CS 221, Fall 2020
Quiz 7 Solutions

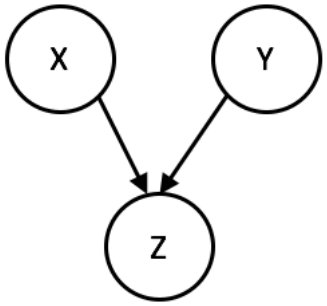
1. **[0 points]** You are allowed to consult online lecture notes, personal notes, and use a basic calculator during this quiz. You may not use the internet to access any websites other than Gradescope or the course website. You may not communicate with others to give or receive aid during or after the quiz. **I agree to abide by these rules and the Stanford Honor Code.**

   (a) Agree

   (b) Disagree

   Answer: (a)

2. Suppose X, Y, Z are binary variables describing the health of a patient Fatima. Let $X = 1$ if Fatima has the Flu and $X = 0$ if Fatima does not have the Flu. Similarly, let $Y = 1$ if Fatima has COVID-19 and $Y = 0$ otherwise. Finally, let $Z = 1$ if Fatima has fever and $Z = 0$ if Fatima does not have a fever. Both COVID-19 and Flu cause a fever, i.e. $p(Z = 1 \mid X = 1) = p(Z = 1 \mid Y = 1) = 1$. The relationship between X, Y and Z is modeled by the Bayesian network below.



   Let $p(X = 1) = \epsilon_1$ and $p(Y = 1) = \epsilon_2$ for some $\epsilon_1, \epsilon_2 < 0.5$.

   (a) **[2 points]** Select the true statement below about our belief that Fatima has COVID-19 if it is known that she has a fever, i.e. $p(Y = 1 \mid Z = 1)$.

   i. $P(Y = 1 \mid Z = 1) > p(Y = 1)$.

   ii. $P(Y = 1 \mid Z = 1) < p(Y = 1)$.

   iii. $P(Y = 1 \mid Z = 1)$ could be smaller or larger than $p(Y = 1)$ depending on the actual values of ϵ_1, ϵ_2.

   Answer: i. From Bayes rule, we have

$$\begin{aligned} p(Y = 1 \mid Z = 1) &= \frac{p(Y = 1, Z = 1)}{p(Z = 1)} \\ &= \frac{p(Y = 1) \times p(Z = 1 \mid Y = 1)}{p(Z = 1)} \\ &= \frac{p(Y = 1)}{p(Z = 1)}. \end{aligned}$$

   Assuming Fatima does not have the Flu or COVID-19, she does not have a fever, $p(Z = 1) < 1$, and thus $p(Y = 1 \mid Z = 1) > p(Y = 1)$.

   If we don't make the commonsense assumption above, we get $p(Y = 1 \mid Z = 1) \geq p(Y = 1)$ which does not correspond to any correct option. However, we would consider option iii. if this point was clarified in the assumption box.

   (b) **[2 points]** Suppose it is known that Fatima has a fever i.e. $Z = 1$, and she is waiting to get tested for COVID-19. In the meantime, she gets a Flu test which would reveals that she has the Flu. Which of the following is true about our belief that Fatima has COVID-19 before (i.e. $p(Y = 1 \mid Z = 1)$) and after the positive Flu test result (i.e. $p(Y = 1 \mid Z = 1, X = 1)$)).

i. $P(Y = 1 \mid Z = 1, X = 1) > p(Y = 1 \mid Z = 1)$.
ii. $P(Y = 1 \mid Z = 1, X = 1) < p(Y = 1 \mid Z = 1)$.
iii. $P(Y = 1 \mid Z = 1, X = 1)$ could be smaller or larger than $p(Y = 1 \mid Z = 1)$ depending on the actual values of ϵ_1, ϵ_2.

Answer: ii. This is due to *explaining away* described in Slide 12 of the Definitions module. Both COVID-19 and Flu positively influence the effect (Fever). Conditioned on the effect, further conditioning on one cause (Flu) reduces the probability of the other cause (Covid 19).

3. **[1 points]** Select the True statement from below.

   (a) Every Markov network can be efficiently converted into a Bayesian network that represents the same probability distribution.

   (b) Every Bayesian network can be efficiently converted into a Markov network that represents the same probability distribution.

   (c) Only Hidden Markov Models can be efficiently converted into Markov networks; other Bayesian networks cannot.

   Answer: (b). y, Bayesian networks are just instances of Markov networks where the normalization constant Z = 1. See Slide 6 of the probabilistic inference module for more details.

4. **[1 points]** For large graphs, we should avoid Laplace smoothing as this significantly increases the computation time for maximum likelihood estimation.

   (a) True

   (b) False.

   Answer: False. Laplace smoothing just uses additional pseudocounts to be added to all counts which does not change the computation time for maximum likelihood estimation.

5. Imagine that you are a climatologist in the year 2080 studying the history of global warming. You cannot find any records of the weather in California for ten consecutive days in the summer of 2020. However, you do find Juan's diary, which lists how many ice creams Juan ate for those ten days that summer when he was in California. Our goal is to use these observations to estimate the temperature every day. We'll simplify this weather task by assuming there are only two kinds of days: Cold and Hot, and Juan eats either 1, 2 or 3 ice creams each day. Let h_i denote the temperature of day i and let e_i denote the number of ice-creams eaten by Juan on day i. The probability distribution of the number of ice-creams eaten by Juan on any day only depends on the temperature of the day, and is as follows. Further, the temperature of a day only depends on the temperature of the preceding day.

$p(e_i = 1 \mid h_i = \text{Hot})$	0.2
$p(e_i = 2 \mid h_i = \text{Hot})$	0.4
$p(e_i = 3 \mid h_i = \text{Hot})$	0.4

$p(e_i = 1 \mid h_i = \text{Cold})$	0.5
$p(e_i = 2 \mid h_i = \text{Cold})$	0.4
$p(e_i = 3 \mid h_i = \text{Cold})$	0.1

   (a) **[2 points]** Juan's ice-cream consumption over the first four days is as follows: $\{1, 3, 2, 3\}$. What is the weight of the edge going from $H_1 = \text{Hot}$ to $H_2 = \text{Cold}$ in the lattice representation?
   Answer: The weight is $p(H_2 = \text{Cold} \mid H_1 = \text{Hot}) \times p(E_2 = 3 \mid H_2 = \text{Cold}) = 0.4 \times 0.1 = 0.04$. See Lattice Representation on Slide 6 of the Forward Backward module for more details.

$\text{p}(\text{h}_{i+1} = \text{Hot} \mid h_i = \text{Hot})$	0.6
$\text{p}(\text{h}_{i+1} = \text{Cold} \mid h_i = \text{Hot})$	0.4
$\text{p}(\text{h}_{i+1} = \text{Hot} \mid h_i = \text{Cold})$	0.5
$\text{p}(\text{h}_{i+1} = \text{Cold} \mid h_i = \text{Cold})$	0.5

(b) **[3 points]** Suppose we have computed the Forward values $F_1(h_1 = H) = 0.1$ and $F_1(h_1 = C) = 0.25$. Compute the forward value $F_2(h_2 = C)$ for the ice-consumption values above.

Answer: The weight of the edge from $H_1 = \text{Hot}$ to $H_2 = \text{Cold}$ is 0.04 calculated above.

Similarly, the weight of the edge from $H_1 = \text{Cold}$ to $H_2 = \text{Cold}$ is $0.5 \times 0.1 = 0.05$.

Applying the formula for forward values where we sum the weight of paths leading to $H_2 = \text{Cold}$, we get $0.04 * 0.1 + 0.05 * 0.25 = 0.0165$. See Slide 8 of the Forward-Backward module.

(c) **[2 points]** To make things simpler for you, we ran the Forward-Backward algorithm on the ice-cream consumption data of all ten days and obtained:

$$F_7(H_7 = \text{Hot}) = 0.8$$
$$F_7(H_7 = \text{Cold}) = 0.5$$
$$B_7(H_7 = \text{Hot}) = 0.25$$
$$B_7(H_7 = \text{Cold}) = 0.2.$$

What is the probability $p(H_7 = \text{Hot} \mid E)$ based on the values above, where E is the observed ice-cream consumption of ten days that was used to compute the Forward-Backward values above.
Answer: We compute $S_7(h_7 = \text{Hot}) = 0.8 \times 0.25 = 0.2$. We compute $S_7(h_7 = \text{Cold}) = 0.5 \times 0.2 = 0.1$. We then normalize to get the answer 0.67. See Slide 8 of the Forward-Backward module.

(d) A fellow climatologist decides to recompute the distribution of the temperatures of the days, this time using particle filtering. Let us start with two candidate particles $C = \{[h_1 = \text{Cold}], [h_1 = \text{Cold}]\}$. Recall that Juan's ice-cream consumption over the first four days is as follows: $\{1, 3, 2, 3\}$.

i. **[2 points]** In the proposal step, we extend each particle. For candidate $[h_1 = \text{Cold}]$, we propose $h_2 = \text{Hot}$ and $h_2 = \text{Cold}$ with equal probability.

A. True

B. False

Answer: True. We sample the new particle from $p(h_2 \mid h_1)$. Both $p(h_2 = \text{Hot} \mid h_1 = \text{Cold})$ and $p(h_2 = \text{Cold} \mid h_1 = \text{Cold})$ are 0.5. Intuitively, the transitions from Cold to both Hot and Cold is equal and hence the two hidden states are proposed with equal probability. See Slide 10 of the Particle Filtering module for details on the proposal distribution.

ii. **[2 points]** Suppose our proposal step resulted in the following two particles: $[h_1 = \text{Cold}, h_2 = \text{Cold}]$ and $[h_1 = \text{Cold}, h_2 = \text{Hot}]$. In the reweighting step where we compute weights of the proposed candidates, $w([h_1 = \text{Cold}, h_2 = \text{Cold}]) = w([h_1 = \text{Cold}, h_2 = \text{Hot}])$.

A. True

B. False

Answer: False. The weight of each proposed particle is $p(e_2 = 3 \mid h_2 = \text{Cold})$ and $p(e_2 = 3 \mid h_2 = \text{Hot})$ respectively. These two quantities are 0.4 and 0.1 and not equal. Intuitively, seeing a high ice-cream consumption makes Cold hidden temperature more likely and hence corresponding particle is weighted higher. See Slide 12 of the Particle Filtering module.

(e) **[2 points]** Confident in our ability to infer temperatures from ice-cream consumption, we decide to repeat this process for inferring the weather of ten days in Texas in the summer of 2020. We obtain Maria's ice-cream consumption over ten days $\tilde{e}_1, \tilde{e}_2, \ldots \tilde{e}_{10}$, but unfortunately do not know how Maria chooses to have ice-creams depending on the temperature. Select the true statement from below about the right way to use Maria's consumption to infer temperature in Texas.

i. We can directly apply supervised learning via maximum likelihood estimation (i.e. simply counting and normalizing) to estimate $p(\tilde{e}_i \mid h_i = \text{Hot})$ and $p(\tilde{e}_i \mid h_i = \text{Cold})$. Then, we perform Forward Backward like we did with Juan's data above.

ii. It is impossible to infer anything about the temperatures given only the data we have (Maria's consumption without accompanying information about how temperature affects Maria's ice-cream consumption).

iii. We can use the EM algorithm to maximize the marginal likelihood as follows. We start with some initial guess for $p(\tilde{e}_i \mid h_i = \text{Hot})$ and $p(\tilde{e}_i \mid h_i = \text{Cold})$. E-step involves performing the same computation we did above with Juan's data (Forward Backward or particle filtering) to obtain weights for different hidden temperatures. M-step revises our estimates for $p(\tilde{e}_i \mid h_i = \text{Hot})$ and $p(\tilde{e}_i \mid h_i = \text{Cold})$ using weights obtained in the E-step.

iv. We can use the EM algorithm to maximize the marginal likelihood as follows. We obtain an estimate for $p(\tilde{e}_i \mid h_i = \text{Hot})$ and $p(\tilde{e}_i \mid h_i = \text{Cold})$ in the E-step. In the M-step, we revise our estimates using the Forward Backward procedure (or particle filtering) similar to the computation we did above for Juan's data.

Answer: iii. This is the definition of the EM algorithm. We can think of the probability distribution of Maria's ice-cream consumption as a function of temperature as the graph parameters that we are estimating. This is a setting of missing data because the hidden temperatures are not observed. Hence, we should perform the EM algorithm. In the E-step we compute the posterior distribution over the hidden temperatures and in the M-step, we do maximum likelihood estimation using the posterior to update our parameters. See Slide 6 of the Expectation Maximization module for more details.

6. Laplace smoothing with pseudocount λ prevents overfitting in maximum likelihood estimation in a Bayesian network. Select True or False for the following two statements describing the role of Laplace smoothing in the context of generalization in Machine Learning.

   (a) **[1 points]** Increasing λ *increases* estimation error

      i. True

      ii. False

   Answer. False. See Slide 12 of the Generalization module (Machine Learning) for a recap of Approximation Error and Estimation Error. By adding pseudocounts in Laplace smoothing, we are forcing the estimated probability distribution to be closer to the uniform distribution—this reduces the hypothesis class (i.e. possible parameters representing the probability distributions that we are learning) and hence increases approximation error. In the limit of $\lambda \to \infty$, the hypothesis class only contains the uniform distribution.

   (b) **[1 points]** Increasing λ *increases* approximation error

      i. True

      ii. False

   Answer. True. See Slide 12 of the Generalization module (Machine Learning) for a recap of Approximation Error and Estimation Error. When we increase λ, we have reduced overfitting to the noise in the training set, and this reduces the estimation error which is the gap in performance on the test set and training set. This is similar to regularizing the norm of the weights in classification (Slide 22 of the generalization module). In the limit $\lambda \to \infty$, both expected train and test error will be the same since the estimate (uniform distribution) does not depend on the training data at all.