6.4110/16.420
Representation, Inference and Reasoning in AI

Quiz 1 Practice B

Solutions

Feburary 16, 2026

Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page.

You are permitted to use a single sheet of paper with notes on (both sides), and a calculator and a timer. You may not use your phone as a calculator. **Box all answers** for free response questions.

Name: _____

MIT email: _____

Question	Points	Score
1	30	
2	45	
3	25	
Total:	100	

6.4110/16.420
Quiz 1 Practice B, Feburary 16, 2026

1 Inverted Hidden Markov Model

1. In this section, we will apply a variant of Hidden Markov Models to a sequence labeling task. Specifically, we are interested in inferring the Part-Of-Speech (POS) labels for an input sentence. Consider the vocabulary containing three words: {the, cat, sat}, and four possible POS taggings {⟨S⟩, ARTICLE, NOUN, VERB} (⟨S⟩ is a special tag representing the begining of a sentence). For example, the groundtruth POS tags for the sentence *the cat sat* should be ARTICLE-NOUN-VERB.

   An "Inverted Hidden Markov Model" (IHMM) can be defined using the following directed graphical model.

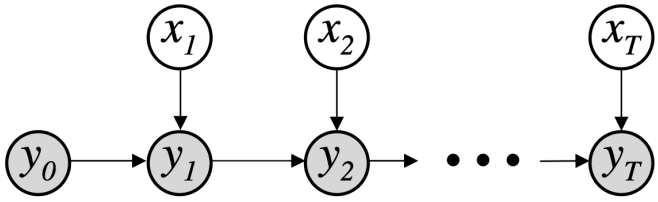


Figure 1: Inverted Hidden Markov Model (IHMM).

In the IHMM model, the y nodes correspond to the output label sequence, and the x nodes correspond to the input sentence.

(a) (4 points) Write down the probability distribution for the following conditional distribution, based on the directed graphical model defined above.

$$p(|) = p(Y_0, Y_1, Y_2, \cdots, Y_T | X_1, X_2, \cdots, X_T).$$

Solution:

$$p(Y_0, Y_1, Y_2, \cdots, Y_T | X_1, X_2, \cdots, X_T) = p(Y_0) \prod_{t=1}^{T} p(Y_t | X_t, Y_{t-1}).$$

For the rest of this section, we will assume $p(Y_0 = \langle S \rangle) = 1$ and focus on inferring $Y_1, Y_2, \cdots, Y_T$. We will develop a compact graphical representation to describe the conditional probability distributions $p(Y_t | X_t, Y_{t-1})$, showing as the weights on the directed edges. For example, for any t, $p(Y_t = \text{NOUN} | X_t = \text{cat}, Y_{t-1} = \text{ARTICLE}) = 0.9$.

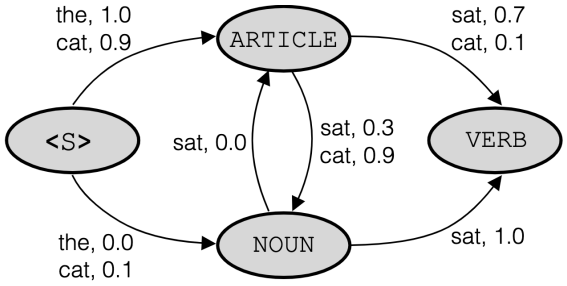


Figure 2: The normalized "conditional probability" function $p(Y_t | X_t, Y_{t-1})$.

(b) (4 points) What's the most likely part-of-speech (POS) sequence $= (Y_1, Y_2, Y_3)$ for the input sentence $= (X_1, X_2, X_3) =$ *the cat sat*? What's the corresponding conditional probability $p(Y_0, Y_1, Y_2, \cdots, Y_T | X_1, X_2, \cdots, X_T)$?

Solution: Most likely label: ARTICLE NOUN VERB.
Conditional probability: $1.0 \times 0.9 \times 1.0 = 0.9$.

(c) (4 points) What's the most likely part-of-speech (POS) sequence $= (Y_1, Y_2)$ for the input sentence $= (X_1, X_2) =$ *cat sat*? What's the corresponding conditional probability $p(Y_0, Y_1, Y_2, \cdots, Y_T | X_1, X_2, \cdots, X_T)$?

Solution: Most likely label: ARTICLE NOUN.
Conditional probability: $0.9 \times 0.7 = 0.63$.

Now, let's consider an "unnormalized" transition model.

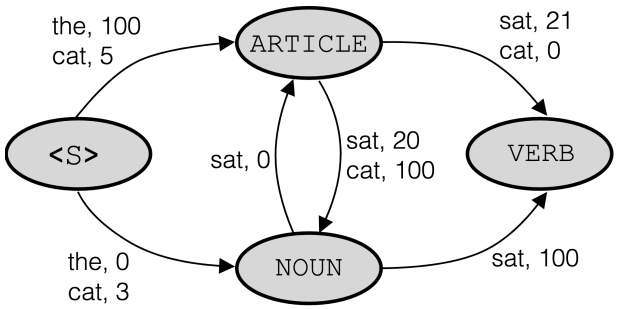


Figure 3: The unnormalized "score" function $s(Y_t, X_t, Y_{t-1})$.

Here, the edge weights will define a score function for transitions $s(Y_t, X_t, Y_{t-1})$. Note: they are not probabilities!

Now let's consider a "highest-score" label sequence. We define the score of an output sequence $y_1, y_2, \cdots, y_T$ as:

$$score(|) = score(Y_0, Y_1, Y_2 \cdots, Y_T | X_1, \cdots, X_T) = s(Y_0) + \sum_{t=1}^{T} s(Y_t, X_t, Y_{t-1}),$$

where $s(Y_0 = \langle S \rangle) = 0$, $s(Y_0 = x) = -\infty$ for any $x \neq \langle S \rangle$. We will use s to represent the score for each individual transition $s(Y_t, X_t, Y_{t-1})$, and *score* to represent the score for the entire sentence $score(|)$

(d) (4 points) What's the highest-score output sequence for the input sentence *the cat sat*? What's the corresponding score?

Solution: Highest-score label: ARTICLE NOUN VERB.
Score: $100 + 100 + 100 = 300$.

(e) (4 points) What's the highest-score output sequence for the input sentence *cat sat*? What's the corresponding score?

Solution: Highest-score label: NOUN VERB.
Score: $3 + 100 = 103$.

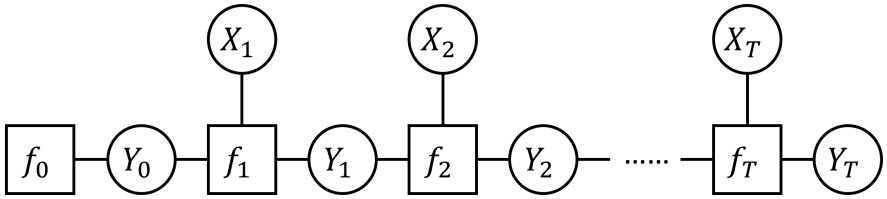
(f) (5 points) We now define the probability of an output sequence as:

$$p'(|) = p'(Y_0, Y_1, Y_2 \cdots, Y_T | X_1, \cdots, X_T) = \frac{\exp(score(|))}{\sum_{'} \exp(score('|))},$$

where the $\sum_{\cdot}$ in the denominator sums over all possible output sequence $'$ of length T (or $T+1$ if you consider y_0).

Draw a factor graph and define the factor potentials based on function s so that this factor graph represents the probability distribution $p'(|)$. (The number of variables in each factor should be strictly smaller than 4.)

Solution:



We will define factor f_0 for Y_0, and $f_0(Y_0 = \langle S \rangle) = 1$; $f_0(Y_0 = x) = 0, \forall x \neq \langle S \rangle$.
We will define factor f_t for each tuple (Y_t, X_t, Y_{t-1}).

$$f_t(Y_t, X_t, Y_{t-1}) = \exp\left(s(Y_t, X_t, Y_{t-1})\right).$$

(g) (5 points) Let's assume the conditional probability $p(Y_t|X_t, Y_{t-1})$ and the score function $s(Y_t, X_t, Y_{t-1})$ are related in the following way:

$$p(Y_t|X_t, Y_{t-1}) = \frac{\exp\left(s(Y_t, X_t, Y_{t-1})\right)}{\sum_{Y'_t} \exp\left(s(Y'_t, X_t, Y_{t-1})\right)}.$$

Under this assumption, is the "score-based" distribution: $p'(|)$ identical to the original IHMM distribution you wrote down in (a): $p(|)$? That is, for any $s(Y_t, X_t, Y_{t-1})$ and $p(Y_t|X_t, Y_{t-1})$ satisfying the condition, $p(|) = p'(|)$.

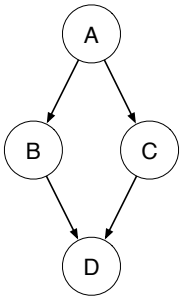
Solution: False. You can see this simply by comparing (c) and (e): the values assigned to edges roughly satisfy this condition. For example:

- for the edge($\langle S \rangle$, cat): $\frac{\exp(5)}{\exp(5)+\exp(3)} \approx 0.9$.
- for the edge(ARTICLE, sat): $\frac{\exp(21)}{\exp(21)+\exp(20)} \approx 0.7$.

The key difference between $p(|)$ and $p'(|)$ is that p' defers the normalization at the global level and p does normalization at each step t.

2 The Deciding Factor

Consider the following directed graphical model:



Assume the variables are all binary and the CPTs are as follows:

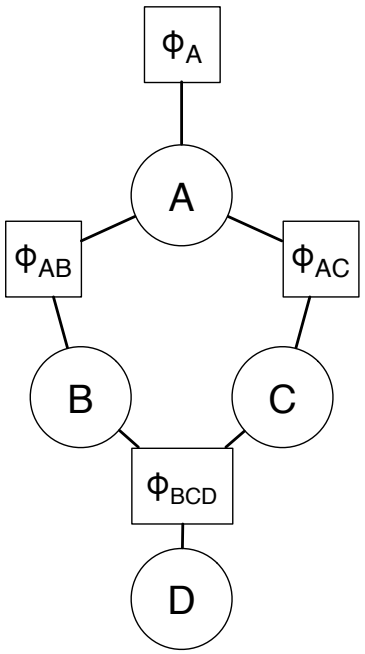
A	P(B = 1)
0	0.3
1	0.6

A	P(C = 1)
0	0.9
1	0.2

B	C	P(D = 1)
0	0	0.9
0	1	0.1
1	0	0.1
1	1	0.9

P(A = 1)
0.3

2. (a) (5 points) Draw its associated factor graph and specify the factors in terms of the CPTs given above.



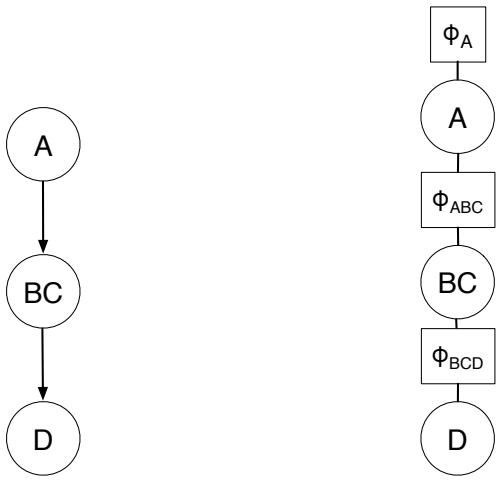
Solution:
The factors are just the CPTs.

(b) (5 points) What algorithm is appropriate for exact inference on this model?

Solution: Variable elimination.

Please go on to the next page. . .

(c) (5 points) Jody suggests converting the original directed graph to the following one, where BC is a random variable that can take on four possible values: 00, 01, 10, 11 which correspond to joint assignments to variables B and C from the original model. We also show its associated factor graph.



How does this directed graph compare in expressive power to the original one?
√ **More** ◯ Less ◯ Same Briefly explain your answer.

Solution: B and C do not have to be conditionally independent given A any more.

(d) (10 points) Provide tables for any factors in the factor graph from part c that differ from the one in part a.

Solution:
Factor ϕ_{ABC} is

A	B	C	
0	0	0	.07
0	0	1	.63
0	1	0	.03
0	1	1	.27
1	0	0	.32
1	0	1	.08
1	1	0	.48
1	1	1	.12

(e) Show how to use belief propagation on the factor graph in part c to compute $P(A \mid D = 0)$, by supplying formulas for each of the messages that is computed. Your expressions may use factor values and values of any previously computed messages. You do not need to do numeric computation.

i. (3 points) Message $\mu_{D \to \phi_{BCD}}(d)$

Solution:
D
0 1
1 0

ii. (3 points) Message $\mu_{\phi_{BCD} \to BC}(b, c)$

Solution: $\sum_D \phi_{BCD}(b, c, d) \mu_{D \to \phi_{BCD}}(d)$

iii. (3 points) Message $\mu_{BC \to \phi_{ABC}}(b, c)$

Solution: $\mu_{\phi_{BCD} \to BC}(b, c)$

iv. (3 points) Message $\mu_{\phi_{ABC} \to A}(a)$

Solution: $\sum_{b,c} \mu_{BC \to \phi_{ABC}}(b, c) \phi_{ABC}(a, b, c)$

v. (3 points) Message $\mu_{\phi_A \to A}(a)$

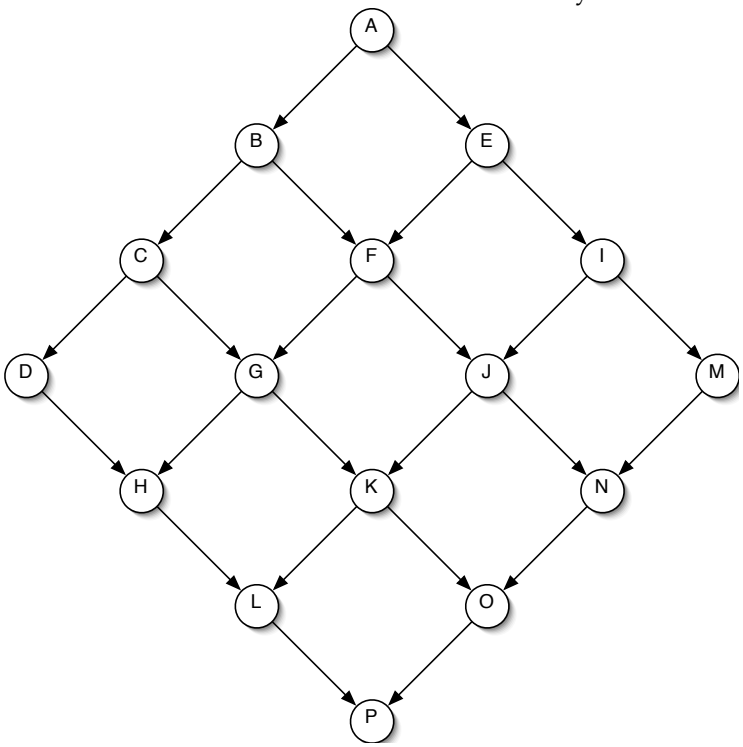
Solution: $\phi_A(a)$

vi. (5 points) Final result: $\Pr(A = a \mid D = 0)$

Solution: $\mu_{\phi_A \to A}(a) \mu_{\phi_{ABC} \to A}(a)$

3 Bayesian network inference

Consider the Bayesian network below where all variables are binary.



3. (a) (5 points) What is the size of the largest CPT in this network? **4**

   (b) (5 points) What nodes can be ignored while computing $\Pr(H|M)$? **JKLNOP**

   (c) (5 points) What is the time complexity of the problem of finding the elimination order that generates the smallest-size largest factor? **Exponential in number of nodes**

   (d) (5 points) If you were computing $\Pr(P|B = b)$ for a very unlikely value of b, would you prefer importance sampling or Gibbs sampling? Why? **importance because it's easy to sample** $B = b$

   (e) (5 points) If you were computing $\Pr(B|P = p)$ for a very unlikely value of p, would you prefer importance sampling or Gibbs sampling? Why? **Gibbs because you will never hit** $P = p$ **by forward sampling**

End of exam.