CS 188 Spring 2015 Introduction to Artificial Intelligence Midterm 2 V1

- You have approximately 2 hours and 50 minutes.
- The exam is closed book, closed calculator, and closed notes except your one-page crib sheet.
- Mark your answers ON THE EXAM ITSELF. If you are not sure of your answer you may wish to provide a *brief* explanation or show your work.
- For multiple choice questions,
  - ☐ means mark **all options** that apply
  - ◯ means mark a **single choice**
- There are multiple versions of the exam. For fairness, this does not impact the questions asked, only the ordering of options within a given question.

First name	
Last name	
SID	
edX username	

First and last name of student to your left	
First and last name of student to your right	

For staff use only:

Q1.	Probability	/14
Q2.	Bayes' Nets: Representation	/8
Q3.	Bayes' Nets: Independence	/8
Q4.	Bayes' Nets: Inference	/6
Q5.	Bayes' Nets: Sampling	/10
Q6.	VPI	/13
Q7.	HMM: Where is the Car?	/13
Q8.	Particle Filtering: Where are the Two Cars?	/11
Q9.	Naive Bayes MLE	/7
Q10.	Perceptron	/10
	Total	/100

THIS PAGE IS INTENTIONALLY LEFT BLANK

Q1. [14 pts] Probability

(a) For the following questions, you will be given a set of probability tables and a set of conditional independence assumptions. Given these tables and independence assumptions, write an expression for the requested probability tables. Keep in mind that your expressions cannot contain any probabilities other than the given probability tables. If it is not possible, mark "Not possible."

(i) [1 pt] Using probability tables $\mathbf{P}(\mathbf{A}), \mathbf{P}(\mathbf{A} \mid \mathbf{C}), \mathbf{P}(\mathbf{B} \mid \mathbf{C}), \mathbf{P}(\mathbf{C} \mid \mathbf{A}, \mathbf{B})$ and no conditional independence assumptions, write an expression to calculate the table $\mathbf{P}(\mathbf{A}, \mathbf{B} \mid \mathbf{C})$.

$\mathbf{P}(\mathbf{A}, \mathbf{B} \mid \mathbf{C}) =$ _____ ● Not possible.

(ii) [1 pt] Using probability tables $\mathbf{P}(\mathbf{A}), \mathbf{P}(\mathbf{A} \mid \mathbf{C}), \mathbf{P}(\mathbf{B} \mid \mathbf{A}), \mathbf{P}(\mathbf{C} \mid \mathbf{A}, \mathbf{B})$ and no conditional independence assumptions, write an expression to calculate the table $\mathbf{P}(\mathbf{B} \mid \mathbf{A}, \mathbf{C})$.

$\mathbf{P}(\mathbf{B} \mid \mathbf{A}, \mathbf{C}) =$ $\frac{P(A)\ P(B|A)\ P(C|A,B)}{\sum_b P(A)\ P(B|A)\ P(C|A,B)}$ ○ Not possible.

(iii) [1 pt] Using probability tables $\mathbf{P}(\mathbf{A} \mid \mathbf{B}), \mathbf{P}(\mathbf{B}), \mathbf{P}(\mathbf{B} \mid \mathbf{A}, \mathbf{C}), \mathbf{P}(\mathbf{C} \mid \mathbf{A})$ and conditional independence assumption $\mathbf{A} \perp\!\!\!\perp \mathbf{B}$, write an expression to calculate the table $\mathbf{P}(\mathbf{C})$.

$\mathbf{P}(\mathbf{C}) =$ $\sum_a P(A \mid B)\ P(C \mid A)$ ○ Not possible.

(iv) [1 pt] Using probability tables $\mathbf{P}(\mathbf{A} \mid \mathbf{B}, \mathbf{C}), \mathbf{P}(\mathbf{B}), \mathbf{P}(\mathbf{B} \mid \mathbf{A}, \mathbf{C}), \mathbf{P}(\mathbf{C} \mid \mathbf{B}, \mathbf{A})$ and conditional independence assumption $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$, write an expression for $\mathbf{P}(\mathbf{A}, \mathbf{B}, \mathbf{C})$.

$\mathbf{P}(\mathbf{A}, \mathbf{B}, \mathbf{C}) =$ _____ ● Not possible.

(b) For each of the following equations, select the *minimal set* of conditional independence assumptions necessary for the equation to be true.

(i) [1 pt] $\mathbf{P}(\mathbf{A}, \mathbf{C}) = \mathbf{P}(\mathbf{A} \mid \mathbf{B})\ \mathbf{P}(\mathbf{C})$

- ☐ $A \perp\!\!\!\perp B \mid C$
- ☐ $B \perp\!\!\!\perp C \mid A$
- ■ $A \perp\!\!\!\perp C$
- ☐ $A \perp\!\!\!\perp C \mid B$
- ☐ $B \perp\!\!\!\perp C$
- ■ $A \perp\!\!\!\perp B$
- ☐ No independence assumptions needed.

(ii) [1 pt] $\mathbf{P}(\mathbf{A} \mid \mathbf{B}, \mathbf{C}) = \frac{\mathbf{P}(\mathbf{A})\ \mathbf{P}(\mathbf{B}|\mathbf{A})\ \mathbf{P}(\mathbf{C}|\mathbf{A})}{\mathbf{P}(\mathbf{B}|\mathbf{C})\ \mathbf{P}(\mathbf{C})}$

- ☐ $A \perp\!\!\!\perp C$
- ☐ $A \perp\!\!\!\perp B \mid C$
- ■ $B \perp\!\!\!\perp C \mid A$
- ☐ $A \perp\!\!\!\perp B$
- ☐ $A \perp\!\!\!\perp C \mid B$
- ☐ $B \perp\!\!\!\perp C$
- ☐ No independence assumptions needed.

(iii) [1 pt] $\mathbf{P}(\mathbf{A}, \mathbf{B}) = \sum_{\mathbf{c}} \mathbf{P}(\mathbf{A} \mid \mathbf{B}, \mathbf{c})\ \mathbf{P}(\mathbf{B} \mid \mathbf{c})\ \mathbf{P}(\mathbf{c})$

- ☐ $B \perp\!\!\!\perp C \mid A$
- ☐ $B \perp\!\!\!\perp C$
- ☐ $A \perp\!\!\!\perp C \mid B$
- ☐ $A \perp\!\!\!\perp C$
- ☐ $A \perp\!\!\!\perp B$
- ☐ $A \perp\!\!\!\perp B \mid C$
- ■ No independence assumptions needed.

(iv) [1 pt] $\mathbf{P(A, B \mid C, D) = P(A \mid C, D)\ P(B \mid A, C, D)}$

- ☐ $A \perp\!\!\!\perp B \mid D$
- ☐ $C \perp\!\!\!\perp D \mid A$
- ☐ $C \perp\!\!\!\perp D \mid B$
- ☐ $C \perp\!\!\!\perp D$
- ☐ $A \perp\!\!\!\perp B$
- ☐ $A \perp\!\!\!\perp B \mid C$
- ■ No independence assumptions needed.

(c) **(i)** [2 pts] Mark **all** expressions that are equal to $\mathbf{P(A \mid B)}$, **given no independence assumptions**.

- ■ $\sum_c P(A, c \mid B)$
- ☐ $\frac{P(A,C|B)}{P(C|B)}$
- ☐ $\sum_c P(A \mid B, c)$
- ■ $\frac{\sum_c P(A,B,c)}{\sum_c P(B,c)}$
- ☐ $\frac{P(B|A)\ P(A|C)}{\sum_c P(B,c)}$
- ☐ $\frac{P(A|C,B)\ P(C|A,B)}{P(C|B)}$
- ☐ None of the provided options.

(ii) [2 pts] Mark **all** expressions that are equal to $\mathbf{P(A, B, C)}$, given that $\mathbf{A \perp\!\!\!\perp B}$.

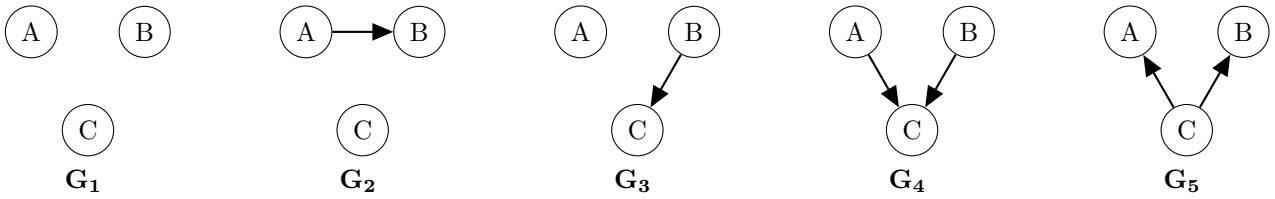
- ■ $P(A)\ P(B)\ P(C \mid A, B)$
- ☐ $P(C)\ P(A \mid C)\ P(B \mid C)$
- ■ $P(A)\ P(B \mid A)\ P(C \mid A, B)$
- ☐ $P(A \mid C)\ P(C \mid B)\ P(B)$
- ☐ $P(A)\ P(C \mid A)\ P(B \mid C)$
- ■ $P(A, C)\ P(B \mid A, C)$
- ☐ None of the provided options.

(iii) [2 pts] Mark **all** expressions that are equal to $\mathbf{P(A, B \mid C)}$, given that $\mathbf{A \perp\!\!\!\perp B \mid C}$.

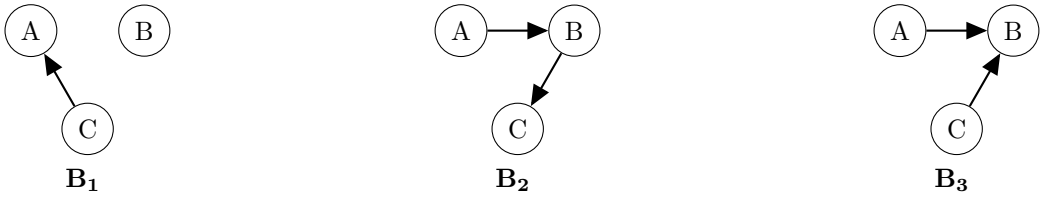
- ■ $P(A \mid C)\ P(B \mid C)$
- ☐ $\frac{\sum_c P(A,B,c)}{P(C)}$
- ☐ $\frac{P(C)\ P(B|C)\ P(A|C)}{P(C|A,B)}$
- ☐ $\frac{P(A)\ P(B|A)\ P(C|A,B)}{\sum_c P(A,B,c)}$
- ■ $\frac{P(C,A|B)\ P(B)}{P(C)}$
- ☐ $P(A \mid B)\ P(B \mid C)$
- ☐ None of the provided options.

Q2. [8 pts] Bayes' Nets: Representation

Assume we are given the following ten Bayes' nets, labeled $\mathbf{G_1}$ to $\mathbf{G_{10}}$:



Assume we are also given the following three Bayes' nets, labeled $\mathbf{B_1}$ to $\mathbf{B_3}$:



Before we go into the questions, let's enumerate all of the (conditional) independence assumptions encoded in all the Bayes' nets above. They are:

- $\mathbf{G_1}$: $A \perp\!\!\!\perp B; A \perp\!\!\!\perp B|C; A \perp\!\!\!\perp C; A \perp\!\!\!\perp C|B; B \perp\!\!\!\perp C; B \perp\!\!\!\perp C|A$
- $\mathbf{G_2}$: $A \perp\!\!\!\perp C; A \perp\!\!\!\perp C|B; B \perp\!\!\!\perp C; B \perp\!\!\!\perp C|A$
- $\mathbf{G_3}$: $A \perp\!\!\!\perp B; A \perp\!\!\!\perp B|C; A \perp\!\!\!\perp C; A \perp\!\!\!\perp C|B$
- $\mathbf{G_4}$: $A \perp\!\!\!\perp B$
- $\mathbf{G_5}$: $A \perp\!\!\!\perp B|C$
- $\mathbf{G_6}$: $A \perp\!\!\!\perp C|B$
- $\mathbf{G_7}$: $B \perp\!\!\!\perp C|A$
- $\mathbf{G_8}$: $A \perp\!\!\!\perp C|B$
- $\mathbf{G_9}$: $\emptyset$
- $\mathbf{G_{10}}$: $\emptyset$
- $\mathbf{B_1}$: $A \perp\!\!\!\perp B; A \perp\!\!\!\perp B|C; B \perp\!\!\!\perp C; B \perp\!\!\!\perp C|A$
- $\mathbf{B_2}$: $A \perp\!\!\!\perp C|B$
- $\mathbf{B_3}$: $A \perp\!\!\!\perp C$

(a) [2 pts] Assume we know that a joint distribution $\mathbf{d_1}$ (over $\mathbf{A}, \mathbf{B}, \mathbf{C}$) can be represented by Bayes' net $\mathbf{B_1}$. Mark all of the following Bayes' nets that are guaranteed to be able to represent $\mathbf{d_1}$.

☐ $\mathbf{G_1}$	☐ $\mathbf{G_2}$	☐ $\mathbf{G_3}$	■ $\mathbf{G_4}$	■ $\mathbf{G_5}$
☐ $\mathbf{G_6}$	■ $\mathbf{G_7}$	☐ $\mathbf{G_8}$	■ $\mathbf{G_9}$	■ $\mathbf{G_{10}}$

☐ None of the above.

Since $\mathbf{B_1}$ can represent $\mathbf{d_1}$, we know that $\mathbf{d_1}$ must satisfy the assumptions that $\mathbf{B_1}$ follows, which are: $A \perp\!\!\!\perp B; A \perp\!\!\!\perp B|C; B \perp\!\!\!\perp C; B \perp\!\!\!\perp C|A$. We cannot assume that $\mathbf{d_1}$ satisfies the other two assumptions, which are $A \perp\!\!\!\perp C$ and $A \perp\!\!\!\perp C|B$, and so a Bayes' net that makes at least one of these two extra assumptions will not be guaranteed to be able to represent $\mathbf{d_1}$. This eliminates the choices $\mathbf{G_1}, \mathbf{G_2}, \mathbf{G_3}, \mathbf{G_6}, \mathbf{G_8}$. The other choices $\mathbf{G_4}, \mathbf{G_5}, \mathbf{G_7}, \mathbf{G_9}, \mathbf{G_{10}}$ are guaranteed to be able to represent $\mathbf{d_1}$ because they do not make any additional independence assumptions that $\mathbf{B_1}$ makes.

(b) [2 pts] Assume we know that a joint distribution $\mathbf{d_2}$ (over $\mathbf{A}, \mathbf{B}, \mathbf{C}$) can be represented by Bayes' net $\mathbf{B_2}$. Mark all of the following Bayes' nets that are guaranteed to be able to represent $\mathbf{d_2}$.

☐ $\mathbf{G_1}$ ☐ $\mathbf{G_2}$ ☐ $\mathbf{G_3}$ ☐ $\mathbf{G_4}$ ☐ $\mathbf{G_5}$

■ $\mathbf{G_6}$ ☐ $\mathbf{G_7}$ ■ $\mathbf{G_8}$ ■ $\mathbf{G_9}$ ■ $\mathbf{G_{10}}$

☐ None of the above.

Since $\mathbf{B_2}$ can represent $\mathbf{d_2}$, we know that $\mathbf{d_2}$ must satisfy the assumptions that $\mathbf{B_2}$ follows, which is just: $A \perp\!\!\!\perp C|B$. We cannot assume that $\mathbf{d_2}$ satisfies any other assumptions, and so a Bayes' net that makes at least one other extra assumptions will not be guaranteed to be able to represent $\mathbf{d_2}$. This eliminates the choices $\mathbf{G_1}, \mathbf{G_2}, \mathbf{G_3}, \mathbf{G_4}, \mathbf{G_5}, \mathbf{G_7}$. The other choices $\mathbf{G_6}, \mathbf{G_8}, \mathbf{G_9}, \mathbf{G_{10}}$ are guaranteed to be able to represent $\mathbf{d_2}$ because they do not make any additional independence assumptions that $\mathbf{B_2}$ makes.

(c) [2 pts] Assume we know that a joint distribution $\mathbf{d_3}$ (over $\mathbf{A}, \mathbf{B}, \mathbf{C}$) ***cannot*** be represented by Bayes' net $\mathbf{B_3}$. Mark all of the following Bayes' nets that are guaranteed to be able to represent $\mathbf{d_3}$.

☐ $\mathbf{G_1}$ ☐ $\mathbf{G_2}$ ☐ $\mathbf{G_3}$ ☐ $\mathbf{G_4}$ ☐ $\mathbf{G_5}$

☐ $\mathbf{G_6}$ ☐ $\mathbf{G_7}$ ☐ $\mathbf{G_8}$ ■ $\mathbf{G_9}$ ■ $\mathbf{G_{10}}$

☐ None of the above.

Since $\mathbf{B_3}$ cannot represent $\mathbf{d_3}$, we know that $\mathbf{d_3}$ is unable to satisfy at least one of the assumptions that $\mathbf{B_3}$ follows. Since $\mathbf{B_3}$ only makes one independence assumption, which is $A \perp\!\!\!\perp C$, we know that $\mathbf{d_3}$ does not satisfy $A \perp\!\!\!\perp C$. However, we can't claim anything about whether or not $\mathbf{d_3}$ makes any of the other independence assumptions. $\mathbf{d_3}$ might not make any (conditional) independence assumptions at all, and so only the Bayes' nets that don't make any assumptions will be guaranteed to be able to represent $\mathbf{d_3}$. Hence, the answers are the fully connected Bayes' nets, which are $\mathbf{G_9}, \mathbf{G_{10}}$.

(d) [2 pts] Assume we know that a joint distribution $\mathbf{d_4}$ (over $\mathbf{A}, \mathbf{B}, \mathbf{C}$) can be represented by Bayes' nets $\mathbf{B_1}$, $\mathbf{B_2}$, and $\mathbf{B_3}$. Mark all of the following Bayes' nets that are guaranteed to be able to represent $\mathbf{d_4}$.

■ $\mathbf{G_1}$ ■ $\mathbf{G_2}$ ■ $\mathbf{G_3}$ ■ $\mathbf{G_4}$ ■ $\mathbf{G_5}$

■ $\mathbf{G_6}$ ■ $\mathbf{G_7}$ ■ $\mathbf{G_8}$ ■ $\mathbf{G_9}$ ■ $\mathbf{G_{10}}$

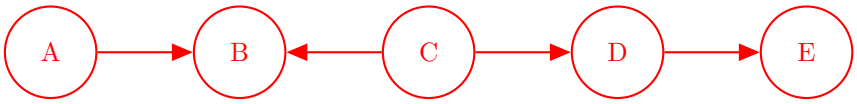
☐ None of the above.

Since $\mathbf{B_1}, \mathbf{B_2}, \mathbf{B_3}$ can represent $\mathbf{d_4}$, we know that $\mathbf{d_4}$ must satisfy the assumptions that $\mathbf{B_1}, \mathbf{B_2}, \mathbf{B_3}$ make. The union of assumptions made by these Bayes' nets are: $A \perp\!\!\!\perp B; A \perp\!\!\!\perp B|C; B \perp\!\!\!\perp C; B \perp\!\!\!\perp C|A, A \perp\!\!\!\perp C, A \perp\!\!\!\perp C|B$. Note that this set of assumptions encompasses all the possible assumptions that you can make with 3 random variables, so any Bayes' net over $\mathbf{A}, \mathbf{B}, \mathbf{C}$ will be able to represent $\mathbf{d_4}$.

Q3. [8 pts] Bayes' Nets: Independence

For the following questions, each edge shown in the Bayes' nets below does not have a direction. For each of the edges shown, assign a direction (by adding an arrowhead at one end of each edge) to ensure that the Bayes' Net structure implies the assumptions provided. You cannot add new edges. The Bayes' nets can imply more assumptions than listed, but they *must* imply the ones listed. If there does not exist an assignment of directions that satisfies all the assumptions listed, clearly mark the *Not Possible* choice. *If you mark the Not Possible choice, the directions that you draw in the Bayes' net will not be looked at.* Keep in mind that Bayes' Nets cannot have directed cycles. You may find it useful to use the front of the next page to work on this problem.

(a) [2 pts]



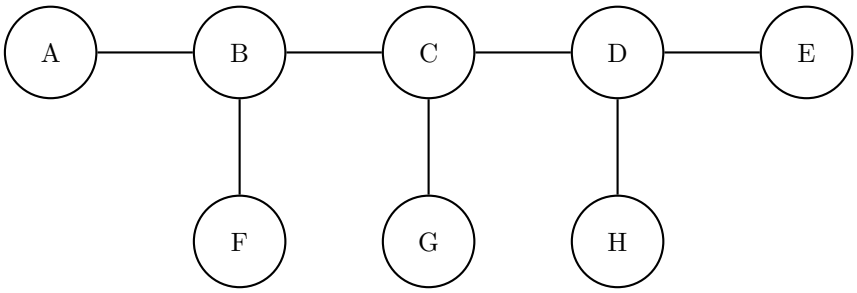
Assumptions:

- $A \perp\!\!\!\perp E$
- $B \perp\!\!\!\perp E \mid D$
- $A \perp\!\!\!\perp E \mid C$

○ Not Possible

This is one of three correct answers. A correct answer must have the common effect triple ABC and must have either the causal chain or common cause triple for CDE. This is because $A \perp\!\!\!\perp E$ implies that there must at least be one common effect triple with the middle node either at B, C or D. If the common effect triple is BCD, then the ABC and CDE triples are either going to be common cause or causal chain triples. Either way, when considering the assumption $A \perp\!\!\!\perp E|C$, the path $ABCDE$ will be active and so this assumption is not guaranteed to hold, so BCD cannot be a common effect triple. Now take the assumption $B \perp\!\!\!\perp E|D$. To make the path $BCDE$ inactive, since we claimed BCD is not a common effect triple, it must be either a causal chain or common cause triple, and either way, with C unobserved, the triple is active. So CDE needs to be inactive, and the only way to do that is to make CDE either a causal chain triple or common cause triple since D is observed. This means that CDE is not a common effect triple, so going back to the fact that there has to be at least one common effect triple to satisfy $A \perp\!\!\!\perp E$, the triple ABC must be a common effect triple.

(b) [3 pts]



Assumptions:

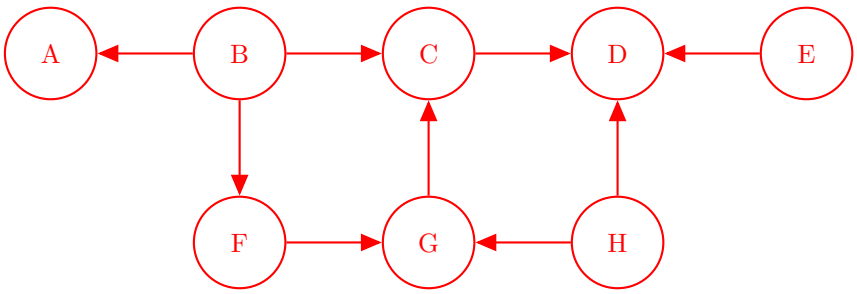
- $C \perp\!\!\!\perp E \mid H$
- $A \perp\!\!\!\perp H \mid B, D$
- $F \perp\!\!\!\perp G \mid A, D$

● Not Possible

There is no possible assignment of edges that satisfies all the assumptions. Starting off with the assumption $C \perp\!\!\!\perp E|H$, this implies that the triple CDE is a common effect triple with the middle node unobserved and none of its descendants observed, which means we need the edge $C \to D, E \to D$, and $H \to D$. Now take the assumption $A \perp\!\!\!\perp H|B, D$. We observe the path $ABCDH$. The triple CDH is a common effect triple with the middle node observed, so it is an active triple. No matter how you draw the direction between B and C, BCD will be either a common cause or causal chain triple with the middle node unobserved, so it is active. For the path to be inactive, we need the triple ABC to be inactive, and since the middle node B is observed, this

means that ABC cannot be a common effect triple. Now we consider the final assumption $F \perp\!\!\!\perp G|A, D$, which means we consider the path $FBCG$. To make this path inactive, one of the triples FBC or BCG must be inactive, and since the middle node is unobserved in both triples, one of these triples must be a common effect triple with no descendants observed. If we make FBC a common effect triple by adding the arrows $F \to B$ and $C \to B$, since we already claimed that ABC is not a common effect triple, then ABC must be a causal chain with the edges $C \to B$ and $B \to A$. Due to the edge $B \to A$, A is a descendant of B, so FBC will still be an active triple. Now let's consider making BCG a common effect triple by adding the arrows $B \to C$ and $G \to C$. But D is a descendant of C, so this triple is also active. Hence, the path $FBCG$ is guaranteed to be active, and so we cannot satisfy all assumptions.

(c) [3 pts]



Assumptions:

- $B \perp\!\!\!\perp H$
- $B \perp\!\!\!\perp G \mid E, F$
- $A \perp\!\!\!\perp E \mid C, H$

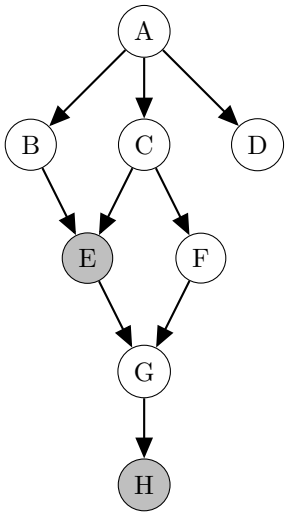
◯ Not Possible

This is one of four correct answers. A correct answer must have all edges the same as the solution above, with the exception that the edges between (A, F) and (A, B) can point in either direction. $B \perp\!\!\!\perp G|E, F$ implies that BCG is a common effect triple, so we add the edges $B \to C$ and $G \to C$. We also need to be wary that none of C's descendants can be observed, so we have to be careful of node E and F since they are observed. This also implies that the triple BFG must either be a causal chain triple or a common effect triple, which will come in handy later. $A \perp\!\!\!\perp E|C, H$ implies that BCD cannot be a common effect triple, so we have to add the arrow $C \to D$. Now, we mentioned that E should not be a descendant of C, so we need the edge $E \to D$. Now consider the assumption $B \perp\!\!\!\perp H$. For the path $BFGH$, we need a common effect triple in either BFG or FGH, but since we claimed earlier that BFG is not a common effect triple, FGH must be a common effect triple and so we add the arrows $F \to G$ and $H \to G$. Because a Bayes' net cannot have cycles, we need the arrow $H \to D$. For any direction for the edges (A, B) and (A, F), you can do a final check to see that all the assumptions will be satisfied in the Bayes' net.

This page is intentionally left blank for scratch work.
Nothing on this page will be graded.

Q4. [6 pts] Bayes' Nets: Inference

Assume we are given the following Bayes' net, and would like to perform inference to obtain $P(B, D \mid E = e, H = h)$.



(a) [1 pt] What is the number of rows in the largest factor generated by *inference by enumeration*, for this query $P(B, D \mid E = e, H = h)$? Assume all the variables are binary.

○ 2^8 ● 2^6 ○ 2^2 ○ 2^3

○ None of the above.

Since the inference by enumeration first joins all the factors in the Bayes' net, that factor will contain six (unobserved) variables. The question assumes all variables are binary, so the answer is 2^6.

(b) [2 pts] Mark all of the following variable elimination orderings that are optimal for calculating the answer for the query $P(B, D \mid E = e, H = h)$. Optimality is measured by the sum of the sizes of the factors that are generated. Assume all the variables are binary.

☐ C, A, F, G ■ G, F, C, A ☐ F, G, C, A ☐ A, C, F, G

☐ None of the above.

The sum of the sizes of factors that are generated for the variable elimination ordering G, F, C, A is $2^1 + 2^1 + 2^2 + 2^2$ rows, which is smaller than for any of the other variable elimination orderings. The ordering F, G, C, A is close but the sum of the sizes of factors is slightly bigger, with $2^2 + 2^1 + 2^2 + 2^2$ rows.

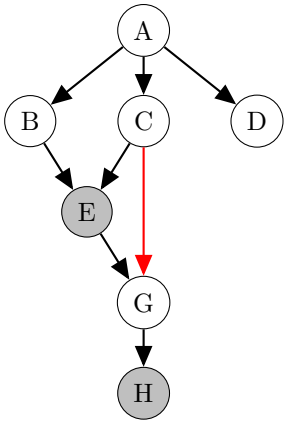
(c) Suppose we decide to perform variable elimination to calculate the query $P(B, D \mid E = e, H = h)$, and choose to eliminate F first.

(i) [2 pts] When F is eliminated, what intermediate factor is generated and how is it calculated? Make sure it is clear which variable(s) come before the conditioning bar and which variable(s) come after.

$f_1($ $G \mid C, e$ $) = \sum_f$ $P(f \mid C)P(G|f, e)$

This follows from the first step of variable elimination, which is to join all factors containing F, and then marginalize over F to obtain the intermediate factor f_1.

(ii) [1 pt] Now consider the set of distributions that can be represented by the remaining factors *after F is eliminated*. Draw the minimal number of directed edges on the following Bayes' Net structure, so that it can represent any distribution in this set. If no additional directed edges are needed, please fill in that option below.

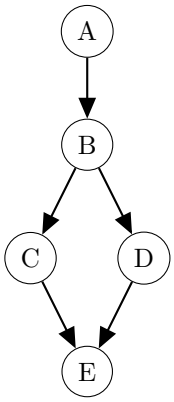


○ No additional directed edges needed

An additional edge from C to G is necessary, because the intermediate factor is of the form $f_1(G|C)$. Without this edge from C to G, the Bayes' net would not be able to express the dependence of G on C. (Note that adding an edge from G to C is not allowed, since that would introduce a cycle.)

Q5. [10 pts] Bayes' Nets: Sampling

Assume we are given the following Bayes' net, with the associated conditional probability tables (CPTs).



A	P(A)
$+a$	0.5
$-a$	0.5

A	B	P(B | A)
$+a$	$+b$	0.2
$+a$	$-b$	0.8
$-a$	$+b$	0.5
$-a$	$-b$	0.5

B	C	P(C | B)
$+b$	$+c$	0.4
$+b$	$-c$	0.6
$-b$	$+c$	0.8
$-b$	$-c$	0.2

B	D	P(D | B)
$+b$	$+d$	0.2
$+b$	$-d$	0.8
$-b$	$+d$	0.2
$-b$	$-d$	0.8

C	D	E	P(E | C, D)
$+c$	$+d$	$+e$	0.6
$+c$	$+d$	$-e$	0.4
$+c$	$-d$	$+e$	0.2
$+c$	$-d$	$-e$	0.8
$-c$	$+d$	$+e$	0.4
$-c$	$+d$	$-e$	0.6
$-c$	$-d$	$+e$	0.8
$-c$	$-d$	$-e$	0.2

You are given a set of the following samples, but are not told whether they were collected with rejection sampling or likelihood weighting.

$$\begin{array}{lllll} -a & -b & +c & +d & +e \\ -a & +b & +c & -d & +e \\ -a & -b & -c & -d & +e \\ -a & -b & +c & -d & +e \\ -a & +b & +c & +d & +e \end{array}$$

Throughout this problem, you may answer as either numeric expressions (e.g. 0.1∗0.5) or numeric values (e.g. 0.05).

(a) [2 pts] Assuming these samples were generated from *rejection sampling*, what is the sample based estimate of $P(+b \mid -a, +e)$?

Answer: 0.4

The answer is the number of samples satisfying the query variable's assignment (in this case, $B = +b$ divided by the total number of samples, so the answer is 2 / 5 = 0.4.

(b) [2 pts] Assuming these samples were generated from *likelihood weighting*, what is the sample-based estimate of $P(+b \mid -a, +e)$?

Answer: $\frac{1}{3}$

Based on likelihood weighting, we know the weight of each sample is $P(A = a) * P(E = e \mid C = c, D = d)$. The weights are: 0.3 (= 0.5 * 0.6), 0.1 (= 0.5 * 0.2), 0.4 (= 0.5 * 0.8), 0.1 (same assignments to C and D as second sample), 0.3 (same assignments to C and D as first sample). The estimate is then (0.1 + 0.3) / (0.3 + 0.1 + 0.4 + 0.1 + 0.3) = 0.4 / 1.20 = 1/3 = 0.333.

(c) [2 pts] Again, assume these samples were generated from *likelihood weighting*. However, you are not sure about the original CPT for $P(E \mid C, D)$ given above being the CPT associated with the Bayes' Net: With 50% chance, the CPT associated with the Bayes' Net is the original one. With the other 50% chance, the CPT is actually the CPT below.

C	D	E	P(E | C, D)
$+c$	$+d$	$+e$	0.8
$+c$	$+d$	$-e$	0.2
$+c$	$-d$	$+e$	0.4
$+c$	$-d$	$-e$	0.6
$-c$	$+d$	$+e$	0.2
$-c$	$+d$	$-e$	0.8
$-c$	$-d$	$+e$	0.6
$-c$	$-d$	$-e$	0.4

Samples from previous page copied below for convenience:

$-a$	$-b$	$+c$	$+d$	$+e$
$-a$	$+b$	$+c$	$-d$	$+e$
$-a$	$-b$	$-c$	$-d$	$+e$
$-a$	$-b$	$+c$	$-d$	$+e$
$-a$	$+b$	$+c$	$+d$	$+e$

Given this uncertainty, what is the sample-based estimate of $P(+b \mid -a, +e)$?

Answer: $\frac{10}{27}$

The weight of each sample is $P(A = a) * (0.5 * P_1(E = e \mid C = c, D = d) + 0.5 * P_2(E = e \mid C = c, D = d))$. The new weights are 0.35 (= 0.5 * (0.5 * 0.6 + 0.5 * 0.8)), 0.15 (= 0.5 * (0.5 * 0.2 + 0.5 * 0.4)), 0.35 (= 0.5 * (0.5 * 0.8 + 0.5 * 0.6)), 0.15, and 0.35. The estimate is then (0.15 + 0.35) / (0.35 * 3 + 0.15 * 2) = 0.5 / 1.35 = 10 / 27

(d) [1 pt] Now assume you can only sample a *small, limited number of samples*, and you want to estimate $P(+b, +d \mid -a)$ and $P(+b, +d \mid +e)$. You are allowed to estimate the answer to one query with likelihood weighting, and the other answer with rejection sampling. In order to obtain the best estimates for both queries, *which query should you estimate with likelihood weighting*? (The other query will have to be estimated with rejection sampling.)

- ○ $P(+b, +d \mid -a)$
- ● $P(+b, +d \mid +e)$
- ○ Either – both choices allow you to obtain the best estimates for both queries.

The evidence +e is at the leaves of the Bayes' net, which means it's possible to sample all the other variables, but have to reject the last node E. We can avoid this problem by using likelihood weighting for sampling, since it fixes the values of observed random variables to that of the fixed evidence.

(e) Suppose you choose to use Gibbs sampling to estimate $P(B, E \mid +c, -d)$. Assume the CPTs are the same as the ones for parts (a) and (b). Currently your assignments are the following:

$$-a \quad -b \quad +c \quad -d \quad +e$$

(i) [1 pt] Suppose the next step is to resample E.
What is the probability that the new assignment to E will be +e?

Answer: 0.2 In order to sample E, we need to calculate $P(E \mid -a, -b, +c, -d)$, which is equal to $P(E \mid +c, -d)$ since E is conditionally independent of A and B, given C and D. The value for $P(+e \mid +c, -d)$ is given directly in the CPT for $P(E|C, D)$, and it is 0.2.

(ii) [1 pt] Instead, suppose the next step is to resample A.
What is the probability that the new assignment to A will be $+a$?

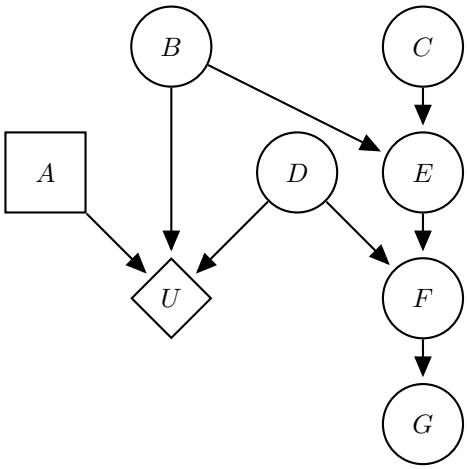
Answer: $\frac{8}{13}$ In order to sample A, we need to calculate $P(A \mid -b, +c, -d, +e)$, which is equal to $P(A \mid B)$ since A is conditionally independent of C, D, and E, given B. We can calculate this using Bayes' rule: $P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$. Thus, $P(+a \mid -b) = \frac{P(-b|+a)P(+a)}{P(-b)} = \frac{P(-b|+a)P(+a)}{\sum_{a \in \{+a,-a\}} P(-b|a)} P(a) = \frac{0.8*0.5}{0.8*0.5+0.5*0.5)} = \frac{0.4}{0.65} = \frac{8}{13}$

(iii) [1 pt] Instead, suppose the next step is to resample B.
What is the probability that the new assignment to B will be +b?

Answer: $\frac{1}{3}$ In order to sample B, we need to calculate $P(B \mid -a, +c, -d, +e)$, which is equal to $P(B \mid -a, +c, -d)$ since B is conditionally independent of E, given C and D. The CPT tables that are involved in calculating $P(B \mid -a + c, -d, +e)$ are $P(A)$, $P(B|A)$, $P(C|B)$, $P(D|B)$. First, we remove rows of the CPTs that do not agree with the evidence -a, +c, -d, +e. We then join the resulting CPTs to obtain $P(-a, B, +c, -d)$. We select $P(-a, +b, +c, -d)$ and $P(-a, -b, +c, -d)$ from this table, and normalize so that the two probabilities sum to one (i.e., to transform them to $P(+b|-a, +c, -d)$ and $P(-b|-a, +c, -d)$.

Q6. [13 pts] VPI

Consider a decision network with the following structure. Node A is an action, and node U is the utility:



(a) [1 pt] Choose the option which is *guaranteed* to be true, or "Neither guaranteed" if no option is guaranteed to be true.

● $VPI(C) = 0$ ○ $VPI(C) > 0$ ○ Neither guaranteed

It is guaranteed that $VPI(X) = 0$ if and only if X is conditionally independent of U, and so knowing X would not change our beliefs about the utility we get. D-Separation shows us that this is true for node C, so $VPI(C) = 0$.

(b) [2 pts] Mark *all* of the following that are *guaranteed* to be true.

■ $VPI(E) \leq VPI(B)$ ☐ $VPI(E) \geq VPI(B)$

☐ $VPI(E) = VPI(B)$ ☐ None of the above

Utility depends on the action A and the values of B and D, so the value of E matters only if it gives us information about B and D. In this case, since E is conditionally independent of D, it can only give us information about B. E can give us noisy information about B, but this cannot be any better than actually knowing the value of B. So, we get that $VPI(E) \leq VPI(B)$. Note that equality can happen when E gives perfect information about B, or when both VPIs are 0, but it is not guaranteed to be happen.

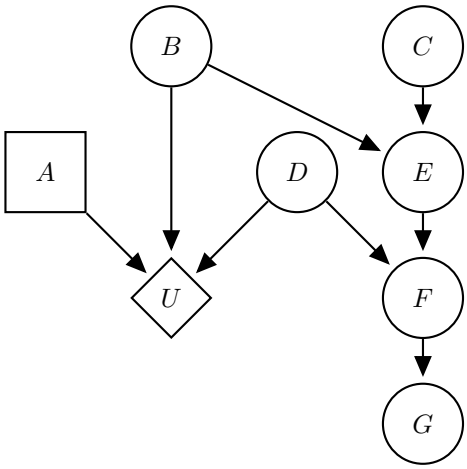
(c) [2 pts] Mark *all* of the following that are *guaranteed* to be true.

☐ $VPI(E \mid F) \leq VPI(B \mid F)$ ☐ $VPI(E \mid F) \geq VPI(B \mid F)$

☐ $VPI(E \mid F) = VPI(B \mid F)$ ■ None of the above

Here, E gives us some information about B and D. This may or may not be more valuable than perfect information about B and less information about D.

The decision network on the previous page has been reproduced below:



(d) [2 pts] Noting that $E \perp\!\!\!\perp G \mid F$, mark *all* of the following that are *guaranteed* to be true.

- ■ $VPI(E, G \mid F) = VPI(E \mid F) + VPI(G \mid E, F)$
- ☐ $VPI(E, G \mid F) = VPI(E \mid F)VPI(G \mid E, F)$
- ■ $VPI(E, G \mid F) = VPI(E \mid F) + VPI(G \mid F)$
- ☐ $VPI(E, G \mid F) = VPI(E \mid F)VPI(G \mid F)$
- ☐ None of the above

Regardless of the Bayes Net structure, it is true that $VPI(E, G \mid F) = VPI(E \mid F) + VPI(G \mid E, F)$. This just says that the information by getting E and then getting G is the same as the information by getting E and G at the same time.

Then, D-separation shows that E is conditionally independent of G given F, and so we know $VPI(G \mid E, F) = VPI(G \mid F)$. Thus, $VPI(E, G \mid F) = VPI(E \mid F) + VPI(G \mid F)$

(e) [3 pts] Suppose we have two actions, a_1 and a_2. In addition, you are given $P(B)$ and $P(D)$ below. Fill in the empty entries in the utility table below **with either** 1 **or** -1 such that $VPI(B) = VPI(D) = 0$, but $VPI(B, D) > 0$.

Note on grading: You will get 1 point for each condition you enforce correctly, that is, you get 1 point each for ensuring $VPI(B) = 0$, $VPI(D) = 0$, and $VPI(B, D) > 0$. If you cannot enforce all three conditions, try to enforce two for partial credit.

$P(B)$	
+b	0.5
-b	0.5

$P(D)$	
+d	0.5
-d	0.5

Action a_1: $U(a_1, B, D)$		
$+b$	$+d$	1
$-b$	$+d$	-1
$+b$	$-d$	-1
$-b$	$-d$	1

Action a_2: $U(a_2, B, D)$		
$+b$	$+d$	-1
$-b$	$+d$	1
$+b$	$-d$	1
$-b$	$-d$	-1

There are many possible answers. The conditions that need to hold are given below:

For $VPI(B, D) > 0$, we need the utilities to be such that for some assignment to B and D, the optimal action is a_1, whereas for some other assignment, the optimal action is a_2. Thus, having the information of B and D allows us to choose a_1 sometimes and a_2 sometimes, giving us a better score than if we only got to choose a_1 or a_2 no matter what the values of B and D are. Thus, there needs to be one row where $U(a_1, b, d) > U(a_2, b, d)$ and another row where $U(a_2, b, d) > U(a_1, b, d)$.

For $VPI(D) = 0$, we need the optimal action to remain the same when we know D and when we don't know D. First, notice that the optimal action when we don't know D is just whichever table has more 1s. (If both tables have the same number of 1s, then both actions have an expected utility of 0 and so they are both optimal.)

Suppose that the optimal action is a_1. Then, when we learn D, in both cases ($D = +d$ and $D = -d$) the optimal action should still be a_1 (if it changes to a_2, that will increase the expected utility). The optimal action for $D = +d$ is the action which has more 1s in the first two rows, and the optimal action for $D = -d$ is the action which has more 1s in the last two rows.

Similarly, for $VPI(B) = 0$, we need the optimal action to remain the same when we know B and when we don't know B. Here, the optimal action for $B = +b$ is the action which has more 1s in the first and third rows, and the optimal action for $B = -b$ is the action which has more 1s in the second and fourth rows.

(f) For this question, assume you did the previous part correctly - that is, you should assume that $VPI(B) = VPI(D) = 0$, and $VPI(B, D) > 0$.

Now taking into account your answer from the previous part, choose the option which is *guaranteed* to be true, or "Neither guaranteed" if no option is guaranteed to be true. **This means that you should fill in one circle for each row.**

(i) [1 pt]	● $VPI(E) = 0$	○ $VPI(E) > 0$	○ Neither guaranteed
(ii) [1 pt]	○ $VPI(G) = 0$	○ $VPI(G) > 0$	● Neither guaranteed
(iii) [1 pt]	○ $VPI(D \mid B) = 0$	● $VPI(D \mid B) > 0$	○ Neither guaranteed

We know that E is conditionally independent of D given no evidence. So, E can only give us information about B, and not about D. However, even perfect information about B is not worth anything (since $VPI(B) = 0$), and so the imperfect information about B given by E is also useless, and so $VPI(E) = 0$.

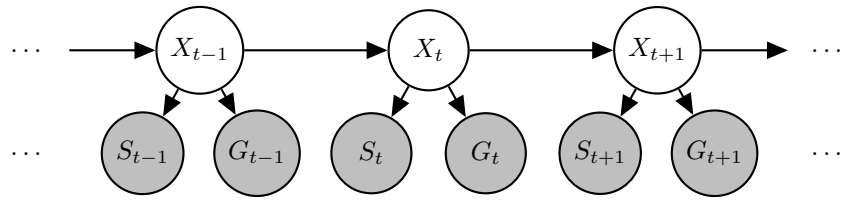
On the other hand, G can affect both B and D (by D-separation, we see that we cannot guarantee that G is conditionally independent of B or D). Thus, G could give us information about both B and D, and since $VPI(B, D) > 0$, we could have $VPI(G) > 0$. However, we are not guaranteed that this is true - perhaps G is conditionally independent of B, even though it cannot be inferred by D-separation. In this case, we would have $VPI(G) = 0$.

Finally, we have $VPI(D \mid B) = VPI(B, D) - VPI(B) > 0$.

Q7. [13 pts] HMM: Where is the Car?

Transportation researchers are trying to improve traffic in the city but, in order to do that, they first need to estimate the location of each of the cars in the city. They need our help to model this problem as an inference problem of an HMM. For this question, assume that only *one* car is being modeled.

(a) The structure of this modified HMM is given below, which includes X, the location of the car; S, the noisy location of the car from the signal strength at a nearby cell phone tower; and G, the noisy location of the car from GPS.



We want to perform filtering with this HMM. That is, we want to compute the belief $P(x_t|s_{1:t}, g_{1:t})$, the probability of a state x_t given all past and current observations.

The **dynamics update** expression has the following form:

$$P(x_t|s_{1:t-1}, g_{1:t-1}) = \underline{\quad \textbf{(i)} \quad} \; \underline{\quad \textbf{(ii)} \quad} \; \underline{\quad \textbf{(iii)} \quad} \; P(x_{t-1}|s_{1:t-1}, g_{1:t-1}).$$

Complete the expression by choosing the option that fills in each blank.

(i)	[1 pt]	○	$P(s_{1:t}, g_{1:t})$	○	$P(s_{1:t-1}, g_{1:t-1})$	○	$P(s_{1:t})P(g_{1:t})$	○	$P(s_{1:t-1})P(g_{1:t-1})$	● 1
(ii)	[1 pt]	○	$\sum_{x_t}$	●	$\sum_{x_{t-1}}$	○	$\max_{x_{t-1}}$	○	$\max_{x_t}$	○ 1
(iii)	[1 pt]	○	$P(x_{t-2}, x_{t-1})$	○	$P(x_{t-1} \mid x_{t-2})$	○	$P(x_{t-1}, x_t)$	●	$P(x_t|x_{t-1})$	○ 1

The derivation of the dynamics update is similar to the one for the canonical HMM, but with two observation variables instead.

$$\begin{aligned} P(x_t|s_{1:t-1}, g_{1:t-1}) &= \sum_{x_{t-1}} P(x_{t-1}, x_t|s_{1:t-1}, g_{1:t-1}) \\ &= \sum_{x_{t-1}} P(x_t|x_{t-1}, s_{1:t-1}, g_{1:t-1}) P(x_{t-1}|s_{1:t-1}, g_{1:t-1}) \\ &= \sum_{x_{t-1}} P(x_t|x_{t-1}) P(x_{t-1}, x_t|s_{1:t-1}, g_{1:t-1}) \end{aligned}$$

In the last step, we use the independence assumption given in the HMM, $X_t \perp\!\!\!\perp S_{1:t-1}, G_{1:t-1}|X_{t-1}$.

The **observation update** expression has the following form:

$$P(x_t|s_{1:t}, g_{1:t}) = \underline{\textbf{(iv)}} \quad \underline{\textbf{(v)}} \quad \underline{\textbf{(vi)}} \quad P(x_t|s_{1:t-1}, g_{1:t-1}).$$

Complete the expression by choosing the option that fills in each blank.

(iv) [1 pt]
- ○ $P(s_t, g_t|s_{1:t-1}, g_{1:t-1})$
- ○ $P(s_{1:t-1}, g_{1:t-1}|s_t, g_t)$
- ○ $P(s_t|s_{1:t-1})P(g_t|g_{1:t-1})$
- ○ $P(s_{1:t-1}|s_t)P(g_{1:t-1}|g_t)$
- ● $\frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})}$
- ○ $\frac{1}{P(s_{1:t-1}, g_{1:t-1}|s_t, g_t)}$
- ○ $\frac{1}{P(s_t|s_{1:t-1})P(g_t|g_{1:t-1})}$
- ○ $\frac{1}{P(s_{1:t-1}|s_t)P(g_{1:t-1}|g_t)}$
- ○ 1

(v) [1 pt]
- ○ $\sum_{x_{t-1}}$
- ○ $\sum_{x_t}$
- ○ $\max_{x_{t-1}}$
- ○ $\max_{x_t}$
- ● 1

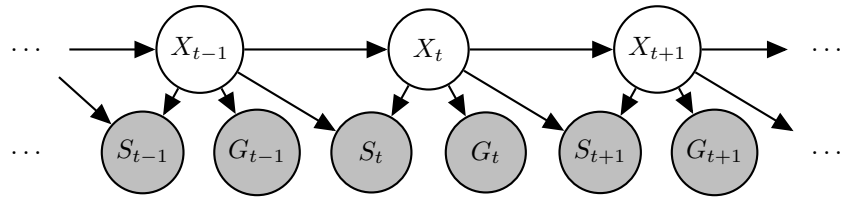
(vi) [1 pt]
- ○ $P(s_{t-1}|x_{t-1})P(g_{t-1}|x_{t-1})$
- ○ $P(x_t, s_t)P(x_t, g_t)$
- ○ $P(x_t, s_t, g_t)$
- ○ $P(x_{t-1}, s_{t-1})P(x_{t-1}, g_{t-1})$
- ○ $P(x_{t-1}, s_{t-1}, g_{t-1})$
- ○ $P(x_t|s_t)P(x_t|g_t)$
- ○ $P(x_{t-1}|s_{t-1})P(x_{t-1}|g_{t-1})$
- ● $P(s_t|x_t)P(g_t|x_t)$
- ○ 1

Again, the derivation of the observation update is similar to the one for the canonical HMM, but with two observation variables instead.

$$\begin{aligned}
P(x_t|s_{1:t}, g_{1:t}) &= P(x_t|s_t, g_t, s_{1:t-1}, g_{1:t-1}) \\
&= \frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} P(x_t, s_t, g_t|s_{1:t-1}, g_{1:t-1}) \\
&= \frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} P(s_t, g_t|x_t, s_{1:t-1}, g_{1:t-1}) P(x_t|s_{1:t-1}, g_{1:t-1}) \\
&= \frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} P(s_t, g_t|x_t) P(x_t|s_{1:t-1}, g_{1:t-1}) \\
&= \frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} P(s_t|x_t) P(g_t|x_t) P(x_t|s_{1:t-1}, g_{1:t-1})
\end{aligned}$$

In the second to last step, we use the independence assumption $S_t, G_t \perp\!\!\!\perp S_{1:t-1}, G_{1:t-1}|X_t$; and in the last step, we use the independence assumption $S_t \perp\!\!\!\perp G_t|X_t$.

(b) It turns out that if the car moves too fast, the quality of the cell phone signal decreases. Thus, the signal-dependent location S_t not only depends on the current state X_t but it also depends on the previous state X_{t-1}. Thus, we modify our original HMM for a new more accurate one, which is given below.



Again, we want to compute the belief $P(x_t|s_{1:t}, g_{1:t})$. In this part we consider an update that combines the dynamics and observation update in a *single* update.

$$P(x_t|s_{1:t}, g_{1:t}) = \underline{\quad\textbf{(i)}\quad}\ \underline{\quad\textbf{(ii)}\quad}\ \underline{\quad\textbf{(iii)}\quad}\ \underline{\quad\textbf{(iv)}\quad}\ P(x_{t-1}|s_{1:t-1}, g_{1:t-1}).$$

Complete the **forward update** expression by choosing the option that fills in each blank.

(i) [1 pt]
- ○ $P(s_t, g_t|s_{1:t-1}, g_{1:t-1})$
- ○ $P(s_{1:t-1}, g_{1:t-1}|s_t, g_t)$
- ○ $P(s_t|s_{1:t-1})P(g_t|g_{1:t-1})$
- ● $\frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})}$
- ○ $\frac{1}{P(s_{1:t-1}, g_{1:t-1}|s_t, g_t)}$
- ○ $P(s_{1:t-1}|s_t)P(g_{1:t-1}|g_t)$
- ○ $\frac{1}{P(s_t|s_{1:t-1})P(g_t|g_{1:t-1})}$
- ○ $\frac{1}{P(s_{1:t-1}|s_t)P(g_{1:t-1}|g_t)}$
- ○ 1

(ii) [1 pt]
- ● $\sum_{x_{t-1}}$
- ○ $\sum_{x_t}$
- ○ $\max_{x_{t-1}}$
- ○ $\max_{x_t}$
- ○ 1

(iii) [1 pt]
- ○ $P(x_{t-2}, x_{t-1}, s_{t-1})P(x_{t-1}, g_{t-1})$
- ○ $P(x_{t-1}, x_t, s_t)P(x_t, g_t)$
- ○ $P(s_{t-1}, g_{t-1}|x_{t-1})$
- ○ $P(s_{t-1}|x_{t-2}, x_{t-1})P(g_{t-1}|x_{t-1})$
- ● $P(s_t|x_{t-1}, x_t)P(g_t|x_t)$
- ○ $P(s_t, g_t|x_t)$
- ○ $P(x_{t-2}, x_{t-1}|s_{t-1})P(x_{t-1}|g_{t-1})$
- ○ $P(x_{t-1}, x_t|s_t)P(x_t|g_t)$
- ○ 1
- ○ $P(x_{t-2}, x_{t-1}, s_{t-1}, g_{t-1})$
- ○ $P(x_{t-1}, x_t, s_t, g_t)$

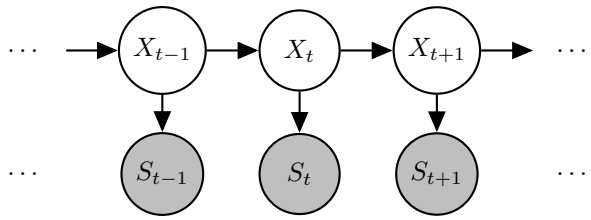
(iv) [1 pt]
- ○ $P(x_{t-1}, x_t)$
- ● $P(x_t|x_{t-1})$
- ○ $P(x_{t-2}, x_{t-1})$
- ○ $P(x_{t-1}|x_{t-2})$
- ○ 1

For this modified HMM, we have the dynamics and observation update in a single update because one of the previous independence assumptions does not longer holds.

$$\begin{aligned}
P(x_t|s_{1:t}, g_{1:t}) &= \sum_{x_{t-1}} P(x_{t-1}, x_t|s_t, g_t, s_{1:t-1}, g_{1:t-1}) \\
&= \frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} \sum_{x_{t-1}} P(x_{t-1}, x_t, s_t, g_t|s_{1:t-1}, g_{1:t-1}) \\
&= \frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} \sum_{x_{t-1}} P(s_t, g_t|x_{t-1}, x_t, s_{1:t-1}, g_{1:t-1})P(x_{t-1}, x_t|s_{1:t-1}, g_{1:t-1}) \\
&= \frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} \sum_{x_{t-1}} P(s_t, g_t|x_{t-1}, x_t)P(x_t|x_{t-1}, s_{1:t-1}, g_{1:t-1})P(x_{t-1}|s_{1:t-1}, g_{1:t-1}) \\
&= \frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} \sum_{x_{t-1}} P(s_t|x_{t-1}, x_t)P(g_t|x_{t-1}, x_t)P(x_t|x_{t-1})P(x_{t-1}|s_{1:t-1}, g_{1:t-1}) \\
&= \frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} \sum_{x_{t-1}} P(s_t|x_{t-1}, x_t)P(g_t|x_t)P(x_t|x_{t-1})P(x_{t-1}|s_{1:t-1}, g_{1:t-1})
\end{aligned}$$

In the third to last step, we use the independence assumption $S_t, G_t \perp\!\!\!\perp S_{1:t-1}, G_{1:t-1}|X_{t-1}, X_t$; in the second to last step, we use the independence assumption $S_t \perp\!\!\!\perp G_t|X_{t-1}, X_t$ and $X_t \perp\!\!\!\perp S_{1:t-1}, G_{1:t-1}|X_{t-1}$; and in the last step, we use the independence assumption $G_t \perp\!\!\!\perp X_{t-1}|X_t$.

(c) The Viterbi algorithm finds the most probable sequence of hidden states $X_{1:T}$, given a sequence of observations $s_{1:T}$, for some time $t = T$. Recall the canonical HMM structure, which is shown below.



For this canonical HMM, the Viterbi algorithm performs the following dynamic programming computations:

$$m_t[x_t] = P(s_t|x_t) \max_{x_{t-1}} P(x_t|x_{t-1}) m_{t-1}[x_{t-1}].$$

We consider extending the Viterbi algorithm for the modified HMM from part **(b)**. We want to find the most likely sequence of states $X_{1:T}$ given the sequence of observations $s_{1:T}$ and $g_{1:T}$. The dynamic programming update for $t > 1$ for the modified HMM has the following form:

$$m_t[x_t] = \underline{\textbf{(i)}} \quad \underline{\textbf{(ii)}} \quad \underline{\textbf{(iii)}} \quad m_{t-1}[x_{t-1}].$$

Complete the expression by choosing the option that fills in each blank.

(i) [1 pt] ○ $\sum_{x_t}$ ○ $\sum_{x_{t-1}}$ ○ $\max_{x_t}$ ● $\max_{x_{t-1}}$ ○ 1

(ii) [1 pt]
○ $P(x_{t-2}, x_{t-1}, s_{t-1})P(x_{t-1}, g_{t-1})$ ○ $P(x_{t-1}, x_t, s_t)P(x_t, g_t)$ ○ $P(s_{t-1}, g_{t-1}|x_{t-1})$
○ $P(s_{t-1}|x_{t-2}, x_{t-1})P(g_{t-1}|x_{t-1})$ ● $P(s_t|x_{t-1}, x_t)P(g_t|x_t)$ ○ $P(s_t, g_t|x_t)$
○ $P(x_{t-2}, x_{t-1}|s_{t-1})P(x_{t-1}|g_{t-1})$ ○ $P(x_{t-1}, x_t|s_t)P(x_t|g_t)$ ○ 1
○ $P(x_{t-2}, x_{t-1}, s_{t-1}, g_{t-1})$ ○ $P(x_{t-1}, x_t, s_t, g_t)$

(iii) [1 pt] ○ $P(x_{t-1}, x_t)$ ● $P(x_t|x_{t-1})$ ○ $P(x_{t-2}, x_{t-1})$ ○ $P(x_{t-1}|x_{t-2})$ ○ 1

If we remove the summation from the forward update equation of part **(b)**, we get a joint probability of the states,

$$P(x_{1:t}|s_{1:t}, g_{1:t}) = \frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} P(s_t|x_{t-1}, x_t)P(g_t|x_t)P(x_t|x_{t-1})P(x_{1:t-1}|s_{1:t-1}, g_{1:t-1}).$$

We can define $m_t[x_t]$ to be the maximum joint probability of the states (for a particular x_t) given all past and current observations, times some constant, and then we can find a recursive relationship for $m_t[x_t]$,

$$\begin{aligned}
m_t[x_t] &= P(s_{1:t}, g_{1:t}) \max_{x_{1:t-1}} P(x_{1:t}|s_{1:t}, g_{1:t}) \\
&= P(s_{1:t}, g_{1:t}) \max_{x_{1:t-1}} \frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} P(s_t|x_{t-1}, x_t)P(g_t|x_t)P(x_t|x_{t-1})P(x_{1:t-1}|s_{1:t-1}, g_{1:t-1}) \\
&= \max_{x_{t-1}} P(s_t|x_{t-1}, x_t)P(g_t|x_t)P(x_t|x_{t-1}) \frac{P(s_{1:t}, g_{1:t})}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} \max_{x_{1:t-2}} P(x_{1:t-1}|s_{1:t-1}, g_{1:t-1}) \\
&= \max_{x_{t-1}} P(s_t|x_{t-1}, x_t)P(g_t|x_t)P(x_t|x_{t-1})P(s_{1:t-1}, g_{1:t-1}) \max_{x_{1:t-2}} P(x_{1:t-1}|s_{1:t-1}, g_{1:t-1}) \\
&= \max_{x_{t-1}} P(s_t|x_{t-1}, x_t)P(g_t|x_t)P(x_t|x_{t-1})m_{t-1}[x_{t-1}].
\end{aligned}$$

Notice that the maximum joint probability of states up to time $t = T$ given all past and current observations is given by

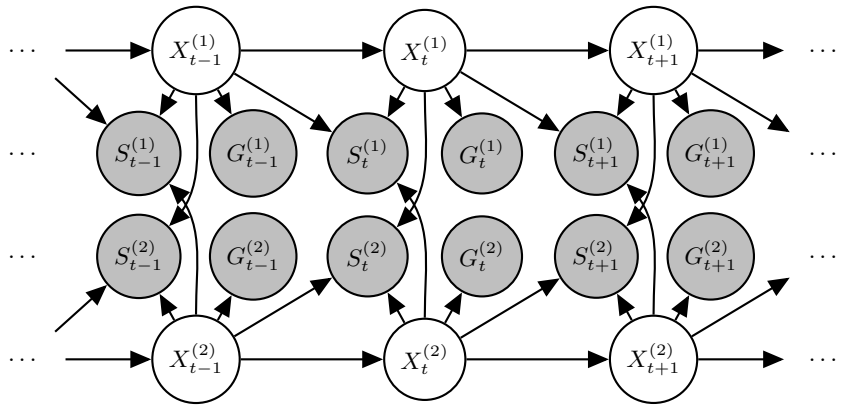
$$\max_{x_{1:T}} P(x_{1:T}|s_{1:T}, g_{1:T}) = \frac{\max_{x_t} m_T[x_t]}{P(s_{1:T}, g_{1:T})}.$$

We can recover the actual most likely sequence of states by bookkepping back pointers of the states the maximized the Viterbi update equations.

Q8. [11 pts] Particle Filtering: Where are the Two Cars?

As before, we are trying to estimate the location of cars in a city, but now, we model two cars jointly, i.e. car i for $i \in \{1, 2\}$. The modified HMM model is as follows:

- $X^{(i)}$ – the location of car i
- $S^{(i)}$ – the noisy location of the car i from the signal strength at a nearby cell phone tower
- $G^{(i)}$ – the noisy location of car i from GPS



d	$D(d)$	$E_L(d)$	$E_N(d)$	$E_G(d)$
-4	0.05	0	0.02	0
-3	0.10	0	0.04	0.03
-2	0.25	0.05	0.09	0.07
-1	0.10	0.10	0.20	0.15
0	0	0.70	0.30	0.50
1	0.10	0.10	0.20	0.15
2	0.25	0.05	0.09	0.07
3	0.10	0	0.04	0.03
4	0.05	0	0.02	0

The signal strength from one car gets noisier if the other car is at the same location. Thus, the observation $S_t^{(i)}$ also depends on the current state of the other car $X_t^{(j)}$, $j \neq i$.

The transition is modeled using a drift model D, the GPS observation $G_t^{(i)}$ using the error model E_G, and the observation $S_t^{(i)}$ using one of the error models E_L or E_N, depending on the car's speed and the relative location of both cars. These drift and error models are in the table above. **The transition and observation models are:**

$$P(X_t^{(i)}|X_{t-1}^{(i)}) = D(X_t^{(i)} - X_{t-1}^{(i)})$$

$$P(S_t^{(i)}|X_{t-1}^{(i)}, X_t^{(i)}, X_t^{(j)}) = \begin{cases} E_N(X_t^{(i)} - S_t^{(i)}), & \text{if } |X_t^{(i)} - X_{t-1}^{(i)}| \geq 2 \text{ or } X_t^{(i)} = X_t^{(j)} \\ E_L(X_t^{(i)} - S_t^{(i)}), & \text{otherwise} \end{cases}$$

$$P(G_t^{(i)}|X_t^{(i)}) = E_G(X_t^{(i)} - G_t^{(i)}).$$

Throughout this problem you may give answers either as unevaluated numeric expressions (e.g. $0.1 \cdot 0.5$) or as numeric values (e.g. 0.05). The questions are decoupled.

(a) Assume that at $t = 3$, we have the single particle $(X_3^{(1)} = -1, X_3^{(2)} = 2)$.

(i) [2 pts] What is the probability that this particle becomes $(X_4^{(1)} = -3, X_4^{(2)} = 3)$ after passing it through the dynamics model?

$$\begin{aligned} P(X_4^{(1)} = -3, X_4^{(2)} = 3 | X_3^{(1)} = -1, X_3^{(2)} = 2) &= P(X_4^{(1)} = -3 | X_3^{(1)} = -1) \cdot P(X_4^{(2)} = 3 | X_3^{(2)} = 2) \\ &= D(-3 - (-1)) \cdot D(3 - 2) \\ &= 0.25 \cdot 0.10 \\ &= 0.025 \end{aligned}$$

Answer: 0.025

(ii) [2 pts] Assume that there are no sensor readings at $t = 4$. What is the joint probability that the *original* single particle (from $t = 3$) becomes $(X_4^{(1)} = -3, X_4^{(2)} = 3)$ and then becomes $(X_5^{(1)} = -4, X_5^{(2)} = 4)$?

$$
\begin{aligned}
&P(X_4^{(1)} = -3, X_5^{(1)} = -4, X_4^{(2)} = 3, X_5^{(2)} = 4 | X_3^{(1)} = -1, X_3^{(2)} = 2) \\
&\quad = P(X_4^{(1)} = -3, X_5^{(1)} = -4 | X_3^{(1)} = -1) \cdot P(X_4^{(2)} = 3, X_5^{(2)} = 4 | X_3^{(2)} = 2) \\
&\quad = P(X_5^{(1)} = -4 | X_4^{(1)} = -3) \cdot P(X_4^{(1)} = -3 | X_3^{(1)} = -1) \cdot P(X_5^{(2)} = 4 | X_4^{(2)} = 3) \cdot P(X_4^{(2)} = 3 | X_3^{(2)} = 2) \\
&\quad = D(-4 - (-3)) \cdot D(-3 - (-1)) \cdot D(4 - 3) \cdot D(3 - 2) \\
&\quad = 0.10 \cdot 0.25 \cdot 0.10 \cdot 0.10 \\
&\quad = 0.00025
\end{aligned}
$$

Answer: 0.00025

For the remaining of this problem, we will be using 2 particles at each time step.

(b) At $t = 6$, we have particles $[(X_6^{(1)} = 3, X_6^{(2)} = 0), (X_6^{(1)} = 3, X_6^{(2)} = 5)]$. Suppose that after weighting, resampling, and transitioning from $t = 6$ to $t = 7$, the particles become $[(X_7^{(1)} = 2, X_7^{(2)} = 2), (X_7^{(1)} = 4, X_7^{(2)} = 1)]$.

(i) [2 pts] At $t = 7$, you get the observations $S_7^{(1)} = 2, G_7^{(1)} = 2, S_7^{(2)} = 2, G_7^{(2)} = 2$. What is the weight of each particle?

Particle	Weight
$(X_7^{(1)} = 2, X_7^{(2)} = 2)$	$P(S_7^{(1)} = 2 \mid X_6^{(1)} = 3, X_7^{(1)} = 2, X_7^{(2)} = 2) \cdot P(G_7^{(1)} = 2 \mid X_7^{(1)} = 2) \cdot$ $P(S_7^{(2)} = 2 \mid X_6^{(2)} = 0, X_7^{(2)} = 2, X_7^{(1)} = 2) \cdot P(G_7^{(2)} = 2 \mid X_7^{(2)} = 2)$ $= E_N(2-2) \cdot E_G(2-2) \cdot E_N(2-2) \cdot E_G(2-2)$ $= 0.30 \cdot 0.50 \cdot 0.30 \cdot 0.50$ $= 0.0225$
$(X_7^{(1)} = 4, X_7^{(2)} = 1)$	$P(S_7^{(1)} = 2 \mid X_6^{(1)} = 3, X_7^{(1)} = 4, X_7^{(2)} = 1) \cdot P(G_7^{(1)} = 2 \mid X_7^{(1)} = 4) \cdot$ $P(S_7^{(2)} = 2 \mid X_6^{(2)} = 5, X_7^{(2)} = 1, X_7^{(1)} = 4) \cdot P(G_7^{(2)} = 2 \mid X_7^{(2)} = 1)$ $= E_L(4-2) \cdot E_G(4-2) \cdot E_N(1-2) \cdot E_G(1-2)$ $= 0.05 \cdot 0.07 \cdot 0.20 \cdot 0.15$ $= 0.000105$

(ii) [2 pts] Suppose both cars' cell phones died so you only get the observations $G_7^{(1)} = 2, G_7^{(2)} = 2$. What is the weight of each particle?

Particle	Weight
$(X_7^{(1)} = 2, X_7^{(2)} = 2)$	$P(G_7^{(1)} = 2 \mid X_7^{(1)} = 2) \cdot P(G_7^{(2)} = 2 \mid X_7^{(2)} = 2)$ $= E_G(2-2) \cdot E_G(2-2)$ $= 0.50 \cdot 0.50$ $= 0.25$
$(X_7^{(1)} = 4, X_7^{(2)} = 1)$	$P(G_7^{(1)} = 2 \mid X_7^{(1)} = 4) \cdot P(G_7^{(2)} = 2 \mid X_7^{(2)} = 1)$ $= E_G(4-2) \cdot E_G(1-2)$ $= 0.07 \cdot 0.15$ $= 0.0105$

(c) [3 pts] To decouple this question, assume that you got the following weights for the two particles.

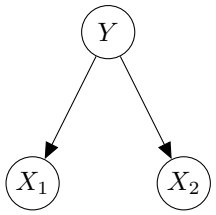
Particle	Weight
$(X_7^{(1)} = 2, X_7^{(2)} = 2)$	0.09
$(X_7^{(1)} = 4, X_7^{(2)} = 1)$	0.01

What is the belief for the location of car 1 and car 2 at $t = 7$?

Location	$P(X_7^{(1)})$	$P(X_7^{(2)})$
$X_7^{(i)} = 1$	$\frac{0}{0.09+0.01} = 0$	$\frac{0.01}{0.09+0.01} = 0.1$
$X_7^{(i)} = 2$	$\frac{0.09}{0.09+0.01} = 0.9$	$\frac{0.09}{0.09+0.01} = 0.9$
$X_7^{(i)} = 4$	$\frac{0.01}{0.09+0.01} = 0.1$	$\frac{0}{0.09+0.01} = 0$

Q9. [7 pts] Naive Bayes MLE

Consider a naive Bayes classifier with two features, shown below. We have prior information that the probability model can be parameterized by λ and p, as shown below: Note that $P(X_1 = 0|Y = 0) = P(X_1 = 1|Y = 1) = p$ and $P(X_1|Y) = P(X_2|Y)$ (they share the parameter p). Call this model M1.



Y	$P(Y)$
0	λ
1	1 - λ

X_1	Y	$P(X_1\|Y)$
0	0	p
1	0	$1-p$
0	1	$1-p$
1	1	p

X_2	Y	$P(X_2\|Y)$
0	0	p
1	0	$1-p$
0	1	$1-p$
1	1	p

We have a training set that contains all of the following:

- n_{000} examples with $X_1 = 0$, $X_2 = 0$, $Y = 0$
- n_{010} examples with $X_1 = 0$, $X_2 = 1$, $Y = 0$
- n_{100} examples with $X_1 = 1$, $X_2 = 0$, $Y = 0$
- n_{110} examples with $X_1 = 1$, $X_2 = 1$, $Y = 0$
- n_{001} examples with $X_1 = 0$, $X_2 = 0$, $Y = 1$
- n_{011} examples with $X_1 = 0$, $X_2 = 1$, $Y = 1$
- n_{101} examples with $X_1 = 1$, $X_2 = 0$, $Y = 1$
- n_{111} examples with $X_1 = 1$, $X_2 = 1$, $Y = 1$

(a) [2 pts] Solve for the maximum likelihood estimate (MLE) of the parameter p with respect to $n_{000}, n_{100}, n_{010}, n_{110}, n_{001}, n_{101}, n_{011}$, and n_{111}.

$$p = \frac{2n_{000}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{111}}{2(n_{000}+n_{001}+n_{010}+n_{011}+n_{100}+n_{101}+n_{110}+n_{111})}$$

We first write down the likelihood of the training data, $T = (Y, X_1, X_1)$.

$$L = \prod_{(y_i, x_{1_i}, x_{2_i})} P(Y = y_i, X_1 = x_{1_i}, X_2 = x_{2_i}|p, \lambda) =$$

$$= \prod_{(y_i, x_{1_i}, x_{2_i})} P(Y = y_i|p, \lambda)P(X_1 = x_{1_i}|p, \lambda)P(X_2 = x_{2_i}|p, \lambda) =$$

$$= \left(\prod_1^{n_{000}} \lambda pp\right)\left(\prod_1^{n_{001}} (1-\lambda)(1-p)(1-p)\right)\left(\prod_1^{n_{010}} \lambda p(1-p)\right)\left(\prod_1^{n_{011}} (1-\lambda)(1-p)p\right) *$$

$$* \left(\prod_1^{n_{100}} \lambda(1-p)p\right)\left(\prod_1^{n_{101}} (1-\lambda)p(1-p)\right)\left(\prod_1^{n_{110}} \lambda(1-p)(1-p)\right)\left(\prod_1^{n_{111}} (1-\lambda)pp\right) =$$

$$= (\lambda^{n_{000}+n_{010}+n_{100}+n_{110}})((1-\lambda)^{n_{001}+n_{011}+n_{101}+n_{111}})(p^{n_{000}+n_{010}+n_{101}+n_{111}})(p^{n_{000}+n_{011}+n_{100}+n_{111}}) *$$

$$*((1-p)^{n_{001}+n_{011}+n_{100}+n_{110}})((1-p)^{n_{001}+n_{010}+n_{101}+n_{110}}) =$$

$$= (\lambda^{n_{000}+n_{010}+n_{100}+n_{110}})((1-\lambda)^{n_{001}+n_{011}+n_{101}+n_{111}}) *$$

$$*(p^{2n_{000}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{111}})((1-p)^{2n_{001}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{110}})$$

We then take the logarithm of the likelihood.

$$\log(L) = (n_{000} + n_{010} + n_{100} + n_{110})\log(\lambda) + (n_{001} + n_{011} + n_{101} + n_{111})\log(1-\lambda) +$$

$$+(2n_{000} + n_{010} + n_{011} + n_{100} + n_{101} + 2n_{111})\log(p) + (2n_{001} + n_{010} + n_{011} + n_{100} + n_{101} + 2n_{110})\log(1-p)$$

We want to take the partial derivative with respect to p and solve for when it is 0 to find the MLE estimate of p. When we do this, the first two terms only depend on λ and not p, so their partial derivative is 0.

$$0 = \frac{\partial}{\partial p}(\log(L)) = (2n_{000}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{111})\frac{1}{p} - (2n_{001}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{110})\frac{1}{1-p}$$

Multiplying both sides by $(p)(1-p)$, we have:

$$0 = (2n_{000}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{111})(1-p) - (2n_{001}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{110})p$$

and simplifying:

$$(2n_{000}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{111}) = 2(n_{000}+n_{001}+n_{010}+n_{011}+n_{100}+n_{101}+n_{110}+n_{111})p$$

so

$$p = \frac{2n_{000}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{111}}{2(n_{000}+n_{001}+n_{010}+n_{011}+n_{100}+n_{101}+n_{110}+n_{111})}$$

(b) [2 pts] For each of the following values of λ, p, X_1, and X_2, classify the value of Y. Hint: No detailed calculation should be necessary.

λ	p	X_1	X_2	Y
3/4	5/8	0	0	0
2/5	4/7	1	0	1

For the first case, $P(Y=0,X_1=0,X_2=0) = \lambda pp$ and $P(Y=1,X_1=0,X_2=0) = (1-\lambda)(1-p)(1-p)$. Since $\lambda > (1-\lambda)$ and $p > (1-p)$, we must have $P(Y=0,X_1=0,X_2=0) > (Y=1,X_1=0,X_2=0)$.

For the second case, we have $P(Y=0,X_1=1,X_2=0) = \lambda(1-p)p$ and $P(Y=1,X_1=1,X_2=0) = (1-\lambda)p(1-p)$. Since both expressions have a $p(1-p)$ term, the question is reduced to $\lambda < (1-\lambda)$ so $P(Y=1,X_1=1,X_2=0) > P(Y=0,X_1=1,X_2=0)$.

(c) [1 pt] For the following value of λ, p, X_1, and X_2, classify the value of Y. Detailed calculation may be necessary.

λ	p	X_1	X_2	Y
3/5	3/7	0	0	1

$P(Y=0,X_1=0,X_2=0) = \lambda pp = \frac{3}{5}\frac{3}{7}\frac{3}{7} = \frac{27}{5*7*7}$ and $P(Y=1,X_1=0,X_2=0) = (1-\lambda)(1-p)(1-p) = \frac{2}{5}\frac{4}{7}\frac{4}{7} = \frac{32}{5*7*7}$. So $\frac{27}{5*7*7} = P(Y=0,X_1=0,X_2=0) < P(Y=1,X_1=0,X_2=0) = \frac{32}{5*7*7}$.

(d) [2 pts] Now let's consider a new model M2, which has the same Bayes' Net structure as M1, but where we have a p_1 value for $P(X_1=0 \mid Y=0) = P(X_1=1 \mid Y=1) = p_1$ and a separate p_2 value for $P(X_2=0 \mid Y=0) = P(X_2=1 \mid Y=1) = p_2$, and we don't constrain $p_1 = p_2$. Let L_{M1} be the likelihood of the training data under model M1 with the maximum likelihood parameters for M1. Let L_{M2} be the likelihood of the training data under model M2 with the maximum likelihood parameters for M2. Which of the following properties are guaranteed to be true?

- ○ $L_{M1} \geq L_{M2}$
- ● $L_{M1} \leq L_{M2}$
- ○ Insufficient information, the above relationships rely on the particular training data.
- ○ None of the above.

$M2$ can represent all of the same probability distributions that $M1$ can, but also some more (when $p_1 \neq p_2$). So in general $M2$ allows for more fitting of the training data, which results in a higher likelihood.

Q10. [10 pts] Perceptron

(a) [1 pt] Suppose you have a binary perceptron in 2D with weight vector $\mathbf{w} = r\ [w_1, w_2]^T$. You are given w_1 and w_2, and are given that $r > 0$, but otherwise not told what r is. *Assume that ties are broken as positive.*

Can you determine the perceptron's classification of a new example x with known feature vector $f(x)$?

● Always ○ Sometimes ○ Never

(b) Now you are learning a multi-class perceptron between 4 classes. The weight vectors are currently $[1, 0]^T$, $[0, 1]^T$, $[-1, 0]^T$, $[0, -1]^T$ for the classes A, B, C, and D. The next training example x has a **label of A** and feature vector $f(x)$.
For the following questions, *do not make any assumptions about tie-breaking.* (Do not write down a solution that creates a tie.)

(i) [1 pt] Write down a feature vector in which no weight vectors will be updated.

$f(x) = \begin{bmatrix} 1 \\ 0 \\ \\ \end{bmatrix}$ ○ Not possible

Any feature vector that points in the direction of w_A more than any other direction, such that $w_A \cdot f(x) > w_i \cdot f(x)$ for $i \neq A$.

(ii) [1 pt] Write down a feature vector in which **only** $\mathbf{w}_A$ will be updated by the perceptron.

$f(x) = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$ ● Not possible

(iii) [1 pt] Write down a feature vector in which **only** $\mathbf{w}_A$ and $\mathbf{w}_B$ will be updated by the perceptron.

$f(x) = \begin{bmatrix} 0 \\ 1 \\ \\ \end{bmatrix}$ ○ Not possible

Any feature vector that points in the direction of w_B more than any other direction, such that $w_B \cdot f(x) > w_i \cdot f(x)$ for $i \neq B$.

(iv) [1 pt] Write down a feature vector in which **only** $\mathbf{w}_A$ and $\mathbf{w}_C$ will be updated by the perceptron.

$f(x) = \begin{bmatrix} -1 \\ 0 \\ \\ \end{bmatrix}$ ○ Not possible

Any feature vector that points in the direction of w_C more than any other direction, such that $w_C \cdot f(x) > w_i \cdot f(x)$ for $i \neq C$.

The weight vectors are the same as before, but now there is a bias feature with value of 1 for all x and the weight of this bias feature is $0, -2, 1, -1$ for classes A, B, C, and D respectively. As before, the next training example x has a **label of A** and a feature vector $f(x)$. The always "1" bias feature is the first entry in $f(x)$.

(v) [1 pt] Write down a feature vector in which **only** $\mathbf{w}_B$ and $\mathbf{w}_C$ will be updated by the perceptron.

$f(x) = \begin{bmatrix} 1 \\ \\ \\ \end{bmatrix}$ ● Not possible

(vi) [1 pt] Write down a feature vector in which **only** $\mathbf{w}_A$ and $\mathbf{w}_C$ will be updated by the perceptron.

$f(x) = \begin{bmatrix} 1 \\ 0 \\ \\ 0 \end{bmatrix}$ ○ Not possible

Any feature vector that points in the direction of w_C more than any other direction, such that $w_C \cdot f(x) > w_i \cdot f(x)$ for $i \neq C$.

(c) Suppose your training data is linearly separable and you are classifying between label Y and label Z. The mistake bound ($\frac{k}{\delta^2}$) is equal to 3, which is the maximum number of weight vector updates the perceptron might have to do before it is guaranteed to converge. There are 100 examples in your training set. Assume the perceptron cycles through the 100 examples in a fixed order.

(i) [1 pt] What is the maximum number of classifications the perceptron might make before it is guaranteed to have converged?

● 300 ○ 103 ○ 100^3 ○ 3 ○ 3^{100} ○ None of the options

The perceptron will make at most 3 mistakes. When it converges, it will be correct for all of the training data. Therefore, in the worst case, it will have to pass through all 100 examples 3 times (with the last mistake being made on the last example in the 3rd pass) before converging. (Note that in this worst case, the algorithm might continue to classify 100 more examples beyond the 300 before *knowing* that it has converged.)

(ii) [2 pts] [*true* or *false*] After convergence, the learned perceptron will correctly classify *all* training examples.

If the data is linearly separable (which it is because the margin is non-zero), perceptrons converge when all of the training examples are correctly classified.