HOMEWORK 8:
GRAPHICAL MODELS

CMU 10601: MACHINE LEARNING (SPRING 2017)
https://piazza.com/cmu/spring2017/10601
OUT: April 17, 2017
DUE: April 24, 2017 11:59 pm
TAs: Dylan Fitzpatrick, Brynn Edmunds, Edward Wang

START HERE: Instructions

- **Collaboration Policy**: Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Jane explained to me what is asked in Question 3.4"). Second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only. See the collaboration policy on the website for more information: http://www.cs.cmu.edu/~mgormley/courses/10601-s17/about.html
- **Late Submission Policy:** See the late submission policy here: http://www.cs.cmu.edu/~mgormley/courses/10601-s17/about.html
- **Submitting your work:** You will use Gradescope to submit answers to all questions for this homework.
  - **Gradescope:** For all problems on this homework we will be using Gradescope. You can access the site here: https://gradescope.com/. Submissions can be handwritten, but should be labeled and clearly legible. If your writing is not legible, you will not be awarded marks. Alternatively, submissions can be written in LaTeX. Upon submission, label each question using the template provided. Regrade requests can be made, however this gives the TA the option to regrade your entire paper, meaning if additional mistakes are found then points will be deducted.

Problem 1: Constructing Bayesian Networks [20 pts]

This section includes descriptions of events which can be represented by Bayesian networks.

1. [8 pts] Assume that Pittsburgh is undergoing construction, and you want to determine whether this will effect if you are late for work. To get to work, you need to cross the Highland Park Bridge. When construction occurs, it is more likely that the Highland Park Bridge will be closed. When the bridge is closed, it is more likely that you will be late for work, because you have to find an alternate route. Construction also often results in a traffic jam, which likewise increases your chance of being late for work. You also want to consider rush hour, which increases the chance of traffic jams.

   (a) Draw the Bayesian network that represents the relationship between the events described above. Use the following notation for your variables: construction (C), rush hour (R), traffic jam (J), bridge closed (B), and late for work (L).

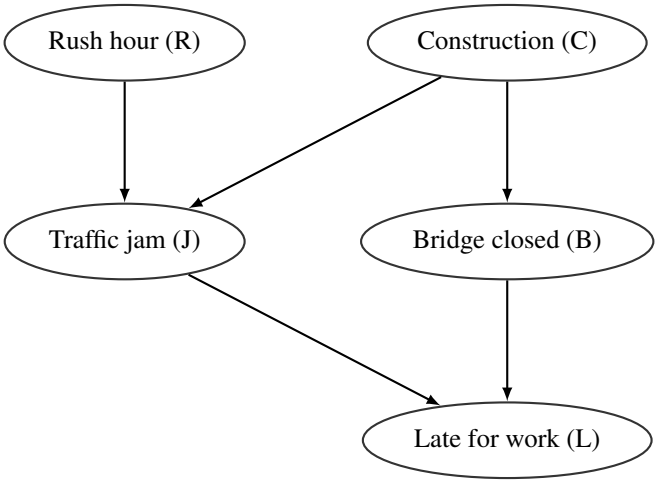


Figure 1: Bayesian network that represents events related to Pittsburgh's construction.

   (b) Assume each event represented in your Bayesian network is binary, e.g. it occurs or it doesn't. State the minimum number of parameters needed to fully specify your model for the described situation.

   As Figure 1 shown, the joint distribution of this model could be represented as:

$$P(C, R, J, B, L) = P(R) * P(C) * P(B|C) * P(J|R, C) * P(L|J, B)$$

   And since each events is binary, so for:
   R: Need 1 parameter, since the possibility add to 1.
   C: Need 1 parameter, since the possibility add to 1.
   B: B conditions on C and C has 2 possible values. For each value, B need 1 parameter to represent its distribution since the possibility add to 1. So totally for B: Need 2 parameters.
   J: J conditions on R and C, there are 4 possible combination of R and C. So totally for J: Need 4 parameters.
   L: L conditions on J and B, same with J: Need 4 parameters.

   So overall we need $(1 + 1 + 2 + 4 + 4) = 12$ parameters to fully specify the model.

2. [12 pts] Harry Potter is a student at Hogwarts school of Witchcraft and Wizardry. At Hogwarts, students are assigned to one of four groups, known as houses, and are given the opportunity to earn points for their assigned house.

   There is a high probability that Harry's Defence Against the Dark Arts teacher is evil, and a low probability that he is not evil. Harry is more likely to sneak out of bed when his professor is evil, as he wants to try and discover his professor's evil scheme. If Harry sneaks out of bed, the probability that he is sent to the infirmary as a result of a near fatal injury increases. Being stuck in the infirmary means Harry is less likely to be in the library studying. Furthermore, sneaking out of his bed after hours will decrease the likelihood of Harry's house earning points. If Harry's professor is not evil, Harry is more likely to spend his time in the library studying. Harry is also more likely to spend his time studying in the library if his friend, Hermione, lectures him to do so. As Harry studies more, he is more likely to earn points for his house.

   (a) Draw the Bayesian network that represents the relationship between the events described above. Use the following notation for your variables: evil professor (P), Hermione lectures Harry (H), Harry sneaks out (S), Harry studies in the library (L), Harry sent to infirmary (I), and Harry earns house points (E).

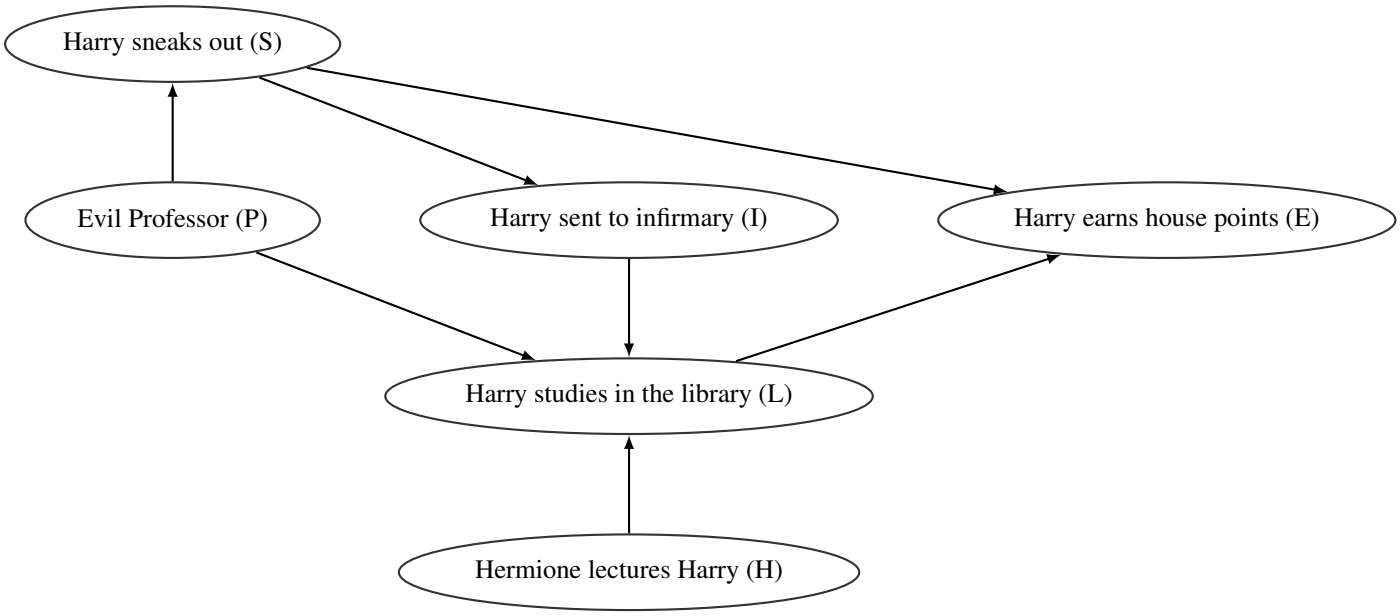


Figure 2: Bayesian network that represents Harry Potter model.

(b) Assume each of event represented in your model is binary, e.g. it occurs or it doesn't. State the minimum number of parameters needed to fully specify your model for the described situation.

The model was shown in Figure 2, the joint distribution of this model could be represented as:

$$P(P, S, I, L, H, E) = P(P) * P(H) * P(S|P) * P(I|S) * P(E|S, L) * P(L|I, P, H)$$

Since each events is binary, same with the former example, total number of parameters needed are (with respect to the sequence of the euqation above):

$$1 + 1 + 2 + 2 + 4 + 8 = 18$$

(c) Suppose that instead of the professor being evil or not evil (binary event), the professor could be good, evil, or neutral. Would this change the minimum number of parameters needed to represent our model? If yes, what is the new number of parameters? If no, why not?

If the possibilities of P turns to 3, 2 parameters will be needed to represent this event instead of 1. As a consequence, all the events conditions on P will have to tune their number of parameters as well.
For P: Number of parameters: $1 \rightarrow 2$.
For S: Number of parameters: $2 \rightarrow 3$.
For L: Number of parameters: $8 \rightarrow 12$.
And others are unchanged. So the total number is :24.

Problem 2: Bayesian Networks for Inference [50 pts]

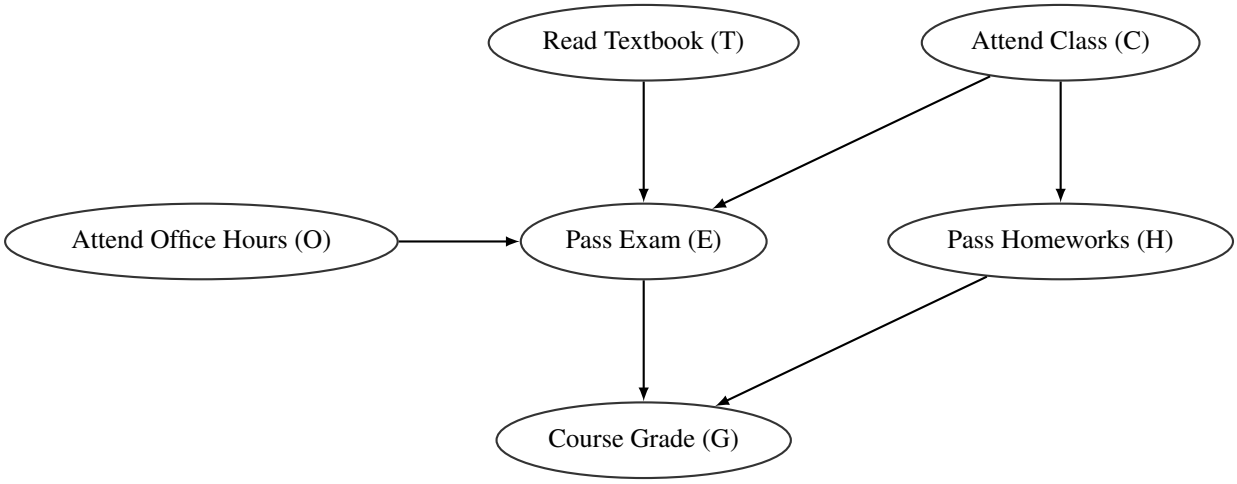


Figure 3: Bayesian network that represents the joint distribution over the variables used to predict a student's 10-601 grade.

As a favor to next semester's 10-601 students, you have decided to construct a model that can predict a student's course grade ahead of time based on the student's study habits and level of participation in the class. For this task, you will consider a Bayesian network that encodes the dependencies between several binary random variables which might affect a student's course grade (G). These binary variables are Read Textbook (T), Attend Class (C), Pass Exam (E), Pass Homeworks (I), and Attend Office Hours (O). The course grade G is a categorical variable with 5 possible values corresponding to letter grades A through F. The joint distribution over these variables is represented by the Bayesian network shown in Figure 3.

2.1 Conditional Independence and D-separation

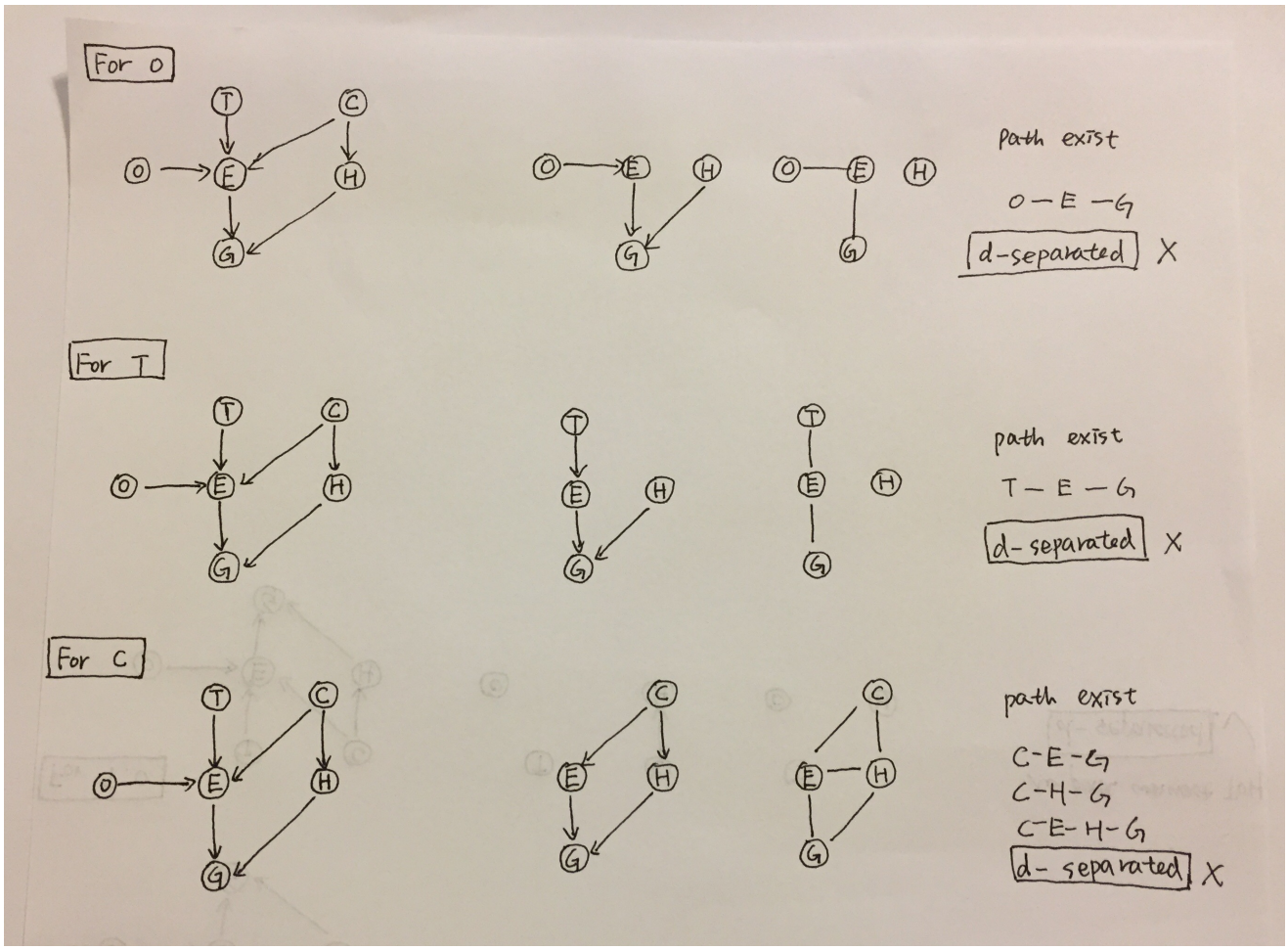
1. [10 pts] Using the graphical model depicted in Figure 3, state whether the following independence statements are true or false.

   (a) $T \perp C$ ——————————True
   (b) $T \perp G$ ——————————False
   (c) $O \perp C$ ——————————True
   (d) $O \perp H$ ——————————True
   (e) $T \perp G \mid E$ ——————————False
   (f) $O \perp C \mid G$ ——————————False
   (g) $E \perp H \mid C$ ——————————True
   (h) $C \perp G \mid O, H$ ——————————False
   (i) $C \perp G \mid O, E, H$ ——————————-True
   (j) $T \perp C \perp O \perp G \mid E, H$ ————-False

2. [4 pts] For the following questions provide your answer and a brief justification.

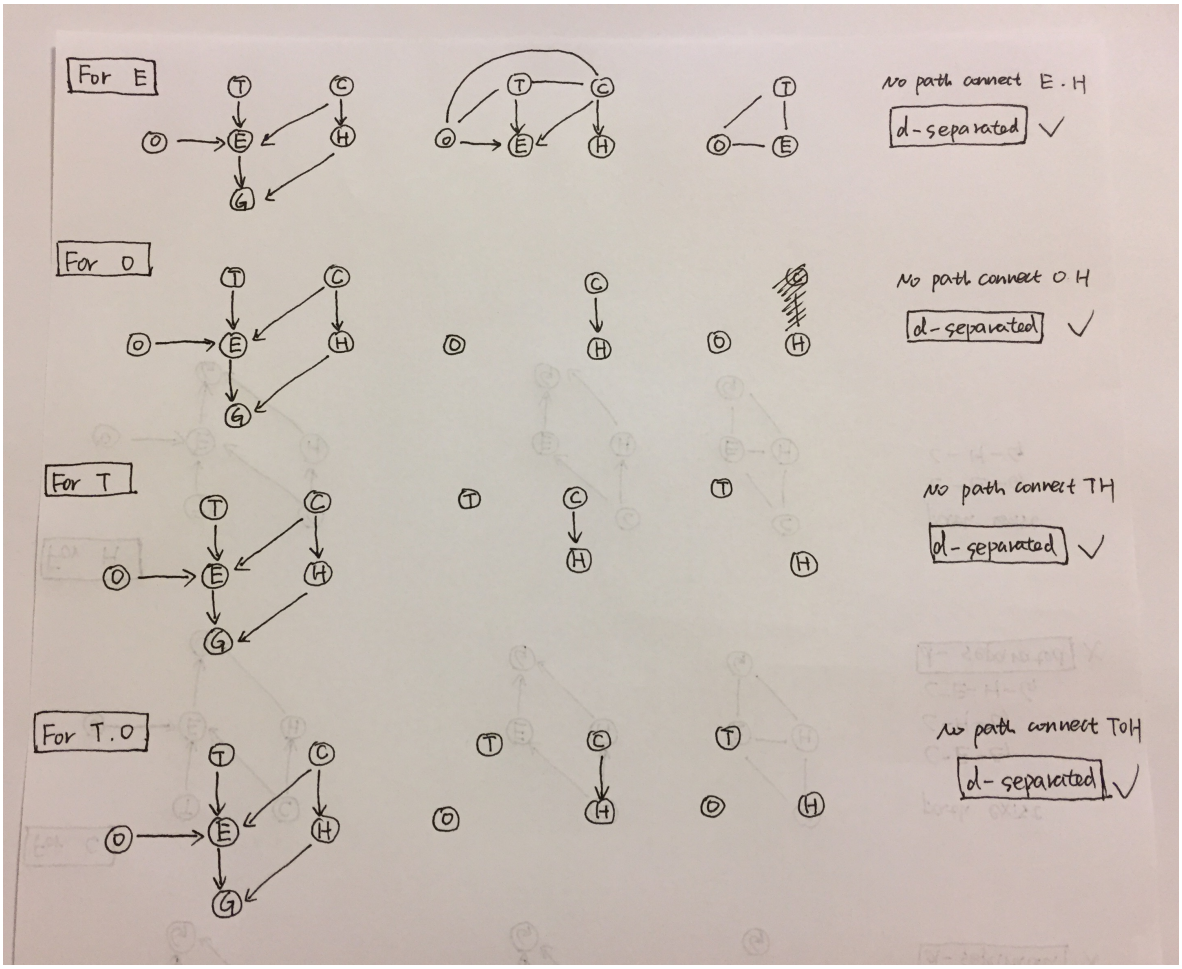
   (a) Which variables are d-separated from G with an empty evidence set?

   For empty evidence set, no one is d-separated from G. Like shown below, from O, T, E, there exists a path connect them to node G and H and E are straight connected to node G. So given an empty evidence set, no one is d-separated from G.



(b) Which variables are d-separated from H given C?

Like shown below, **T, E, O and (T,O)are d-separated from H given C.** Since G is straight connected to H, G is not d-separated from H given C.



3. [4 pts] Write down the factorized form of the joint distribution over all the variables represented in Figure 3, $P(T, C, O, E, H, G)$.

$P(T, C, O, E, H, G) = P(T) * P(C) * P(O) * P(H|C) * P(G|E, H) * P(E|T, C, O)$

2.2 Inference

The 10-601 TAs are interested in using your Bayesian network to perform inference regarding 10-601 lecture attendance. Specifically, the TAs would like to know the probability that a student attended class given that the student received an overall grade of a B and was observed attending office hours (i.e., $P(C = C^1|O = O^1, G = G^B)$). After collecting data on previous 10-601 students, you learn the conditional probability tables (CPTs) shown below.

$P(T^0)$	0.45
$P(T^1)$	0.55

$P(C^0)$	0.2
$P(C^1)$	0.8

$P(O^0)$	0.7
$P(O^1)$	0.3

	$T^0C^0O^0$	$T^0C^0O^1$	$T^0C^1O^0$	$T^0C^1O^1$	$T^1C^0O^0$	$T^1C^0O^1$	$T^1C^1O^0$	$T^1C^1O^1$
$P(E^0)$	0.85	0.8	0.4	0.3	0.6	0.5	0.15	0.1
$P(E^1)$	0.15	0.2	0.6	0.7	0.4	0.5	0.85	0.9

	C^0	C^1
$P(H^0)$	0.75	0.2
$P(H^1)$	0.25	0.8

	E^0H^0	E^0H^1	E^1H^0	E^1H^1
$P(G^A)$	0.05	0.1	0.1	0.6
$P(G^B)$	0.05	0.3	0.3	0.2
$P(G^C)$	0.1	0.3	0.3	0.1
$P(G^D)$	0.2	0.2	0.2	0.05
$P(G^F)$	0.6	0.1	0.1	0.05

1. [8 pts] Use exact inference to estimate the probability of attending class given that the course grade was a B and office hours were attended. Report $P(C = C^1|O = O^1, G = G^B)$, and show your work.

$$P(C = C^1|O = O^1, G = G^B) = \frac{P(C^1, O^1, G^B)}{P(O^1, G^B)}$$

$$P(C^1, O^1, G^B) = \sum^{T,E,H} P(T, C^1, O^1, E, H, G^B)$$

$$= \sum^{T,E,H} P(T)P(C^1)P(O^1)P(H|C^1)P(G^B|E, H)P(E|T, C^1, O^1)$$

$$= P(C^1)P(O^1) \sum^{T,E,H} P(T)P(H|C^1)P(G^B|E, H)P(E|T, C^1, O^1)$$

$$= P(C^1)P(O^1) * [P(T^0)P(H^0|C^1)P(G^B|E^0, H^0)P(E^0|T^0, C^1, O^1)+$$
$$P(T^0)P(H^1|C^1)P(G^B|E^0, H^1)P(E^0|T^0, C^1, O^1)+P(T^0)P(H^0|C^1)P(G^B|E^1, H^0)P(E^1|T^0, C^1, O^1)$$
$$+P(T^0)P(H^1|C^1)P(G^B|E^1, H^1)P(E^1|T^0, C^1, O^1)+P(T^1)P(H^0|C^1)P(G^B|E^0, H^0)P(E^0|T^1, C^1, O^1)$$
$$+P(T^1)P(H^1|C^1)P(G^B|E^0, H^1)P(E^0|T^1, C^1, O^1)+P(T^1)P(H^0|C^1)P(G^B|E^1, H^0)P(E^1|T^1, C^1, O^1)+$$
$$P(T^1)P(H^1|C^1)P(G^B|E^1, H^1)P(E^1|T^1, C^1, O^1)]$$
$$= 0.8 * 03 * (0.45 * (0.2 * 0.05 * 0.3 + 0.8 * 0.3 * 0.3 + 0.2 * 0.3 * 0.7 + 0.8 * 0.2 * 0.7)+$$
$$0.55 * (0.2 * 0.05 * 0.1 + 0.8 * 0.3 * 0.1 + 0.2 * 0.3 * 0.9 + 0.8 * 0.2 * 0.9))$$
$$= 0.8 * 0.3 * (0.45 * 0.229 + 0.55 * 0.223) = 0.054168$$

$$P(O^1, G^B) = \sum^{T,C,E,H} P(T)P(C)P(O^1)P(H|C)P(G^B|E, H)P(E|T, C, O^1)$$

$$= P(O^1) \sum^{T,C,E,H} P(T)P(C)P(H|C)P(G^B|E, H)P(E|T, C, O^1)$$

$$= 0.054168 + P(C^0, O^1, G^B) = 0.054168 + 0.01032 = 0.064485$$

$$P(C = C^1|O = O^1, G = G^B) = \frac{0.054168}{0.064485} \approx 0.84$$

2. [10 pts] We will now answer the same question using an approach for approximate inference which we will call "brute force sampling." Following the Bayesian network in Figure 3 and using the CPTs given above, generate k samples for each variable. You should reject sets of samples that do not match the observed values for O and G. For example, if on one sampling iteration you draw $\{T = T^1, C = C^0, O = O^0, E = E^1, H = H^1, G = G^B\}$, you should throw away this entire set of samples because O does not match the observed value of O^1. Using only the valid samples, estimate the probability that a student attended class given that office hours were attended and the course grade was a B.

   You should generate $k = 10000$ samples (including rejections), and estimate $P(C = C^1|O = O^1, G = G^B)$ at each sampling iteration. Plot your estimate for $P(C = C^1|O = O^1, G = G^B)$ against the number of total (accepted and rejected) samples. Your estimate for $P(C = C^1|O = O^1, G = G^B)$ should be zero until you have obtained at least one valid set of samples. You do not need to submit any code for this question.

   I initialized the conditional possibility to be 0 until there come a new value to replace it. To generate the sample, I generated T, C and O based on their own distribution. E was generated based on the value of T, C and O generated before and the conditional distribution of E, and same with H and G. Values were generated by a random value bounded by the possibilities in each distribution. Below are two figures generated by my algorithm. As we can see the final value of joint possibility is about 0.84 which is nearly same as the output of inference.

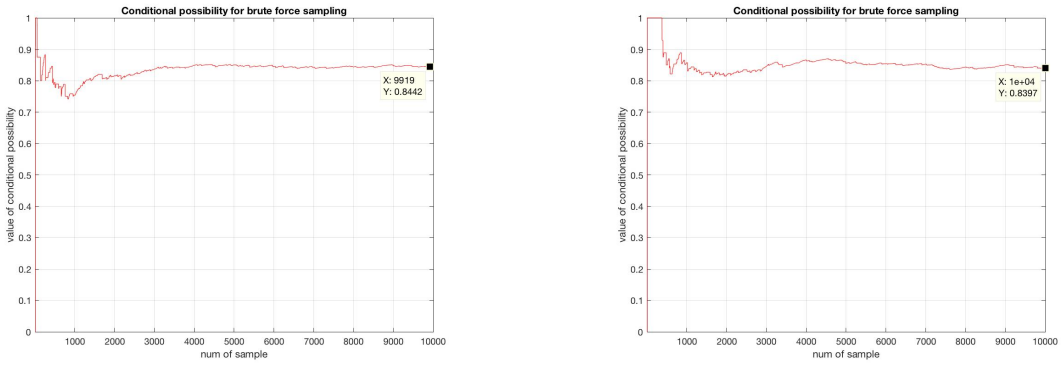


Figure 6: Conditional possibility for Brute force sampling

3. [10 pts] Now we will consider an alternate approach to estimating our joint distribution known as Gibbs sampling. The Gibbs sampling procedure approximates the joint distribution by sampling from conditional distributions for each variable. The sampling proceeds as follows:

   (a) Begin with any initial assignment for all unobserved variables, T, C, E, and H

   (b) Draw the next set of samples, which we will call T_{i+1}, C_{i+1}, E_{i+1}, and H_{i+1}. Sample each of the variables *in order*, conditioning on the most recent sample from other variables. Formally,

      i. Sample $T_{i+1}|C_i, E_i, H_i, O = O^1, G = G^B$
      ii. Sample $C_{i+1}|T_{i+1}, E_i, H_i, O = O^1, G = G^B$
      iii. Sample $E_{i+1}|T_{i+1}, C_{i+1}, H_i, O = O^1, G = G^B$
      iv. Sample $H_{i+1}|T_{i+1}, C_{i+1}, E_{i+1}, O = O^1, G = G^B$

   (c) Repeat step (b) k times to produce k samples for each variable.

   You will use the CPTs given above to draw samples from the full conditional distributions. You should generate $k = 10000$ samples for each variable, and plot your estimate for $P(C = C^1|O = O^1, G = G^B)$ against the number of samples on the same plot as brute force sampling. Submit your plot. You do not need to submit any code for this question.

   In my algorithm I initialized $T_1, C_1, E_1, H_1 = 0$ and sample T,C,E,H in a iteration of 10000. Since full conditions only need to condition on the Markov blanket, so the possibility of $T_{i+1} = 1$ is

$$P(T^1_{i+1}) = \frac{P(T^1_{i+1})P(E|O^1, T_{i+1^1,C_i})}{P(T^1_{i+1})P(E|O^1, T_{i+1^1,C_i}) + P(T^0_{i+1})P(E|O^1, T_{i+1^0,C_i})}$$

$$P(C^1_{i+1}) = \frac{P(C^1_{i+1})P(H|C^1_{i+1})P(E|O^1, T_{i+1}, C^1_{i+1})}{P(C^1_{i+1})P(H|C^1_{i+1})P(E|O^1, T_{i+1}, C^1_{i+1}) + P(C^0_{i+1})P(H|C^0_{i+1})P(E|O^1, T_{i+1}, C^0_{i+1})}$$

$$P(E^1_{i+1}) = \frac{P(E^1_{i+1}|o^1, T_{i+1}, C_{i+1})P(G^B|E^1_{i+1}, H_i)}{P(E^1_{i+1}|o^1, T_{i+1}, C_{i+1})P(G^B|E^1_{i+1}, H_i) + P(E^0_{i+1}|o^1, T_{i+1}, C_{i+1})P(G^B|E^0_{i+1}, H_i)}$$

$$P(H^1_{i+1}) = \frac{P(H^1_{i+1}|C_{i+1})P(G^B|E_{i+1}, H^1_{i+1})}{P(H^1_{i+1}|C_{i+1})P(G^B|E_{i+1}, H^1_{i+1}) + P(H^0_{i+1}|C_{i+1})P(G^B|E_{i+1}, H^0_{i+1})}$$

   And then generate G based on E and H, and generate O based on O's distribution. The image below is the output. Below are some images generated from my algorithm. As we can see the final joint possibility of brute force sampling is nearly same with the output of inference which is about 0.84. And the output of Gibbs sampling is about 0.8-0.82 which is approximate to the joint possibility.

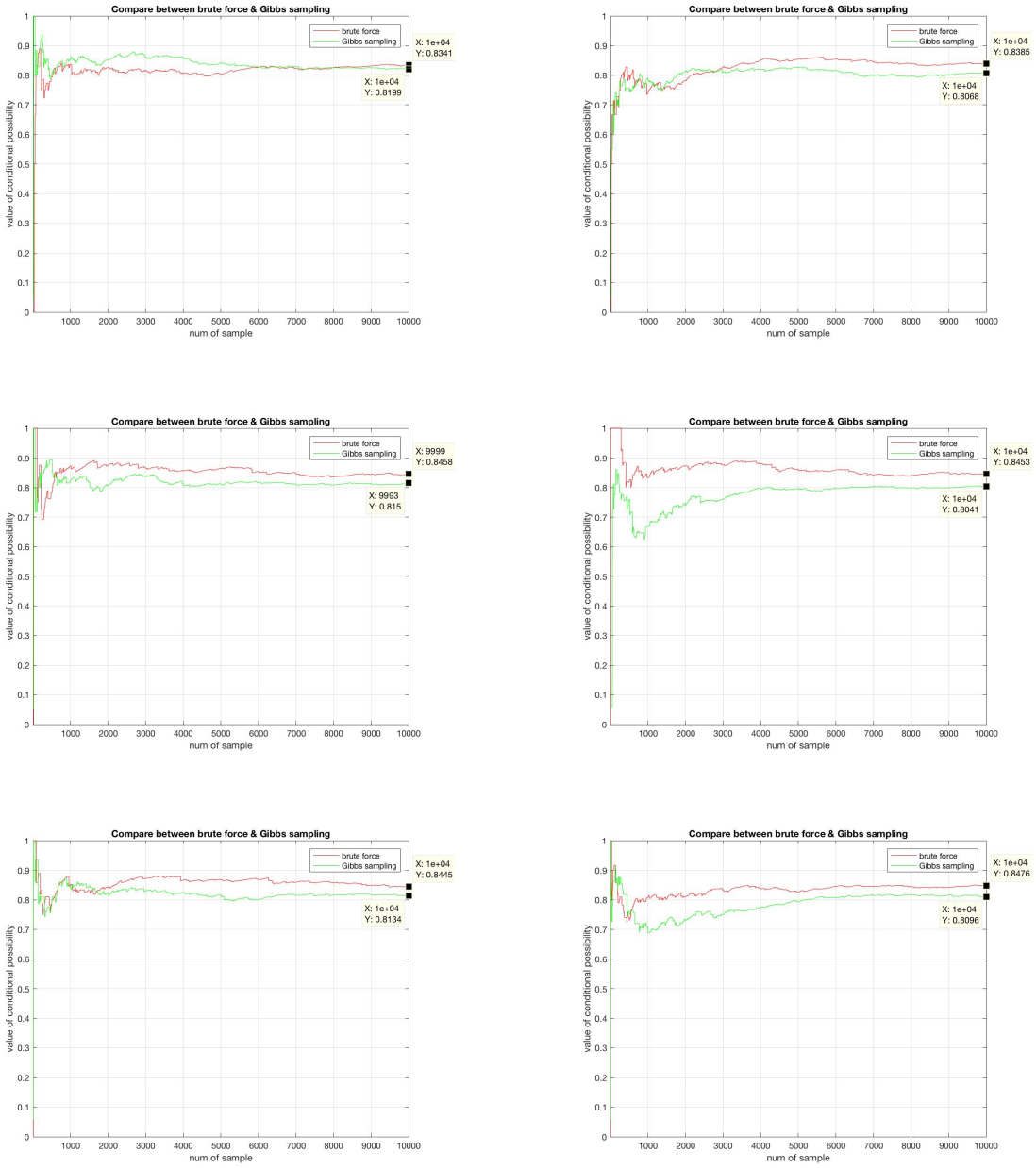


Figure 9: Conditional possibility for Brute force sampling and Gibbs sampling

4. [4 pts] Why might we prefer Gibbs sampling over brute force sampling? Provide a 2-3 sentence justification.

   In real practice with large scale sample, compute the joint distribution in the brute force way is NP hard. Since the joint distribution is what we want and the conditional distribution is easier to have, we could make use of it and draw sample from the conditional distribution, it's easier to approximate the joint distribution by this way with less computation.

Problem 3: Hidden Markov Models [30 pts]

A HMM defines a joint probability distribution over sequences of state-observation pairs, and can be represented with a graphical model as shown in Figure 10.

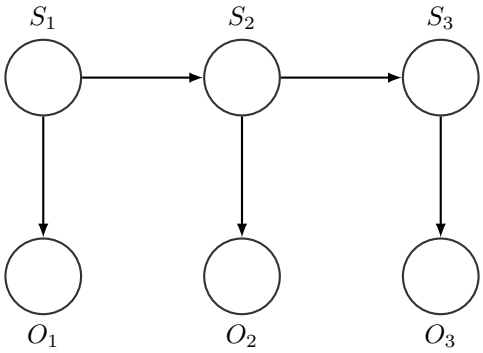


Figure 10: An HMM with three state-observation pairs.

3.1 Structure

1. [2 pts] List two types of conditional independence statements from the HMM graphical model structure, using S_t and O_t to represent the t^{th} state-observation pair.

   Common parent: $S_{t+1} \perp O_t | S_t$
   Cascade: $S_{t-1} \perp S_{t+1} | S_t$

2. [2 pts] Assuming each state S can take k different values, and observation O can take on p different values, what is the minimum number of parameters needed to fully define the HMM shown in Figure 10? Make sure to consider that some parameters must add to 1.
   Joint distribution:

$$P(S_1)P(O_1|S_1)P(S_2|S_1)P(O_2|S_2)P(S_3|S_2)P(O_3|S_3)$$

   With sequence of this equation, total number of parameters are:

$$\begin{aligned}
&(k-1) + k(p-1) + k(k-1) + k(p-1) + k(k-1) + k(p-1) \\
&= (k-1) + 3k(p-1) + 2k(k-1) \\
&= (2k+1)(k-1) + 3k(p-1) \\
&= 2k^2 - 4K + 3pk - 1
\end{aligned}$$

3.2 Inference

Assume that we have the HMM shown in Figure 10, our hidden state variables can take on one of two states ($S \in \{A, B\}$), and at each time t we can observe one of two outcomes ($O \in \{C, D\}$). This model is completely specified by the initial state probabilities, transition probabilities, and emission probabilities shown in the tables below.

State	$P(S_1)$
A	0.8
B	0.2

(a) Initial probabilities

S_t	S_{t+1}	$P(S_{t+1}\|S_t)$
A	A	0.4
A	B	0.6
B	A	0.5
B	B	0.5

(b) Transition probabilities

S	O	$P(O\|S)$
A	C	0.75
A	D	0.25
B	C	0.2
B	D	0.8

(c) Emission probabilities

1. [8 pts] Assume we observe the sequence $O_1 = D, O_2 = D, O_3 = C$. Apply the forward algorithm to compute the probability of observing this sequence. Show your work and report $P(O_1 = D, O_2 = D, O_3 = C)$, as well as the values of unscaled α parameters α_1^A, α_1^B, α_2^A, α_2^B, α_3^A, and α_3^B.

   $\alpha_1^A = P(S_1^A) * P(O_1^D|S_1^A) = 0.8 * 0.25 = 0.2$

   $\alpha_1^B = P(S_1^B) * P(O_1^D|S_1^B)0.2 * 0.8 = 0.16$

   $\alpha_2^A = P(O_2^D|S_2^A) * [\alpha_1^A * P(S_2^A|S_1^A) + \alpha_1^B * P(S_2^A|S_1^B)] = 0.25 * (0.2 * 0.4 + 0.16 * 0.5) = 0.04$

   $\alpha_2^B = P(O_2^D|S_2^B) * [\alpha_1^A * P(S_2^B|S_1^A) + \alpha_1^B * P(S_2^B|S_1^B)] = 0.8 * (0.2 * 0.6 + 0.16 * 0.5) = 0.16$

   $\alpha_3^A = P(O_3^C|S_3^A) * [\alpha_2^A * P(S_3^A|S_2^A) + \alpha_2^B * P(S_3^A|S_2^B)] = 0.75 * (0.04 * 0.4 + 0.16 * 0.5) = 0.072$

   $\alpha_3^B = P(O_3^C|S_3^B) * [\alpha_2^A * P(S_3^B|S_2^A) + \alpha_2^B * P(S_3^B|S_2^B)] = 0.2 * (0.04 * 0.6 + 0.16 * 0.5) = 0.0208$

   Since $\alpha_3^A = P(O_1^D, O_2^D, O_3^C, S_3^A)$ and $\alpha_3^B = P(O_1^D, O_2^D, O_3^C, S_3^B)$

   So $P(O_1^D, O_2^D, O_3^C) = \alpha_3^A + \alpha_3^B = 0.0928$

2. [8 pts] Still assuming we observe the sequence $O_1 = D, O_2 = D, O_3 = C$, apply the backward algorithm to compute the probability of observing this sequence. Show your work and report $P(O_1 = D, O_2 = D, O_3 = C)$, as well as the values of unscaled β parameters β_3^A, β_3^B, β_2^A, β_2^B, β_1^A, and β_1^B.

Assume $\beta_4^{END} = 1$

$\beta_3^A = P(O_4|S_4^A) * P(S_4^A|S_3^A) + P(O_4|S_4^B) * P(S_4^B|S_3^A) = 1$

$\beta_3^B = P(O_4|S_4^A) * P(S_4^A|S_3^B) + P(O_4|S_4^B) * P(S_4^B|S_3^B) = 1$

$\beta_2^A = P(O_3^C|S_3^A) * \beta_3^A * P(S_3^A|S_2^A) + P(O_3^C|S_3^B) * \beta_3^B * P(S_3^B|S_2^A) = 0.75 * 1 * 0.4 + 0.2 * 1 * 0.6 = 0.42$

$\beta_2^B = P(O_3^C|S_3^A) * \beta_3^A * P(S_3^A|S_2^B) + P(O_3^C|S_3^B) * \beta_3^B * P(S_3^B|S_2^B) = 0.75 * 1 * 0.5 + 0.2 * 1 * 0.5 = 0.475$

$\beta_1^A = P(O_2^D|S_2^A) * \beta_2^A * P(S_2^A|S_1^A) + P(O_2^D|S_2^B) * \beta_2^B * P(S_2^B|S_1^A)$

$= 0.25 * 0.42 * 0.4 + 0.8 * 0.475 * 0.6 = 0.27$

$\beta_1^B = P(O_2^D|S_2^A) * \beta_2^A * P(S_2^A|S_1^B) + P(O_2^D|S_2^B) * \beta_2^B * P(S_2^B|S_1^B)$

$= 0.25 * 0.42 * 0.5 + 0.8 * 0.475 * 0.5 = 0.2425$

Since $\beta_1^A = P(O_2^D.O_3^C|S_1^A)$ and $\beta_1^B = P(O_2^D.O_3^C|S_1^B)$

So $P(O_1^D, O_2^D, O_3^C) = \beta_1^A * P(S_1^A) * P(O_1^D|S_1^A) + \beta_1^B * P(S_1^B) * P(O_1^D|S_1^B)$

$= 0.25 * 0.27 * 0.8 + 0.8 * 0.2425 * 0.2 = 0.0928$

3. [8 pts] Still assuming we observe the sequence $O_1 = D, O_2 = D, O_3 = C$, complete the forward-backward algorithm to compute marginal probabilities for the hidden state variables. For each of the state variables S_1, S_2, and S_3, report the value with highest marginal probability.

$$P(S_1^A|O_1^D, O_2^D, O_3^C) = \frac{\alpha_1^A * \beta_1^A}{P(O_1^D, O_2^D, O_3^C)} = 0.2 * 0.27/0.0928 = 0.5819$$

$$P(S_1^B|O_1^D, O_2^D, O_3^C) = \frac{\alpha_1^B * \beta_1^B}{P(O_1^D, O_2^D, O_3^C)} = 0.16 * 0.2425/0.0928 = 0.4181$$

$$P(S_2^A|O_1^D, O_2^D, O_3^C) = \frac{\alpha_2^A * \beta_2^A}{P(O_1^D, O_2^D, O_3^C)} = 0.04 * 0.42/0.0928 = 0.181$$

$$P(S_2^B|O_1^D, O_2^D, O_3^C) = \frac{\alpha_2^B * \beta_2^B}{P(O_1^D, O_2^D, O_3^C)} = 0.16 * 0.475/0.0928 = 0.819$$

$$P(S_3^A|O_1^D, O_2^D, O_3^C) = \frac{\alpha_3^A * \beta_3^A}{P(O_1^D, O_2^D, O_3^C)} = 0.072 * 1/0.0928 = 0.7759$$

$$P(S_3^B|O_1^D, O_2^D, O_3^C) = \frac{\alpha_3^B * \beta_3^B}{P(O_1^D, O_2^D, O_3^C)} = 0.0208 * 1/0.0928 = 0.2241$$

So the value with highest marginal probability is $S_1^A = 0.5891, S_2^B = 0.819, S_3^A = 0.7759$.

4. [2 pts] Another common inference task is to find the most probable assignment of states given a sequence of observations. For any HMM, is the most probable assignment of states equal to the sequence of values with highest marginal probability found using the forward-backward algorithm? Provide a 1-2 sentence justification.

In HMM and Viterbi algorithm we both care about the observation possibility and the transition possibility. To find the most probable assignments, use Viterbi algorithm

$$\omega_1^A = \max_{A,B} P(O_1^D|S_1^A)P(S_1^A) = 0.25 * 0.8 = 0.2$$

$$\omega_1^B = \max_{A,B} P(O_1^D|S_1^B)P(S_1^B) = 0.8 * 0.2 = 0.16$$

$$\omega_2^A = \max_{A,B}[P(O_2^D|S_2^A)\omega_1^A P(S_2^A|S_1^A); P(O_2^D|S_2^A)\omega_1^B P(S_2^A|S_1^B)] = \max_{A,B}[0.02; 0.02] = 0.02$$

$$\omega_2^B = \max_{A,B}[P(O_2^D|S_2^B)\omega_1^A P(S_2^B|S_1^A); P(O_2^D|S_2^B)\omega_1^B P(S_2^B|S_1^B)] = \max_{A,B}[0.096; 0.064] = 0.096$$

$$\omega_3^A = \max_{A,B}[P(O_3^C|S_3^A)\omega_2^A P(S_3^A|S_2^A); P(O_3^C|S_3^A)\omega_2^B P(S_3^A|S_2^B)] = \max_{A,B}[0.006; 0.036] = 0.036$$

$$\omega_3^B = \max_{A,B}[P(O_3^C|S_3^B)\omega_2^A P(S_3^B|S_2^A); P(O_3^C|S_3^B)\omega_2^B P(S_3^B|S_2^B)] = \max_{A,B}[0.0024; 0.0096] = 0.0096$$

In Viterbi algorithm we pick the state with higher ω in a time and use it to compute the next ω. From the output above we can see the most probable assignment is $S_1 = A, S_2 = B, S_3 = A$ which is the same as sequence of values with the highest marginal probability yield by forward-backward algorithm.

Collaboration Questions

Please complete the following questions and submit your answers in Gradescope:

Collaboration

- Did you receive any help whatsoever from anyone in solving this assignment?
  No.
- If you answered *yes*, give full details: _____
  (e.g. *Jane explained to me what is asked in Question 3.4*)
- Did you give any help whatsoever to anyone in solving this assignment?
  No.
- If you answered *yes*, give full details: _____
  (e.g. *I pointed Joe to section 2.3 to help him with Question 2*).

Time Spent

- How many hours did this assignment take: 30h