pi_p^*}(A) = V^{\pi_p^*}(B)$
- ☐ $V^{\pi_p^*}(A) \leq Q^{\pi_p^*}(A, \texttt{left})$
- ☐ $V^{\pi_p^*}(A) \geq Q^{\pi_p^*}(A, \texttt{left})$
- ☐ $V^{\pi_p^*}(A) = Q^{\pi_p^*}(A, \texttt{left})$
- ☐ $V^{\pi_p^*}(A) \leq Q^{\pi_p^*}(A, \texttt{right})$
- ☐ $V^{\pi_p^*}(A) \geq Q^{\pi_p^*}(A, \texttt{right})$
- ☐ $V^{\pi_p^*}(A) = Q^{\pi_p^*}(A, \texttt{right})$

(d) Assume $p \geq 0.5$ below.

(i) [3 pts] $V^*(B) = \alpha V^*(A) + \beta$. Find α and β in terms of p.

- $\alpha =$ __________________________

- $\beta =$ __________________________

(ii) [3 pts] $Q^{\pi_p^*}(A, \texttt{left}) = \alpha V^*(B) + \beta$. Find α and β in terms of p.

- $\alpha =$ __________________________

- $\beta =$ __________________________

(iii) [3 pts] $Q^{\pi_p^*}(A, \texttt{right}) = \alpha V^*(B) + \beta$. Find α and β in terms of p.

- $\alpha =$ __________________________

- $\beta =$ __________________________

Q7. [22 pts] Take Actions

An agent is acting in the following gridworld MDP, with the following characteristics.

- Discount factor $\gamma < 1$.
- Agent gets reward $R > 0$ for **entering** the terminal state T, and 0 reward for all other transitions.
- When in terminal state T, the agent has no more action and gets no more reward.
- In non-terminal states $\{A, B, C\}$, the agent can take an action $\{Up, Down, Left, Right\}$.
- Assume perfect transition dynamics. For example, taking action *Right* at state A will always result in state C in the next time step.
- If the agent hits an edge, it stays in the same state in the next time step. For example, after taking action *Right* at C, the agent remains in state C.

B	T
A	C

(a) **(i)** [3 pts] What are all the optimal deterministic policies? Each cell should contain a single action $\{Up, Down, Left, Right\}$. Each row corresponds to a different optimal policy. You may not need all rows.

State	A	B	C
Optimal policy 1			
Optimal policy 2 (if needed)			
Optimal policy 3 (if needed)			

(ii) [2 pts] Suppose the agent uniformly randomly chooses between the optimal policies in (i). In other words, at each state, the agent picks randomly between the actions in the corresponding column with equal probability. The agent's location at each time step is then a Markov process where state $X_t \in \{A, B, C, T\}$. Fill in the following transition probabilities for the Markov process.

- $P(X_{t+1} = B | X_t = A) =$ ______________________
- $P(X_{t+1} = A | X_t = B) =$ ______________________
- $P(X_{t+1} = T | X_t = C) =$ ______________________

(b) Suppose the agent is acting in the same gridworld as above, but does not get to observe their exact state X_t. Instead, the agent only observes $O_t \in \{black, green, pink\}$. The observation probability as a function of the state $P(O_t|X_t)$ is specified in the table below. This becomes a partially-observable Markov decision process (POMDP). The agent is equally likely to start in non-terminal states $\{A, B, C\}$.

B 0.5, black 0.5, green	T
A 0.5, black 0.5, pink	C 0.5, pink 0.5, green

(i) [3 pts] If the agent can only act based on its current observation, what are all deterministic optimal policies? You may not need all rows.

	Black	Green	Pink
Optimal policy 1			
Optimal policy 2 (if needed)			
Optimal policy 3 (if needed)			

(ii) [3 pts] Suppose that the agent follows the policy $\pi(\text{Black}) = \text{Right}$, $\pi(\text{Green}) = \text{Right}$, and $\pi(\text{Pink}) = Up$. Let $V(S)$ be the agent's expected reward from state S. Your answer should be in terms of γ and R. Note that $V(S)$ is the expected value before we know the observation, so you must consider all possible observations at state S.

- V(A) = ________________________
- V(B) = ________________________
- V(C) = ________________________

Now suppose that the agent's policy can also depend on all past observations and actions. Assume that when the agent is starting (and has no past observations), it behaves the same as the policy in the previous part: $\pi([\text{Black}]) = \text{Right}$, $\pi([\text{Green}]) = \text{Right}$, $\pi([\text{Pink}]) = \text{Up}$. In all cases where the agent has more than one observation (for example, observed Pink in the previous time step and now observes Green), π acts optimally.

(iii) [3 pts] For each of the following sequences of two observations, write the optimal action that the policy π would take.

Black Pink	Black Green	Green Pink	Green Green	Pink Black	Pink Green

(iv) [3 pts] In this part only, let $V(S)$ refer to the expected sum of discounted rewards following π if we start from state S (and thus have no previous observations yet). As in the previous part, this is the expected value before knowing the observation, so you must consider all possible observations at S.

Hint: since π now depends on sequences of observations, the way we act at states after the first state may be different, and this affects the value at the first state.

- V(A) = ________________________
- V(B) = ________________________
- V(C) = ________________________

(c) Boba POMDP May is a venture capitalist who knows that Berkeley students love boba. She is picking between investing in Sharetea or Asha. If she invests in the better one, she will make a profit of \$1000. If she invests in the worse one, she will make no money.

At the start, she believes both Asha and Sharetea have an equal chance of being better. However, she can pay to have students taste test. At each time step, she can either choose to invest or to pay for a student taste test. Each student has a $p = 0.9$ probability of picking the correct place (independent of other students).

(i) [1 pt] What is the expected profit if May invests optimally after one (free) student test?

____

(ii) [1 pt] If she had to invest after one student test, what is the highest May should pay for the test?

____

(iii) [1 pt] Suppose after n student tests, it turns out that all students have chosen the same store. What is her expected profit after after observing these n student tests?

- ◯ $1000(0.9^n)$
- ◯ $1000(\frac{0.9^n}{0.9^n+0.1^n})$
- ◯ $1000(0.9^n - 0.1^n)$
- ◯ $1000(\frac{0.9^n-0.1^n}{0.9^n})$
- ◯ $1000(\frac{0.9^n-0.1^n}{0.9^n+0.1^n})$
- ◯ $1000(1 - 0.1^n)$

(iv) [2 pts] How many tests should May pay for if each one costs \$100?

Hint: Think about the maximum possible value of information. How does this compare to the expected value of information?

____

Q8. [18 pts] Tracking a cyclist

We are trying to track cyclists as they move around a self-driving car. The car is equipped with 4 "presence detectors" corresponding to:

- Front of the car (F),
- Back of the car (B),
- Left side of the car (L),
- Right side of the car (R).

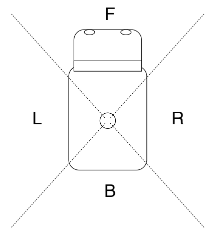


Figure 3: Autonomous vehicle and detection zones

Unfortunately, the detectors are not perfect and feature the following conditional probabilities for detection $D \in \{0, 1\}$ ("no detection" or "detection", respectively) given cyclist presence $C \in \{0, 1\}$ ("no cyclist" or "cyclist", respectively).

Front detector

$P_F(D\|C)$	$d = 1$	$d = 0$
$c = 1$	0.8	0.2
$c = 0$	0.1	0.9

Back detector

$P_B(D\|C)$	$d = 1$	$d = 0$
$c = 1$	0.6	0.4
$c = 0$	0.4	0.6

Left & Right detectors

$P_L(D\|C) = P_R(D\|C)$	$d = 1$	$d = 0$
$c = 1$	0.7	0.3
$c = 0$	0.2	0.8

(a) Detection and dynamics

(i) [2 pts] If you could freely choose any detector to equip all four detection zones, which one would be best?

○ The front detector. ○ The detector at the back. ○ The left/right detector.

Dynamics: We have measured the following transition probabilities for cyclists moving around the car when driving. Assume any dynamics are Markovian. Variable $X_t \in \{f, l, r, b\}$ denotes the location of the cyclist at time t, and can be in front, left, right, or back of the car.

$P(X_{t+1}\|X_t)$	$X_{t+1} = f$	$X_{t+1} = l$	$X_{t+1} = r$	$X_{t+1} = b$
$X_t = f$	p_{ff}	p_{fl}	p_{fr}	p_{fb}
$X_t = l$	p_{lf}	p_{ll}	p_{lr}	p_{lb}
$X_t = r$	p_{rf}	p_{rl}	p_{rr}	p_{rb}
$X_t = b$	p_{bf}	p_{bl}	p_{br}	p_{bb}

(ii) [3 pts] Which criterion does this table have to satisfy for it to be a well defined CPT? (Select all that apply).

☐ Each row should sum to 1. ☐ Each column should sum to 1. ☐ The table should sum to 1.

(b) Let's assume that we have been given a sequence of observations $d_1, d_2, \ldots, d_t$ and computed the posterior probability $P(X_t|d_1, d_2, \ldots, d_t)$, which we represent as a four-dimensional vector.

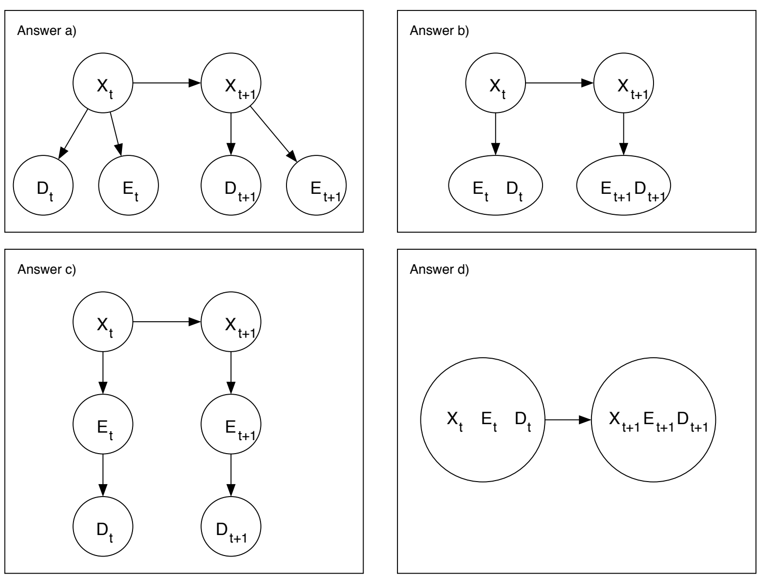
(i) [3 pts] What is vector $P(X_{t+1}|d_1, d_2, \ldots, d_t)$ as a function of $P(X_{t+1}|X_t)$ (a 4×4 matrix written above) and $P(X_t|d_1, d_2, \ldots, d_t)$?

____

(ii) [2 pts] What is the computational complexity of computing $P(X_t|D_1 = d_1, D_2 = d_2, \ldots, D_t = d_t)$ as a function of t and the number of states S (using big O notation)?

____

(c) **(i)** [4 pts] We now add a radar to the system (random variable $E \in \{f, l, r, b\}$). Assuming the detection by this device is independent from what happens with the pre-existing detectors, which of the probabilistic models *could* you use? If several variables are in the same node, the node represents a tuple of random variables, which itself is a random variable.



Select all that apply.

☐ a) ☐ b) ☐ c) ☐ d)

(ii) [4 pts] ERRATUM: Which of the following values for Z are correct?

$$P(X_{t+1}|D_1, \ldots, D_{t+1}, E_1, \ldots, E_{t+1}) = \sum_{x=f,l,r,b} \frac{Z \cdot P(X_t = x|D_1, \ldots, D_t, E_1, \ldots, E_t) \cdot P(X_{t+1}|X_t = x)}{P(D_{t+1}, E_{t+1}|D_1, \ldots, D_t, E_1, \ldots, E_t)}.$$

☐ $Z = P(E_{t+1} D_{t+1}|X_{t+1}, X_t, D_1, \ldots, D_{t+1}, E_1, \ldots, E_{t+1})$

☐ $Z = P(E_{t+1}|X_{t+1})P(D_{t+1}|X_{t+1})$

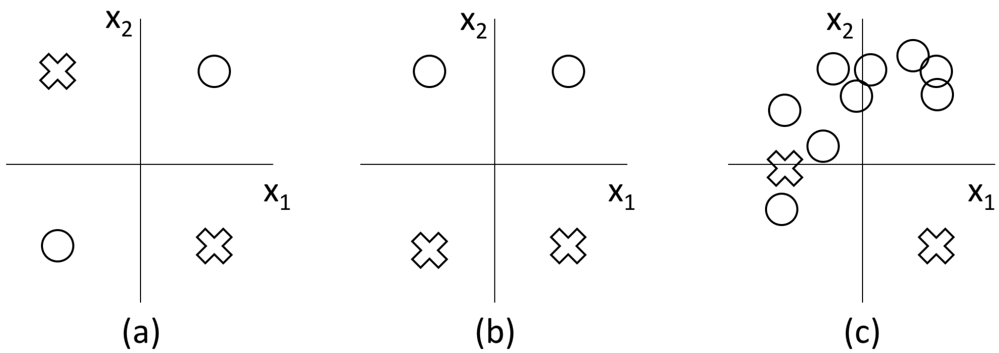
☐ $Z = P(E_{t+1}|X_t)P(D_{t+1}|X_t)$

☐ $Z = P(E_{t+1}|E_t)P(D_{t+1}|D_t)$

- ☐ $Z = P(E_{t+1}, D_{t+1} | X_{t+1})$
- ☐ $Z = P(E_{t+1}, D_{t+1} | X_t)$

Q9. [23 pts] Deep Learning

(a) [5 pts] **Data Separability**



The plots above show points in feature space (x_1, x_2), also referred to as feature vectors $\mathbf{x} = [x_1 \;\; x_2]^T$.

For each of the following, we will define a function $h(\mathbf{x})$ as a composition of some functions f_i and g_i. For each one, consider the decision rule

$$y(\mathbf{x}) = \begin{cases} \times & h(\mathbf{x}) \geq 0 \\ \bigcirc & h(\mathbf{x}) < 0. \end{cases}$$

Under each composition of functions h, select the datasets for which there exist some **linear** functions f_i and some **nonlinear** functions g_i such that the corresponding decision rule perfectly classifies the data. (Select all that apply)

(i) $h(\mathbf{x}) = f_1(\mathbf{x})$

(a) ☐ (b) ☐ (c) ☐

(ii) $h(\mathbf{x}) = f_2(f_1(\mathbf{x}))$

(a) ☐ (b) ☐ (c) ☐

(iii) $h(\mathbf{x}) = f_2(g_1(f_1(\mathbf{x})))$

(a) ☐ (b) ☐ (c) ☐

(iv) $h(\mathbf{x}) = f_4(f_3(f_2(f_1(\mathbf{x}))))$

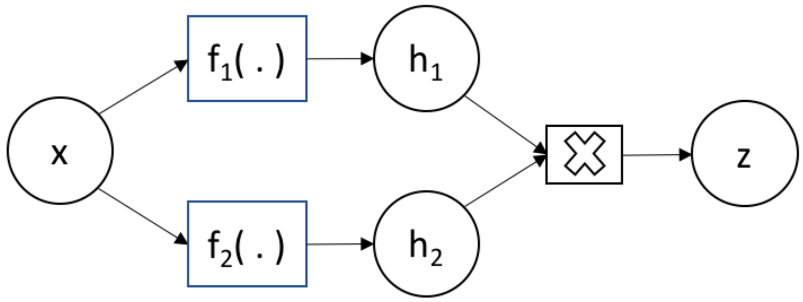
(a) ☐ (b) ☐ (c) ☐

(v) $h(\mathbf{x}) = g_2(g_1(\mathbf{x}))$

(a) ☐ (b) ☐ (c) ☐

(b) Backpropagation Below is a deep network with input x. Values x, h_1, h_2, z are all scalars.

$$h_1 = f_1(x), h_2 = f_2(x), z = h_1 h_2 \tag{1}$$



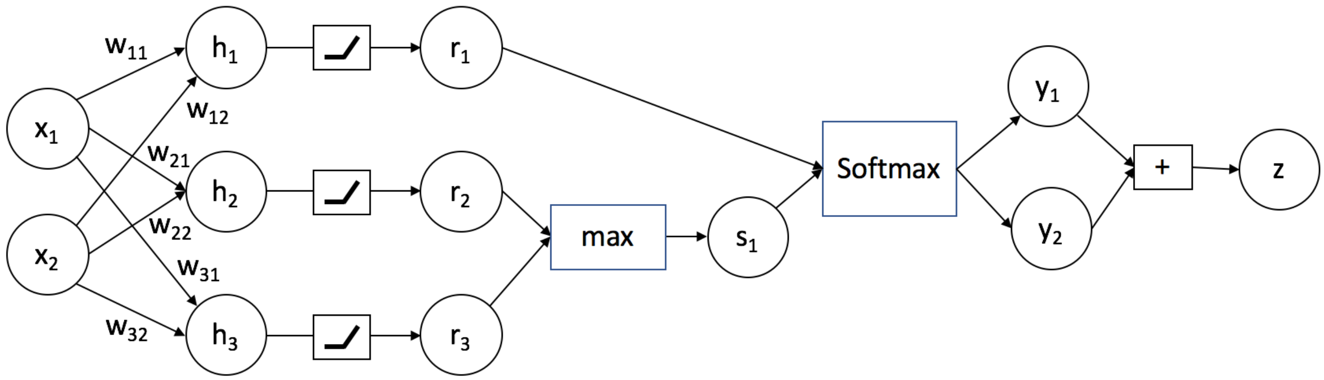
Derive the following gradients in terms of $x, h_1, h_2, \frac{\partial f_1}{\partial x}, \frac{\partial f_2}{\partial x}$.

(i) [1 pt] Derive $\frac{\partial z}{\partial h_1}$ ________________________

(ii) [1 pt] Derive $\frac{\partial z}{\partial h_2}$ ________________________

(iii) [3 pts] Derive $\frac{\partial z}{\partial x}$ ________________________

(c) Deep Network Below is a deep network with inputs x_1, x_2. The internal nodes are computed below. All variables are scalar values.



$$
\begin{aligned}
&h_1 = w_{11}x_1 + w_{12}x_2 \qquad && h_2 = w_{21}x_1 + w_{22}x_2 \qquad && h_3 = w_{31}x_1 + w_{32}x_2 \\
&r_1 = \max(h_1, 0) && r_2 = \max(h_2, 0) && r_3 = \max(h_3, 0) \\
& & s_1 = \max(r_2, r_3) & & \\
&y_1 = \frac{\exp(r_1)}{\exp(r_1) + \exp(s_1)} && y_2 = \frac{\exp(s_1)}{\exp(r_1) + \exp(s_1)} & & \\
& & z = y_1 + y_2 & &
\end{aligned} \tag{2}
$$

(i) [5 pts] **Forward propagation** Now, given $x_1 = 1, x_2 = -2, w_{11} = 6, w_{12} = 2, w_{21} = 4, w_{22} = 7,$ $w_{31} = 5, w_{32} = 1$, and the same values for x_1, x_2 above, compute the values of the internal nodes. Please simplify any fractions.

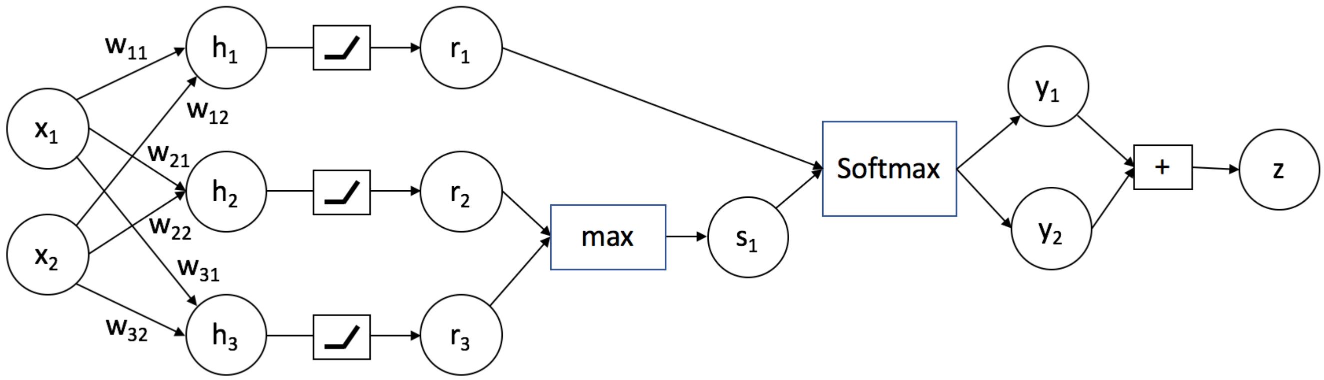
h_1	h_2	h_3	r_1	r_2

r_3	s	y_1	y_2	z

(ii) [2 pts] **Bounds on variables.**

Find the tightest bounds on y_1. ________

Find the tightest bounds on z. ________



$$
\begin{aligned}
&h_1 = w_{11}x_1 + w_{12}x_2 \qquad h_2 = w_{21}x_1 + w_{22}x_2 \qquad h_3 = w_{31}x_1 + w_{32}x_2 \\
&r_1 = \max(h_1, 0) \qquad r_2 = \max(h_2, 0) \qquad r_3 = \max(h_3, 0) \\
&s_1 = \max(r_2, r_3) \\
&y_1 = \frac{\exp(r_1)}{\exp(r_1) + \exp(s_1)} \qquad y_2 = \frac{\exp(s_1)}{\exp(r_1) + \exp(s_1)} \\
&z = y_1 + y_2
\end{aligned} \tag{3}
$$

(iii) [6 pts] **Backpropagation** Compute the following gradients analytically. The answer should be an expression of any of the nodes in the network $(x_1, x_2, h_1, h_2, h_3, r_1, r_2, r_3, s_1, y_1, y_2, z)$ or weights $w_{11}, w_{12}, w_{21}, w_{22}, w_{31}, w_{32}$. Hint: Recall that for functions of the form $g(x) = \frac{1}{1+exp(a-x)}$, $\frac{\partial g}{\partial x} = g(x)\,(1 - g(x))$. Also, your answer may be a constant or a piecewise function.

$\frac{\partial h_1}{\partial w_{12}}$	$\frac{\partial h_1}{\partial x_1}$	$\frac{\partial r_1}{\partial h_1}$	$\frac{\partial y_1}{\partial r_1}$

$\frac{\partial y_1}{\partial s_1}$	$\frac{\partial z}{\partial y_1}$	$\frac{\partial z}{\partial x_1}$	$\frac{\partial s_1}{\partial r_2}$

THIS PAGE IS INTENTIONALLY LEFT BLANK