CS 188 Spring 2014 Introduction to Artificial Intelligence Midterm 2

- You have approximately 2 hours and 50 minutes.
- The exam is closed book, closed notes except your two-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. All short answer sections can be successfully answered in a few sentences AT MOST.

First name	
Last name	
SID	
edX username	

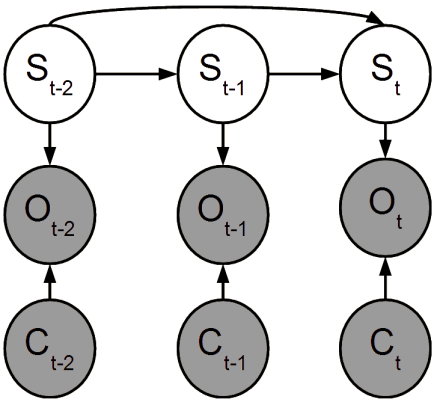
First and last name of student to your left	
First and last name of student to your right	

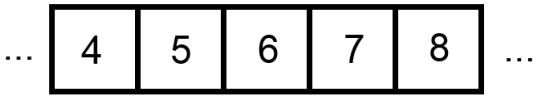
For staff use only:

Q1.	Particle Filtering	/12
Q2.	Probabilities	/15
Q3.	Need That CPT	/15
Q4.	Fun with Bayes Nets	/14
Q5.	Value of Gambling and Bribery	/12
Q6.	Bayes Net Modeling	/12
Q7.	Most Likely Estimates in HMMs	/20
	Total	/100

THIS PAGE IS INTENTIONALLY LEFT BLANK

Q1. [12 pts] Particle Filtering





Pacman is trying to hunt a ghost in an infinite hallway with positions labeled as in the picture above. He's become more technologically savvy, and decided to locate find the ghosts actual position, $\mathbf{S_t}$, using some sensors he set up. From the sensors, Pacman can find, at each time step, a noisy reading of the ghost's location, $\mathbf{O_t}$. However, just as Pacman has gained technology, so has the ghost. It is able to cloak itself at each time step, given by $\mathbf{C_t}$, adding extra noise to Pacman's sensor readings.

Pacman has generated an error model, given in the table below, for the sensor depending on whether the ghost is cloaked or not.

Pacman has also generated a dynamics model, given in the table below, that takes into account the position of the ghost at the two previous timesteps.

Dynamics model:

$P(S_t|S_{t-1}, S_{t-2}) = F(D_1, D_2)$
$D_1 = |S_t - S_{t-1}|$
$D_2 = |S_t - S_{t-2}|$

D_1	D_2	F(D_1, D_2)
0	0	0.7
0	1	0.2
0	2	0
1	0	0.3
1	1	0.3
1	2	0.5

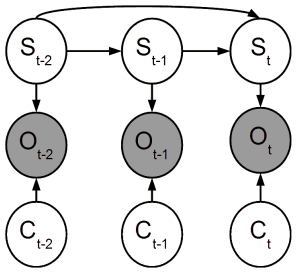
Observation model:

$P(O_t|S_t, C_t) = E(C_t, D)$
$D = |O_t - S_t|$

C	D	E(C, D)
+	0	0.4
+	1	0.2
+	2	0.1
-	0	0.6
-	1	0.2
-	2	0

(a) [2 pts] Assume that you currently have the following two particles: $(S_6 = 7, S_7 = 8)$ and $(S_6 = 6, S_7 = 6)$. Compute the weights for each particle given the observations $C_6 = +, C_7 = -, O_6 = 5$, $O_7 = 8$:

$(S_6 = 7, S_7 = 8)$	$Pr(O_6 = 5\|C_6 = +, S_6 = 7) * Pr(O_7 = 8\|C_7 = -, S_7 = 8) = 0.1 * 0.6 = 0.06$
$(S_6 = 6, S_7 = 6)$	$Pr(O_6 = 5\|C_6 = +, S_6 = 6) * Pr(O_7 = 8\|C_7 = -, S_7 = 6) = 0.2 * 0 = 0$



C	P(C)
+	0.5
-	0.5

(b) [4 pts] Assume that Pacman can no longer see whether the ghost is cloaked or not, but assumes that it will be cloaked at each timestep with probability 0.5. Compute the weights for each particle given the observations $O_6 = 5$, $O_7 = 8$:

For each of the particle's states: we want to find $Pr(o_t|s_t)$, as this is the contribution to the weight of the sample. However, we have C unobserved, so: $Pr(o_t|s_t) = \sum_{c_t} Pr(o_t, c_t|s_t) = \sum_{c_t} Pr(c_t|s_t)Pr(o_t|s_t, c_t) = \sum_{c_t} Pr(c_t)Pr(o_t|s_t, c_t)$. Last equality is due to independence between C_t and S_t.

$(S_6 = 7, S_7 = 8)$	$\sum_{c_6} Pr(c_6)Pr(O_6 = 5\|S_6 = 7, c_6) * \sum_{c_7} Pr(c_7)Pr(O_7 = 8\|S_7 = 8, c_7) = 0.05 * 0.5 = 0.025$
$(S_6 = 6, S_7 = 6)$	$\sum_{c_6} Pr(c_6)Pr(O_6 = 5\|S_6 = 6, c_6) * \sum_{c_7} Pr(c_7)Pr(O_7 = 8\|S_7 = 6, c_7) = 0.2 * 0.05 = 0.01$

(c) [4 pts] To prevent error propagation, assume that after weighting the particles and resampling, one of the particles you end up with is $(S_6 = 6, S_7 = 7)$.

(i) [2 pts] What is the probability that after passing this particle through the dynamics model it becomes $(S_7 = 6, S_8 = 6)$?
0. It's invalid for a particle to start at $S_7 = 7$ and after one transition, become $S_7 = 6$.

(ii) [2 pts] What is the probability the particle becomes $(S_7 = 7, S_8 = 8)$?
0.5. This is just $Pr(S_8 = 8|S_6 = 6, S_7 = 7) = F(D_1 = 1, D_2 = 2) = 0.5$.

(d) [2 pts] To again decouple this part from previous parts, assume that you have the following three particles with the specified weights.

Particle	weight
$(S_7 = 5, S_8 = 6)$	.1
$(S_7 = 7, S_8 = 6)$	.25
$(S_7 = 7, S_8 = 7)$	.3

What is Pacman's belief for the ghost's position at time $t = 8$?

Position	P(S_8)
$S_8 = 5$	$\frac{0}{.1+.25+.3} = 0$
$S_8 = 6$	$\frac{.1+.25}{.1+.25+.3} = \frac{.35}{.65} = \frac{7}{13}$
$S_8 = 7$	$\frac{.3}{.1+.25+.3} = \frac{.3}{.65} = \frac{6}{13}$
$S_8 = 8$	$\frac{0}{.1+.25+.3} = 0$

Q2. [15 pts] Probabilities

(a) [3 pts] Fill in the circles of **all** expressions that are equal to **1**, **given no independence assumptions**:

- ● $\sum_a P(A = a \mid B)$
- ○ $\sum_b P(A \mid B = b)$
- ● $\sum_a \sum_b P(A = a, B = b)$
- ○ $\sum_a \sum_b P(A = a \mid B = b)$
- ● $\sum_a \sum_b P(A = a)\ P(B = b)$
- ○ $\sum_a P(A = a)\ P(B = b)$
- ○ None of the above.

Probability distributions, including conditional distributions, sum to one. We are testing this axiom when applied to multivariate distributions and conditionals.

(b) [3 pts] Fill in the circles of **all** expressions that are equal to $\mathbf{P(A, B, C)}$, **given no independence assumptions**:

- ● $P(A \mid B, C)\ P(B \mid C)\ P(C)$
- ○ $P(C \mid A, B)\ P(A)\ P(B)$
- ● $P(A, B \mid C)\ P(C)$
- ● $P(C \mid A, B)\ P(A, B)$
- ○ $P(A \mid B)\ P(B \mid C)\ P(C)$
- ○ $P(A \mid B, C)\ P(B \mid A, C)\ P(C \mid A, B)$
- ○ None of the above.

We are testing the chain rule when applied to more than two variables.

(c) [3 pts] Fill in the circles of **all** expressions that are equal to $\mathbf{P(A \mid B, C)}$, **given no independence assumptions**:

- ● $\frac{P(A,B,C)}{\sum_a P(A=a,B,C)}$
- ● $\frac{P(B,C|A)\ P(A)}{P(B,C)}$
- ● $\frac{P(B|A,C)\ P(A|C)}{P(B|C)}$
- ○ $\frac{P(B|A,C)\ P(A|C)}{P(B,C)}$
- ○ $\frac{P(B|A,C)\ P(C|A,B)}{P(B,C)}$
- ○ $\frac{P(A,B|C)}{P(B|A,C)}$
- ○ None of the above.

This is Bayes' rule applied to distributions over multiple variables. $P(A \mid B, C) = P(A, B, C)/P(B, C)$

(d) [3 pts] Fill in the circles of **all** expressions that are equal to $\mathbf{P(A \mid B)}$, **given that** $\mathbf{A \perp\!\!\!\perp B \mid C}$:

- ○ $\frac{P(A|C)\ P(B|C)}{P(B)}$
- ○ $\frac{P(A|C)\ P(B|C)}{P(B|C)}$
- ● $\frac{\sum_c P(A|C=c)\ P(B|C=c)\ P(C=c)}{\sum_{c'} P(B|C=c')\ P(C=c')}$
- ○ $\frac{P(A|B,C)}{P(A|C)}$
- ● $\frac{\sum_c P(B|A,C=c)\ P(A,C=c)}{P(B)}$
- ○ $\frac{\sum_c P(A,C=c)\ P(B|C=c)}{\sum_{c'} P(A,B,C=c')}$
- ○ None of the above.

Apply Bayes' rule to get $P(A \mid B) = P(A, B)/P(B) = \sum_c P(A, B \mid C = c)P(C = c)/P(B)$ and conditional independence $P(A, B \mid C) = P(A \mid C)\ P(B \mid C)$

(e) [3 pts] Fill in the circles of **all** expressions that are equal to $\mathbf{P(A, B, C)}$, **given that $\mathbf{A \perp\!\!\!\perp B \mid C}$ and $\mathbf{A \perp\!\!\!\perp C}$**:

- ○ $P(A)\ P(B)\ P(C)$
- ● $P(A)\ P(B, C)$
- ● $P(A \mid B)\ P(B \mid C)\ P(C)$
- ○ $P(A \mid B, C)\ P(B \mid A, C)\ P(C \mid A, B)$
- ● $P(A \mid C)\ P(B \mid C)\ P(C)$
- ○ $P(A \mid C)\ P(B \mid C)$
- ○ None of the above.

If $A \perp\!\!\!\perp B \mid C$ and $A \perp\!\!\!\perp C$ it can be proven that $A \perp\!\!\!\perp B$ but not that $B \perp\!\!\!\perp C$. Here is the proof:

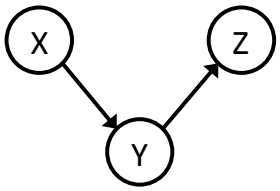
$$
\begin{aligned}
P(A, B) &= \sum_c P(A, B \mid C = c) P(C = c) \\
&= \sum_c P(A \mid C = c) P(B \mid C = c) P(C = c) \\
&= P(A) \sum_c P(B \mid C = c) P(C = c) \\
&= P(A) \sum_c P(B, C = c) \\
&= P(A)\ P(B)
\end{aligned}
$$

Q3. [15 pts] Need That CPT

For each of the following questions, mark the **minimum** set of variables whose associated probability table is **needed** to answer the query.

For example, for part (a), the probability table associated with X is $P(X)$, the probability table associated with Y is $P(Y \mid X)$, and the probability table associated with Z is $P(Z \mid Y)$. The query for part (a) (i) is $P(X \mid Y)$.

(a) [3 pts]



(i) [1 pt] $P(X \mid Y)$ ● X ● Y ○ Z

$P(X \mid Y) \propto_Y P(Y \mid X)P(X)$

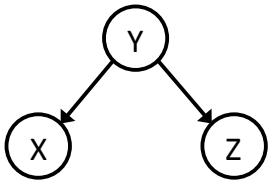
(ii) [1 pt] $P(X \mid Z)$ ● X ● Y ● Z

Y is in between X and Z in the Bayes' net, so the distribution for Y is needed to compute the query.

(iii) [1 pt] $P(Y \mid Z)$ ● X ● Y ● Z

Like in part (i), we compute the query using Bayes' rule. However, because we do not directly know $P(Y)$, we need to sum out over X.

(b) [3 pts]



(i) [1 pt] $P(X \mid Y)$ ● X ○ Y ○ Z

$P(X \mid Y)$ is exactly the probability table for X.

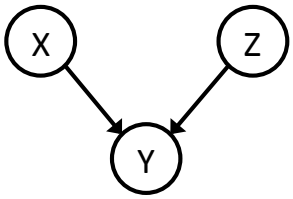
(ii) [1 pt] $P(X \mid Z)$ ● X ● Y ● Z

Y is in between X and Z in the Bayes' net, so the distribution for Y is needed to compute the query.

(iii) [1 pt] $P(Z \mid X)$ ● X ● Y ● Z

Same as in part (ii).

(c) [3 pts]



(i) [1 pt] $P(X \mid Y)$ ● X ● Y ● Z

Because we do not directly know $P(Y \mid X)$, we need to sum out over Z to calculate it.

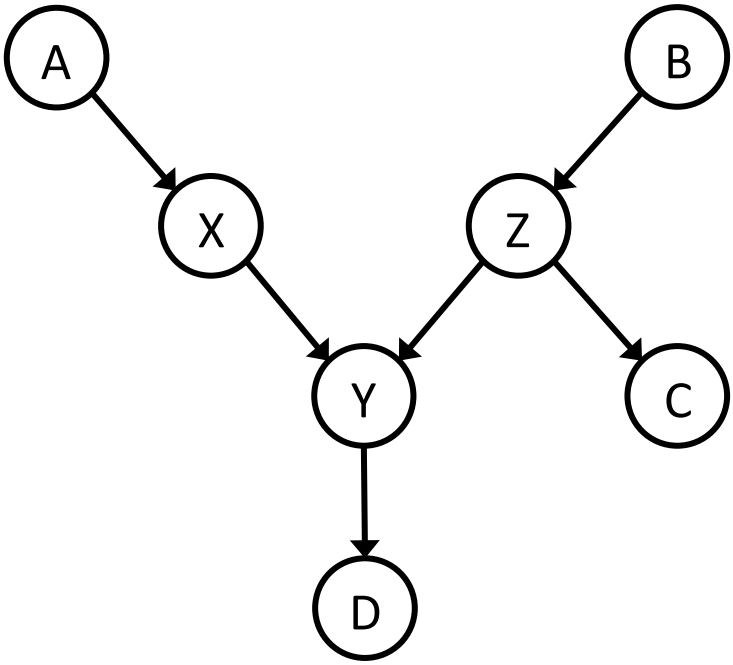
(ii) [1 pt] $P(X \mid Z)$ ● X ○ Y ○ Z

X is independent from Z, so $P(X \mid Z) = P(X)$.

(iii) [1 pt] $P(Y \mid Z)$ ● X ● Y ○ Z

$\Sigma_x P(X)P(Y \mid X, Z) = \Sigma_x P(X \mid Z)P(Y \mid X, Z) = \Sigma_x P(X, Y \mid Z) = P(Y \mid Z)$. Therefore, we only need the probability tables for $P(X)$ and $P(Y \mid X, Z)$.

(d) [6 pts]



(i) [2 pts] $P(X \mid Z)$ ● X ○ Y ○ Z ● A ○ B ○ C ○ D

X is independent from Z, so we only need to calculate $P(X)$.

(ii) [2 pts] $P(X \mid D)$ ● X ● Y ● Z ● A ● B ○ C ● D

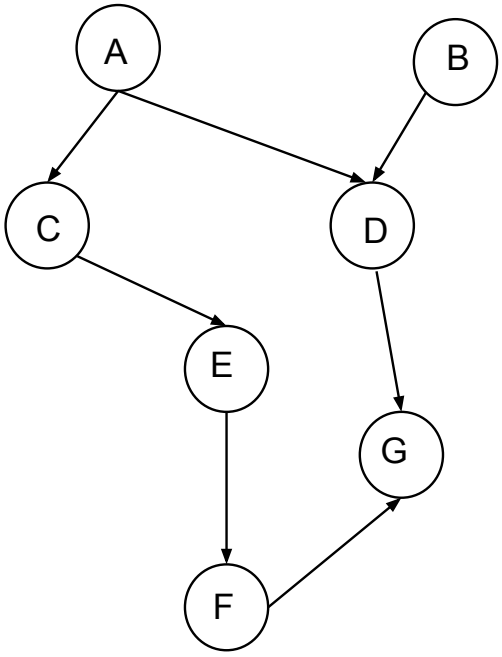
Eliminating C requires us to join and sum over all factors involving c, which is $\sum_c Pr(c \mid Z)$. This factor's value is 1 for any value of Z, since $Pr(c \mid Z)$ is a distribution.

(iii) [2 pts] $P(X \mid Z, D)$ ● X ● Y ○ Z ● A ○ B ○ C ● D

As in part (c)(iii), we do not need to sum out over Z in order to compute the query. Therefore, we don't need the probability table for Z, and as B and C are independent from X given Z, we don't need their probability tables either.

Q4. [14 pts] Fun with Bayes Nets

(a) [6 pts] Consider a Bayes Net with the following graph:

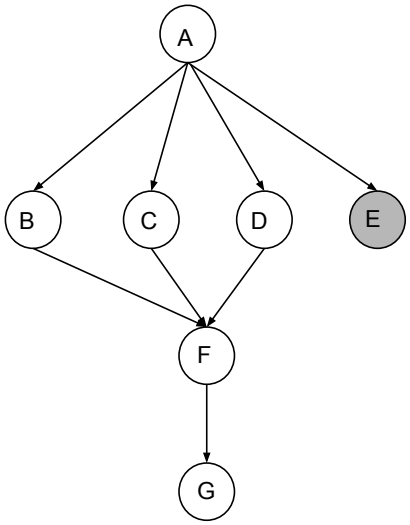


Which of the following are guaranteed to be true without making any additional conditional independence assumptions, other than those implied by the graph? (Mark all true statements)

- [x] $P(A \mid C, E) = P(A \mid C)$
- [ ] $P(A, E \mid G) = P(A \mid G) * P(E \mid G)$
- [x] $P(A \mid B = b) = P(A)$
- [ ] $P(A \mid B, G) = P(A \mid G)$
- [ ] $P(E, G \mid D) = P(E \mid D) * P(G \mid D)$
- [x] $P(A, B \mid F) = P(A \mid F) * P(B \mid F)$

This question deals with (conditional) independence of a Bayes Net.
Option 1: $A \perp\!\!\!\perp E \mid C$, since with C observed no path between A and C is active.
Option 2: there's no conditional independence between A and E given G, since A-C-E is active.
Option 3: $A \perp\!\!\!\perp B$, since no path between A and B is active.
Option 4: there's no conditional independence between A and B given G, since the path A, B, G is an active.
Option 5: there's no conditional independence between E and G, since E-F-G is active.
Option 6: $A \perp\!\!\!\perp B \mid F$, since A, B, F is active.

(b) [8 pts] Now consider a Bayes Net with the following graph:



The factors associated with the Bayes Net are $P(A)$, $P(B \mid A)$, $P(C \mid A)$, $P(D \mid A)$, $P(E \mid A)$, $P(F \mid B, C, D)$ and $P(G \mid F)$. We will consider variable elimination to answer the query $P(G \mid +e)$.

(i) [2 pts] Suppose the first variable we eliminate is A. Provide an expression for the resulting factor as a function of the original factors.

$f_1(B, C, D, +e) = \sum_a P(a)P(B \mid a)P(C \mid a)P(D \mid a)P(+e \mid a)$

(ii) [2 pts] Suppose the first variable we eliminate is B. Provide an expression for the resulting factor generated as a function of the original factors.

$f_2(A, F, C, D) = \sum_b P(b \mid A)P(F \mid b, C, D)$

(iii) [2 pts] Suppose the first variable we eliminate is F. Provide an expression for the resulting factor as a function of the original factors.

$f_3(B, C, D, G) = \sum_f P(f \mid B, C, D)P(G \mid f)$

(iv) [2 pts] Suppose we eliminated the variables A, B, C, D, F in that order, and the single remaining factor is $f(+e, G)$. How do we obtain $P(G \mid +e)$ from this remaining factor? (Your answer should be in the form of an equation.)

$P(G \mid +e) = \frac{f(+e,G)}{\sum_g f(+e,g)}$

Q5. [12 pts] Value of Gambling and Bribery

The local casino is offering a new game. There are two biased coins that are indistinguishable in appearance. There is a *head-biased* coin, which yields head with probability 0.8 (and tails with probability 0.2). There is a *tail-biased* coin, which yields tail with probability 0.8 (and head with probability 0.2).

At the start of the game, the dealer gives you one of the two coins at random, with equal probability. You get to flip that coin once. Then you decide if you want to stop or continue. If you choose to continue, you flip it 10 more times. In those 10 flips, each time it yields head, you get \$1, and each time it yields tail, you lose \$1.

(a) [1 pt] What is the expected value of your earnings if continuing to play with a head-biased coin?

$10 \cdot (0.8 \cdot 1 + 0.2 \cdot -1) = 6$

(b) [1 pt] What is the expected value of your earnings if continuing to play with a tail-biased coin?

$10 \cdot (0.8 \cdot -1 + 0.2 \cdot 1) = -6$

(c) [3 pts] Suppose the first flip comes out head.

(i) [1 pt] What is the posterior probability that the coin is *head-biased*? 4/5

$$\frac{P(hb|heads)}{P(tb|heads)} = \frac{P(heads|hb)P(hb)/P(heads)}{P(heads|tb)P(tb)} = \frac{P(heads|hb)P(hb)}{P(heads|tb)P(tb)} = \frac{(.8)\cdot(1/2)}{(.2)\cdot(1/2)} = 4 \quad (1)$$

Thus $\frac{P(hb|heads)}{P(hb|heads)+P(hb|tails)} = \frac{4}{1+4} = \frac{4}{5}$.

(ii) [1 pt] What is the expected value of your earnings for continuing to play? $\frac{4}{5} \cdot 6 + \frac{1}{5} \cdot -6 = 3.6$

(iii) [1 pt] Which is the action that maximizes the expected value of your earnings? ● Continue ○ Stop

(d) Suppose the first flip comes out tail.

(i) [1 pt] What is the posterior probability that the coin is *tail-biased*? 4/5

(ii) [1 pt] What is the expected value of your earnings for continuing to play? -3.6

(iii) [1 pt] Which is the action that maximizes the expected value of your earnings? ○ Continue ● Stop

(e) [1 pt] What is the expected value of your earnings after playing the game optimally one time?

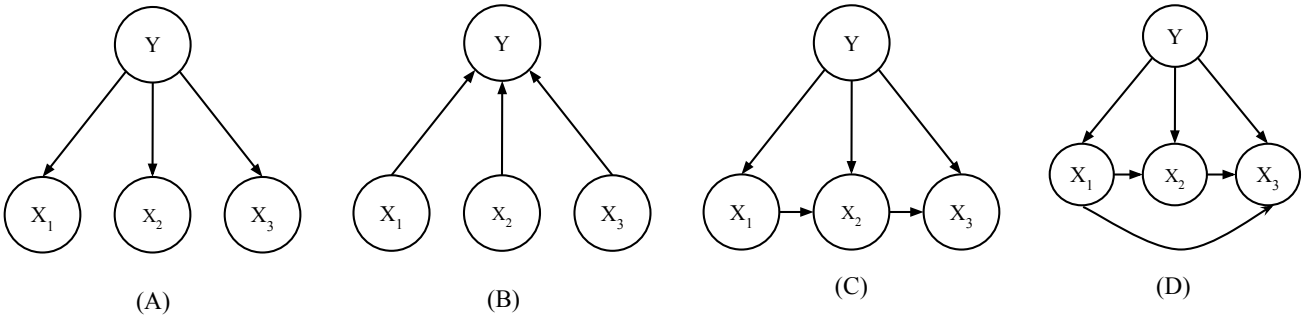
$0.5 \cdot 3.6 + .5 \cdot 0 = 1.8$

(f) [3 pts] Suppose again that the first flip yields head. The dealer knows which coin you picked. How much are you willing to pay the dealer to find out the type of the coin?

6/5 You only change your action (from continue to stop) if the dealer tells you the coin is tails biased. The probability that it's a tails-biased coin is 1/5. The expected returns from a tails-biased coin is -6, so your payout improves by 6 by switching from continue to stop. So the answer is $\frac{1}{5} \cdot 6$.

Q6. [12 pts] Bayes Net Modeling

Abhishek has been getting a lot of spam recently(!) and is not satisfied with his email client's Naive Bayes spam classifier. Thankfully he knows about Bayes Nets and has decided to implement his own spam classifier. It is your job to help him model his Bayes Nets and train them. The following are 4 Bayes Nets he is considering. Each variable can take on the values {0, 1}.



(a) [6 pts] Abhishek wants to know how much memory he needs to store each of these Bayes Nets on disk. The amount of memory depends on the number of values he would need to store in the CPTs for that Bayes Net. For each of the nets above give the **least** number of parameters he would need to store to completely specify the Bayes Net.

A = 7 C = 11

B = 11 D = 15

The number of parameters in the Bayes Net is the total number of probability values you need in the CPTs of the Bayes Net. You also need to remember that if A takes k values, $P(A)$ can be represented by $k-1$ values as the last value can be chosen so that the probabilities sum to 1. For example in (A), $P(X_1|Y=0)$ has 1 parameter and so does $P(X_1|Y=1)$ and thus total number of parameters are $2+2+2+1=7$

(b) It's now time to train the Bayes Nets. Abhishek has training datasets D_l, D_m, D_s and a test dataset D_t. The number of training examples in each set vary as $|D_l| > |D_m| > |D_s|$. e_{tr} and e_{te} represent train and test errors and their arguments represent which dataset a model was trained on. For example, $e_{tr}(A, D_l)$ refers to the training error of model A on dataset D_l.

(i) [4 pts] Abhishek tries a bunch of experiments using model D. In a typical scenario[1] , which of the following can be expected to be true? (Mark all that apply)

- ● $e_{tr}(D, D_l) \geq e_{tr}(D, D_m) \geq e_{tr}(D, D_s)$
- ○ $e_{tr}(D, D_s) \geq e_{tr}(D, D_m) \geq e_{tr}(D, D_l)$
- ○ $e_{te}(D, D_l) \geq e_{te}(D, D_m) \geq e_{te}(D, D_s)$
- ● $e_{te}(D, D_s) \geq e_{te}(D, D_m) \geq e_{te}(D, D_l)$
- ○ $e_{tr}(D, D_l) \geq e_{te}(D, D_l)$
- ● $e_{tr}(D, D_l) \leq e_{te}(D, D_l)$
- ○ $e_{tr}(D, D_s) \geq e_{te}(D, D_s)$
- ● $e_{tr}(D, D_s) \leq e_{te}(D, D_s)$

[1] Train and test data sampled from same underlying distribution

(ii) [2 pts] Abhishek is now trying to compare performance across different models. In a typical scenario, which of the following can be expected to be true? (Mark all that apply)

- ● $e_{tr}(A, D_l) \geq e_{tr}(B, D_l) \geq e_{tr}(D, D_l)$
- ○ $e_{tr}(A, D_l) \leq e_{tr}(B, D_l) \leq e_{tr}(D, D_l)$
- ● $e_{te}(A, D_l) \geq e_{te}(B, D_l) \geq e_{te}(D, D_l)$
- ○ $e_{te}(A, D_l) \leq e_{te}(B, D_l) \leq e_{te}(D, D_l)$

The rationale is to test knowledge about "power" of a model and overfitting/underfitting. Bayes Nets with less number of parameters have low capacity/power as they can model a more restricted class of distributions due to greater independence assumptions. A more powerful model would overfit on less data and a weak model would underfit. In this question, we are assuming that D_l is *very* large. Rationale for correct answers(column major):

(i) - 1: A larger dataset would be tougher to model than a smaller. Thus training error would be greater for a larger dataset.

(i) - 4: A smaller dataset would overfit for a complex model like D and thus testing errors would be more.

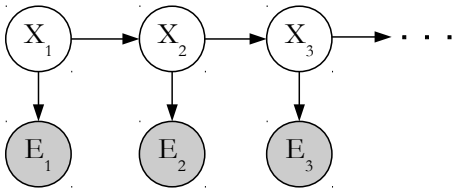
(i) - 6: Training error less than test error in general because we explicitly model the training data and need to generalize to test data.

(i) - 8: Overfitting for smaller data and complex model

(ii) - 1: Weaker model underfits on large dataset

(ii) - 3: Same as above. Weaker model cant model the distribution well enough if we have a lot of data.

Q7. [20 pts] Most Likely Estimates in HMMs



The Viterbi algorithm finds the most probable sequence of hidden states $X_{1:T}$, given a sequence of observations $e_{1:T}$. Throughout this question you may assume there are no ties. Recall that for the canonical HMM structure, the Viterbi algorithm performs the following **dynamic programming**[2] computations:

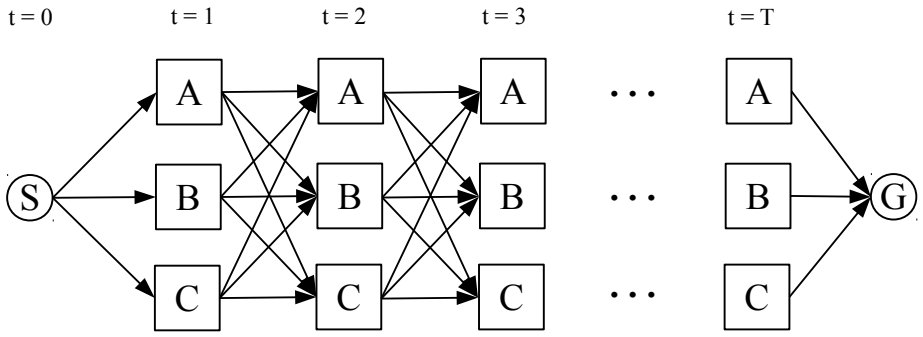
$$m_t[x_t] = P(e_t|x_t) \max_{x_{t-1}} P(x_t|x_{t-1}) m_{t-1}[x_{t-1}]$$

(a) [3 pts] For the HMM structure above, which of the following probabilities are maximized by the sequence of states returned by the Viterbi algorithm? Mark **all** the correct option(s).

- ○ $P(X_{1:T})$
- ○ $P(X_T|e_T)$
- ● $P(X_{1:T}|e_{1:T})$
- ● $P(X_{1:T}, e_{1:T})$
- ● $P(X_1)P(e_1|X_1)\prod_{t=2}^{T} P(e_t|X_t)P(X_t|X_{t-1})$
- ○ $P(X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})$
- ○ None of the above

The sequence of states returned by the Viterbi Algorithm maximizes the conditional = $P(X_{1:T}|e_{1:T})$. Since, $P(e_{1:T})$ is just a constant, $P(X_{1:T}, e_{1:T}) \propto P(X_{1:T}|e_{1:T})$. Hence, the full joint $P(X_{1:T}, e_{1:T})$ is also maximized. The third option $P(X_1)P(e_1|X_1)\prod_{t=2}^{T} P(e_t|X_t)P(X_t|X_{t-1}) == P(X_{1:T}, e_{1:t})$ due to the conditional independences implied by the HMM structure; therefore, this is also maximized.

(b) Consider an HMM structure like the one in part (a) above. Say for all time steps t, the state X_t can take on one of the three values $\{A, B, C\}$. Then, we can represent the state transitions through the following directed graph, also called a *Trellis Diagram*.



We wish to formulate the most probable sequence of hidden state query as a **graph search problem**.

Note in the diagram above, dummy nodes S and G have been added to represent the *start* state and the *goal* state respectively. Further, the transition from the starting node S to the first state X_1 occurs at time step $t = 0$; transition from X_T (the last HMM state) to the goal state G occurs at time step $t = T$.

(The questions for this section are on the following page)

[2] If you're not familiar with dynamic programming, it is essentially a recursive relation in which the current value is defined as a function of previously computed values. In this case, the value at time t is defined as a function of the values at time $t - 1$

Definition : Let $w^t_{Y\to Z}$, be the cost of the edge for the transition from state Y at time t to state Z at time $t+1$. For example, $w^1_{A\to B}$ is the cost of the edge for transition from state A at time 1 to state B at time 2.

(i) [3 pts] For which **one** of the following values for the weights $w^t_{Y\to Z}$, $1 \le t < T$, would the minimum cost path be exactly the same as most likely sequence of states computed by the Viterbi algorithm.

- ○ $w^t_{Y\to Z} = -P(X_{t+1} = Z|X_t = Y)$
- ○ $w^t_{Y\to Z} = -P(e_{t+1}|X_{t+1} = Z)P(X_{t+1} = Z|X_t = Y)$
- ○ $w^t_{Y\to Z} = -\log(P(X_{t+1} = Z|X_t = Y))$
- ● $w^t_{Y\to Z} = -\log(P(e_{t+1}|X_{t+1} = Z)P(X_{t+1} = Z|X_t = Y))$
- ○ $w^t_{Y\to Z} = \dfrac{1}{P(X_{t+1} = Z|X_t = Y)}$
- ○ $w^t_{Y\to Z} = \dfrac{1}{P(e_{t+1}|X_{t+1} = Z)P(X_{t+1} = Z|X_t = Y)}$

We want the solution to maximize the joint $P(X_{1:T}, e_{1:T}) = P(X_1)P(e_1|X_1)\prod_{t=2}^{T} P(e_t|X_t)P(X_t|X_{t-1})$.
Since, a search algorithm **minimizes** the cost, we want to pose this problem as a minimization problem. Hence, we can equivalently say that we want to minimize $\dfrac{1}{P(X_1)P(e_1|X_1)\prod_{t=2}^{T} P(e_t|X_t)P(X_t|X_{t-1})}$.
A search algorithm in its native form can only work with **additive** costs. Therefore, to turn the above products of probabilities into sums, we take the log.

Hence, we wish to minimize :

$$\log\left(\frac{1}{P(X_1)P(e_1|X_1)\prod_{t=2}^{T} P(e_t|X_t)P(X_t|X_{t-1})}\right) = -\log\left(P(X_1)P(e_1|X_1)\prod_{t=2}^{T} P(e_t|X_t)P(X_t|X_{t-1})\right)$$

$$= -\log\left(P(X_1)P(e_1|X_1)\right) - \sum_{t=2}^{T}\log\left(P(e_t|X_t)P(X_t|X_{t-1})\right).$$

If for $t > 1$, the edge cost is set to $-\log P(e_{t+1}|X_{t+1} = Z)P(X_{t+1} = Z|X_t = Y)$, we get the above as the total cost of the path.

(ii) [3 pts] The initial probability distribution of the state at time $t = 1$ is given $P(X_1 = Y), Y \in \{A, B, C\}$.

Which **one** of the following should be the value of $w^0_{S\to Y}, Y \in \{A, B, C\}$ — these are the cost on the edges connecting S to the states at time $t = 1$?

- ○ $w^0_{S\to Y} = -P(X_1 = Y)$
- ○ $w^0_{S\to Y} = -P(e_1|X_1 = Y)P(X_1 = Y)$
- ○ $w^0_{S\to Y} = -\log(P(X_1 = Y))$
- ● $w^0_{S\to Y} = -\log(P(e_1|X_1 = Y)P(X_1 = Y))$
- ○ $w^0_{S\to Y} = \dfrac{1}{P(X_1 = Y)}$
- ○ $w^0_{S\to Y} = \dfrac{1}{P(e_1|X_1 = Y)P(X_1 = Y)}$
- ○ $w^0_{S\to Y} = \alpha$, $\alpha \in \mathbb{R}$: (some constant)

The reasoning is essentially the same as the previous question. One exception is that for the first node X_1, there is no state transition probability. Instead, we use the prior $P(X_1)$. Therefore, we modify the answer to the previous question to use the prior probability instead.

(iii) [3 pts] Which **one** of the following should be the value of $w^T_{Y\to G}, Y \in \{A, B, C\}$ — these are the cost on the edges connecting the states at last time step $t = T$ to the goal state G?

- ○ $w^T_{Y\to G} = -P(X_T = Y)$
- ○ $w^T_{Y\to G} = -P(e_T|X_T = Y)P(X_T = Y)$
- ○ $w^T_{Y\to G} = -\log(P(X_T = Y))$
- ○ $w^T_{Y\to G} = -\log(P(e_T|X_T = Y)(P(X_T = Y))$
- ○ $w^T_{Y\to G} = \dfrac{1}{P(X_T = Y)}$
- ○ $w^T_{Y\to G} = \dfrac{1}{P(e_T|X_T = Y)P(X_T = Y)}$
- ● $w^T_{Y\to G} = \alpha,\ \alpha \in \mathbb{R}$: (some constant)

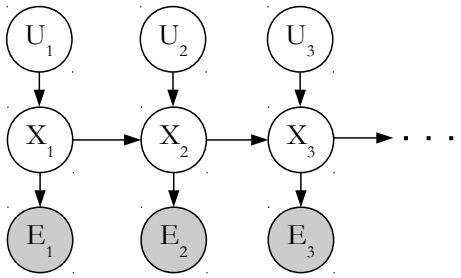
As long as the costs on all the three edges : $w^T_{A\to G}, w^T_{B\to G}, w^T_{C\to G}$ is the same, we will get the optimal answer. Note, it does not matter if the constant α is postive or negative because the total probability is only scaled by a positive constant $= e^{\alpha}$ (that is, the total cost of the path is (`sum of all log probabilities`) $+\ w^T_{Y\to G}(== \alpha)$. When you exponentiate this to get the probabilities, you get $e^{\text{sum of log probabilities}} * e^{\alpha}$ — which is just a scaling by some positive number).

(c) We consider extending the Viterbi algorithm for finding the most likely sequence of states in modified HMMs.

For your convenience, the computations performed by the Viterbi algorithm for the canonical HMM structure, like in part (a), are repeated below:

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

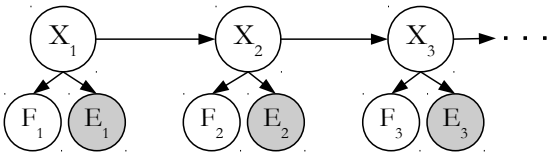
(i) [4 pts] Consider the HMM below with additional variables U_t. The HMM can be interpreted as : The state X_t at time t is caused due to some action U_t and previous state X_{t-1}. The state X_t emits an observation E_t. Both U_t and X_t are unobserved.



We want to find the most likely sequence of states $X_{1:T}$ and actions $U_{1:T}$, given the sequence of observations $e_{1:T}$. Write a dynamic programming update for $t > 1$ **analogous** to the one for the canonical HMM structure.

$m_t[x_t, u_t] =$ $P(e_t|x_t)P(u_t) \max_{x_{t-1},u_{t-1}} P(x_t|u_t, x_{t-1})m[x_{t-1}, u_{t-1}]$

(ii) [4 pts] Consider the HMM below with two emission variables at each time step F_t and E_t. E_t is observed while X_t and F_t are unobserved.



We want to find the most likely sequence of states $X_{1:T}$ and the unobserved emissions $F_{1:T}$, given the sequence of observations $e_{1:T}$. Write a dynamic programming update for $t > 1$ **analogous** to the one for the canonical HMM structure.

$m_t[x_t, f_t] =$ $P(e_t|x_t)P(f_t|x_t) \max_{x_{t-1},f_{t-1}} P(x_t|x_{t-1})m[x_{t-1}, f_{t-1}]$