CSE 150A/250A

AI: Probabilistic Methods

Spring 2025 - Practice Final Exam

Instructor: Trevor Bonjour

34 Points

Thursday, June 5th, 2025

Instructions

- **Do not turn to any other page of this exam until instructed to begin.**
- Once instructed to begin, write your name (in all caps) and PID on all odd pages.
- You may use both sides of a one 8.5"x11" (or A4-sized) sheet of handwritten notes during the exam. Write your name and PID on the sheet of notes. No other notes or resources are allowed.
- There are a total of **5** questions. For any multiple choice or true/false question, please **fill in the bubble completely** to mark your answer.
- You have **75 minutes** to complete the exam.
- The pages at the end of the exam may be used as a scratch paper. No other scratch paper is allowed.
- In the name slot below, write your name in **ALL CAPITAL LETTERS** as it appears on Gradescope.

Declaration:

I understand and will abide by the above instructions.

Signature: ____

Name (ALL CAPS): ____

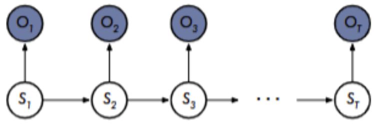
PID: ____

1 T/F [10 points]

For each statement, select either true or false. Please fill in the bubble completely to mark your answer.

1. Value Iteration algorithm converges in a finite number of steps. (T) (F)
2. A Markov decision process (MDP) is defined by a state space, action space, transition probabilities, and reward function. (T) (F)
3. In most reinforcement learning problems, we should maximize the best-case return over the expected return. (T) (F)
4. The CPTs of a belief network grows linearly when applying node clustering. (T) (F)
5. All belief networks are polytrees. (T) (F)
6. The run time of the EM algorithm on Hidden Markov Models is quadratic with respect to the number of states. (T) (F)
7. $P(O_{t+1} = k|S_t = i)$ is a parameter of hidden Markov models (HMMs) and can be represented by the emission matrix. (T) (F)
8. In an n-gram model, we should always try to maximize n to model long-range dependencies between words. (T) (F)
9. Markov Chain Monte Carlo (MCMC) is generally faster than Likelihood Weighting and Rejection Sampling. (T) (F)
10. The objective function optimized by the EM algorithm can also be optimized by a gradient descent algorithm which will find the global optimal solution, whereas EM finds its solution more quickly but may return only a locally optimal solution. (T) (F)

2 Viterbi Algorithm (3 pts)



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

$$\boxed{t = 1} \quad \ell^*_{i1} = \log \pi_i + \log b_i(O_1)$$
$$\boxed{t > 1} \quad \ell^*_{j,t+1} = \max_i \left[\ell^*_{it} + \log a_{ij}\right] + \log b_j(O_{t+1})$$

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 options.

A. $P(S_{1:T})$

B. $P(S_T \mid O_T)$

C. $P(S_{1:T} \mid O_{1:T})$

D. $P(S_{1:T}, O_{1:T})$

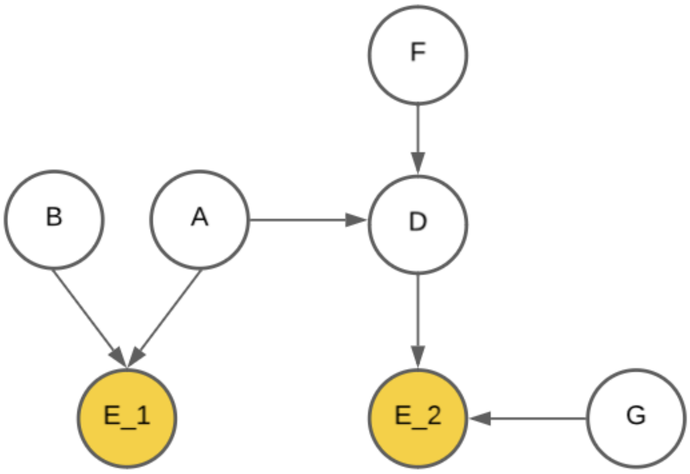
E. $P(S_1)P(O_1 \mid S_1) \prod_{t=2}^{T} P(O_t \mid S_t)P(S_t \mid S_{t-1})$

F. $P(S_1) \prod_{t=2}^{T} P(S_t \mid S_{t-1})$

G. None of the above

3 D-Separation [7 points]

Answer the following questions referring to the Bayesian network shown below:



1. **(3 points)** Assume that the shaded nodes, E1 and E2, are observed evidence. Which of the following statements must be true?
   - Ⓐ Nodes A and F would be conditionally independent if D was also observed
   - Ⓑ $P(B = 1|E_1, E_2) \neq P(D = 1|E_1, E_2)$
   - Ⓒ $P(F|D, G) = P(F)$
   - Ⓓ Nodes B and G are conditionally independent

2. **(1 point)** Is the graph above a Polytree? If not, which 2 nodes could we merge to make the graph a Polytree?

3. **(3 points)** Assume that all variables only take on the values of 1 or 0. Further, for this question, assume that E_2 is **NOT** observed. We know that:

   (a) P(G=1) = $\frac{3}{5}$
   (b) P(F=1) = $\frac{1}{2}$
   (c) P(A=1) = $\frac{1}{2}$
   (d) $P(D = 0|A = 1, F = 1) = \frac{1}{3}$
   (e) $P(D = 1|A = 0, F = 1) = \frac{2}{3}$
   (f) $P(D = 0|A = 1, F = 0) = 0$
   (g) $P(D = 1|A = 0, F = 0) = 1$

   What is the probability that D = 0? You may leave your answer as a fraction.

4 Expectation Maximization [8 points]

1. **(1 point)** During the Expectation (E) step of the EM algorithm applied to a Bayesian network, the algorithm:
   Ⓐ Maximizes the expected complete-data log-likelihood with respect to the parameters.
   Ⓑ Updates the parameters to maximize the likelihood given the observed data.
   Ⓒ Estimates the expected values of hidden variables using current parameter estimates.
   Ⓓ Introduces regularization to prevent overfitting.

2. **(1 point)** Which of the following is a common criterion for convergence of the EM algorithm in Bayesian networks?
   Ⓐ The algorithm stops after a fixed number of iterations regardless of convergence.
   Ⓑ The change in parameter estimates between iterations falls below a predefined threshold.
   Ⓒ All missing data have been accurately predicted.
   Ⓓ The likelihood of the observed data decreases between iterations.

3. **(6 points)** Consider the belief network shown in Figure 1, with observed nodes B and C and hidden node A. For each of the following questions, please select **only one** answer. Fill in the bubble completely to mark your answer. The formula should be in its simplest form, expressing them in terms of the posterior probabilities $P(a \mid b, c)$, as well as the functions $I(b, b_t)$ and $I(c, c_t)$.

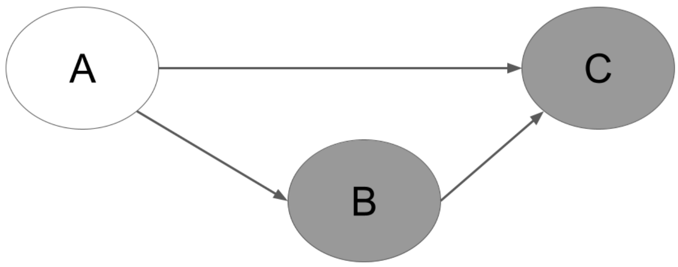


Figure 1: Belief network with observed nodes B and C, and hidden node A.

   (a) Updating $P_{\text{new}}(A = a)$
   Which of the following is the correct EM update equation for $P_{\text{new}}(A = a)$?

   Ⓐ
   $$P_{\text{new}}(A = a) = \frac{\sum_t P(a \mid b_t, c_t)}{T}$$

   Ⓑ
   $$P_{\text{new}}(A = a) = \frac{\sum_t P(a \mid b_t, c_t) \cdot I(b_t, b_t) \cdot I(c_t, c_t)}{T}$$

Ⓒ

$$P_{\text{new}}(A = a) = \frac{\sum_t I(a, a_t)}{T}$$

Ⓓ

$$P_{\text{new}}(A = a) = \frac{\sum_t P(a \mid b_t, c_t)}{\sum_t P(b_t, c_t)}$$

(b) Updating $P_{\text{new}}(B = b \mid A = a)$
Which of the following is the correct EM update equation for $P_{\text{new}}(B = b \mid A = a)$?

Ⓐ

$$P_{\text{new}}(B = b \mid A = a) = \frac{\sum_t P(a \mid b_t, c_t) \cdot I(b, b_t)}{\sum_t P(a \mid b_t, c_t)}$$

Ⓑ

$$P_{\text{new}}(B = b \mid A = a) = \frac{\sum_t I(a, a_t) \cdot I(b, b_t)}{\sum_t I(a, a_t)}$$

Ⓒ

$$P_{\text{new}}(B = b \mid A = a) = \frac{\sum_t P(b \mid a_t) \cdot I(b, b_t)}{\sum_t P(a \mid b_t, c_t)}$$

Ⓓ

$$P_{\text{new}}(B = b \mid A = a) = \frac{\sum_t P(a \mid b_t, c_t) \cdot I(b, b_t) \cdot I(c, c_t)}{\sum_t P(a \mid b_t, c_t)}$$

(c) Updating $P_{\text{new}}(C = c \mid A = a, B = b)$
Which of the following is the correct EM update equation for $P_{\text{new}}(C = c \mid A = a, B = b)$?

Ⓐ

$$P_{\text{new}}(C = c \mid A = a, B = b) = \frac{\sum_t P(a \mid b_t, c_t) \cdot I(b, b_t) \cdot I(c, c_t)}{\sum_t P(a \mid b_t, c_t) \cdot I(b, b_t)}$$

Ⓑ

$$P_{\text{new}}(C = c \mid A = a, B = b) = \frac{\sum_t I(a, a_t) \cdot I(b, b_t) \cdot I(c, c_t)}{\sum_t I(a, a_t) \cdot I(b, b_t)}$$

Ⓒ

$$P_{\text{new}}(C = c \mid A = a, B = b) = \frac{\sum_t P(a \mid b_t, c_t) \cdot I(c, c_t)}{\sum_t P(a \mid b_t, c_t)}$$

Ⓓ

$$P_{\text{new}}(C = c \mid A = a, B = b) = \frac{\sum_t P(a \mid b_t, c_t) \cdot I(b, b_t)}{\sum_t P(a \mid b_t, c_t)}$$

5 HMMs [6 points]

A casino alternates between using a fair dice ($S_t = 1$) and a loaded dice ($S_t = 2$). The probabilities which will be used for calculation in this question are defined as below:

- When using the **fair dice** ($S_t = 1$), all faces have an equal probability of showing up:

$$P(O_t = k|S_t = 1) = \frac{1}{6}, \quad k \in \{1, 2, 3, 4, 5, 6\}.$$

- When using the **loaded dice** ($S_t = 2$), the face 6 is more likely, we have:

$$P(O_t = k|S_t = 2) = \begin{cases} 0.5 & \text{if } k = 6, \\ 0.1 & \text{if } k \in \{1, 2, 3, 4, 5\}. \end{cases}$$

- The probability of switching between dice:
  - If the current dice is fair ($S_t = 1$):

$$P(S_{t+1} = 1|S_t = 1) = 0.8, \quad P(S_{t+1} = 2|S_t = 1) = 0.2.$$

  - If the current dice is loaded ($S_t = 2$):

$$P(S_{t+1} = 1|S_t = 2) = 0.1, \quad P(S_{t+1} = 2|S_t = 2) = 0.9.$$

- The initial probabilities:

$$P(S_1 = 1) = 0.9, \quad P(S_1 = 2) = 0.1.$$

Given the observation sequence $O = [1, 6, 5]$, please solve the following problems.

1. **(2 points)** Calculate the likelihood that the first roll of die observed is a 1.

2. **(3 points)** Recall that in the forward algorithm, we use a matrix to keep track of all intermediary values α_{it}, where $\alpha_{it} = P(o_1, o_2, \cdots, o_t, S_t = i)$. How will you calculate the value of α_{12} given the second observation is a 6? Assume you are also given α_{11} and α_{21}. You may leave your answer as a formula (using the terms given in the question above), no need for any calculations.

3. **(1 point)** Once you have computed the whole matrix of α_{it} for the forward algorithm, how will you extract the probability of the observation sequence O?