NATIONAL UNIVERSITY OF SINGAPORE

CS5340 - Uncertainty Modeling in AI

(Quiz 2, Semester 2 AY2020/21)

SOLUTIONS

Time Allowed: 1 hour

Instructions

- This is an open-book quiz. You may refer to any of the lecture slides and tutorials.
- You may *not* refer to any external online material or use any software to help you answer the questions.
- Please do not cheat; your answers *must* be your own. Do *not* collaborate with anyone else.
- Please put all your answers in Luminus.
- Read each question *carefully*. Don't get stuck on any one problem. The questions are *not* in any particular order of difficulty.
- Don't panic. The problems often look more difficult than they really are.
- Good luck!

Student Number.: ______________________________

Common Probability Distributions

Distribution (Parameters)	PDF/PMF
Normal (μ, σ^2)	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$
Bernoulli (r)	$r^x(1-r)^{(1-x)}$
Categorical (π)	$\prod_{k=1}^{K} \pi_k^{x_k}$
Binomial (μ, N)	$\binom{N}{x}\mu^x(1-\mu)^{N-x}$
Poisson (λ)	$\frac{\lambda^x \exp[-\lambda]}{x!}$
Beta (α, β)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
Gamma (a, b)	$\frac{1}{\Gamma(a)} b^a x^{a-1} \exp[-bx]$
Dirichlet $(\boldsymbol{\alpha})$	$\frac{\Gamma(\sum_k^K \alpha_k)}{\Gamma(\alpha_1)...\Gamma(\alpha_K)} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$
Multivariate Normal $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	$\frac{1}{(2\pi)^{D/2}\lvert\Sigma\rvert^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$
Uniform (a, b)	$\frac{1}{b-a}$
Cauchy (x_0, γ)	$\frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}$

Note: $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$ is the Gamma function.

1 True or False?

For the following questions, please answer TRUE or FALSE.

Grading Policy: Each problem is worth 1 point.

Problem 1. [1 points] Let $Z = aX + bY$ where $X \sim \mathcal{N}(3, 2)$ and $Y \sim \mathcal{N}(1, 2)$ are Gaussian distributed random variables. Then

$$\mathbb{E}[Z] = 3a + b$$

Solution: True since Z is a Gaussian with mean $a\mu_X + b\mu_Y$.

Problem 2. [1 points] Let the function $f(x) = -(x^2)$, then

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$$

Solution: False. Since $-(x^2)$ is concave. See Jensen's Inequality which we have discussed in the context of variational inference.

Problem 3. [1 points] At SoC, students were given two different types of educational activities X or Y. The manner in which students were assigned to X or Y is unknown. We observe that students who received X performed better than students who received Y in their CS5340 quizzes. True or False: Given the information above, activity X is better than Y in helping students perform better in the CS5340 quizzes.

Solution: False. See Tutorial 5 on Simpson's paradox.

Problem 4. [1 points] The Markov blanket for a node in a Bayesian Network is the set containing its parents and children.

Solution: False. The Markov blanket must also contain the co-parents.

Problem 5. [1 points] Consider an estimator $R_N(f) = \frac{1}{N}\sum_i^N f(x^{(i)})^2$ where the samples $x^{(i)}$ are drawn from $p(x)$. For the special case where, $p(x) = \mathcal{N}(0, \sigma^2)$, $R_N(f)$ is an unbiased estimator for the expectation of $f(x)$.

Solution: False. Work this out using in a similar way as slide 26 in Lecture 9 on MCMC.

Problem 6. [1 points] It is possible to model the outcome of a fair die roll using a categorical distribution. Assume a standard 6-sided die.

Solution: True. You could also use a uniform distribution.

Problem 7. [1 points] True or False:

$$(X \perp Y|Z) \wedge (X \perp Y|W) \Rightarrow (X \perp Y|Z, W)$$

In other words, $(X \perp Y|Z)$ and $(X \perp Y|W)$ implies $(X \perp Y|Z, W)$.

Solution: False. Try to come up with a counterexample where you need both Z and W to render X and Y dependent.

Problem 8. [1 points] The variance of a sum of two independent random variables X and Y is always given by:

$$\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y]$$

Solution: True. Since X and Y are independent, their covariance is zero.

Problem 9. [1 points] When we perform variational inference, we minimize the backward KL divergence $\mathbb{D}[q\|p]$. The optimization results in a distribution q that attempts to "cover" all areas of p that have non-zero probability, e.g., across different modes.

Solution: False. $\mathbb{D}[q\|p]$ is zero avoiding. Look at the lecture notes which compares the backward and forward KL divergence.

Problem 10. [1 points] Consider a fully-connected Bayesian Network with three random variables $\{X, Y, A\}$. Then,

$$\log p(X|Y) = \sum_A p(A) \log p(X|Y)$$

where $p(A)$ is a probability mass function.

Solution: True. Note that

$$\sum_a p(A) \log p(X|Y) = \log p(X|Y) \sum_A p(A) = \log p(X|Y)$$

Problem 11. [1 points] Consider the Gaussian Mixture Model that we learnt in the lectures. Suppose that the mixing coefficients $\{\pi_k\}_{k=1}^K$ and component variances $\{\Sigma_k\}_{k=1}^K$ are known. We place a standard normal prior on each of the means $\{\mu_k\}_{k=1}^K$, so $\mu_k \sim \mathcal{N}(0,1)$ for $k = 1, \ldots, K$. Then learning the parameters $\{\mu_k\}_{k=1}^K$ via EM is identical to learning the parameters via variational inference.

Solution: False. EM learns a point estimate (corresponding to the MLE) but variational inference infers a posterior distribution over the parameters.

2 Valid Transitions

For each of the matrices below, select True if the matrix is ergodic. Select False otherwise.

Problem 12. [1 points]

$$T = \begin{bmatrix} 0.8 & 0.2 & 0.0 \\ 0.2 & 0.3 & 0.5 \\ 0.5 & 0.2 & 0.3 \end{bmatrix}$$

Solution: Yes, the matrix is aperiodic and irreducible.

Problem 13. [1 points]

$$T = \begin{bmatrix} 0.0 & 0.5 & 0.5 \\ 1.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \end{bmatrix}$$

Solution: No. The matrix is not aperiodic.

Problem 14. [1 points]

$$T = \begin{bmatrix} 0.7 & 0.3 & 0.0 \\ 0.1 & 0.7 & 0.2 \\ 0.0 & 0.1 & 0.4 \end{bmatrix}$$

Solution: No. The last row does not sum to 1.

Problem 15. [1 points]

$$T = \begin{bmatrix} 0.0 & 0.2 & 0.8 \\ 0.1 & 0.1 & 0.8 \\ 0.8 & 0.1 & 0.1 \end{bmatrix}$$

Solution: Yes. The matrix is aperiodic and irreducible.

Problem 16. [1 points]

$$T = \begin{bmatrix} 0.5 & 0.0 & 0.5 \\ 0.9 & 0.0 & 0.1 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$$

Solution: No, The matrix is not irreducible.

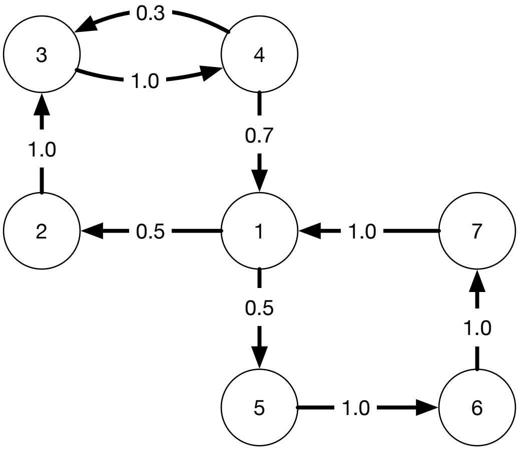
3 More MCMC

Problem 17. [2 points] Which of the following adaptive proposal distributions will not work well in an MCMC sampler (it will not lead to sampling from the stationary distribution in general)? If hyperparameters are used, you can assume that they are properly set. Assume the target distribution to be a continuous univariate distribution.

A. $q(x'|x) = \text{Poisson}(x)$
B. $q(x'|x) = \text{Normal}(x, \sigma^2)$
C. $q(x'|x) = \text{Uniform}(x - a, x + b)$
D. $q(x'|x) = \text{Cauchy}(x, \gamma)$

Solution: $q(x'|x) = \text{Poisson}(x)$ is not a valid proposal to use since there is zero probability of sampling negative x. Which condition is not satisfied by using this proposal distribution?

Problem 18. [2 points] Consider the transition graph below that shows the probabilities of transitioning between the different states. Here, the nodes are not random variables but state values. For example, $p(s_{t+1} = 2|s_t = 1) = 0.5$, $p(s_{t+1} = 3|s_t = 4) = 0.3$ and $p(s_{t+1} = 1|s_t = 3) = 0$. You can create a transition matrix from the graph by enumerating through each pair of states. Is the Markov Chain that results from this transition graph irreducible and aperiodic?

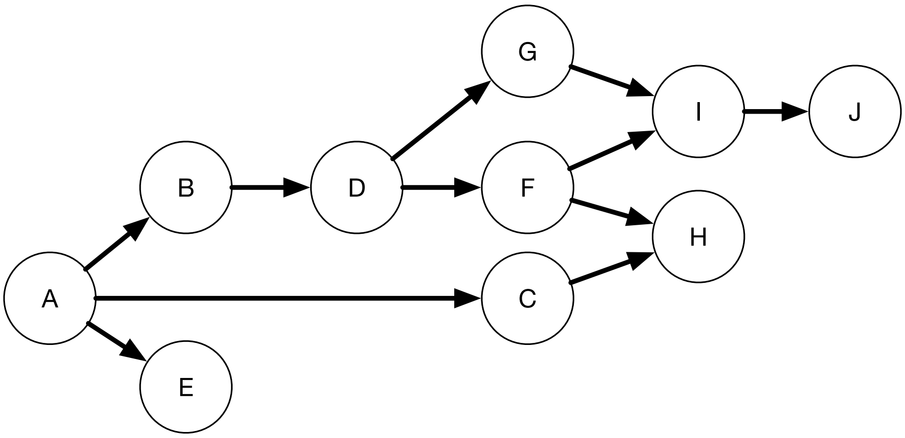


A. Irreducible and Aperiodic
B. Irreducible and Not Aperiodic
C. Not Irreducible and Aperiodic
D. Not Irreducible and Not Aperiodic

Solution: The chain is irreducible but not aperiodic.

4 Gibbs Sampling

You want to run Gibbs sampling on the following graphical model. For each of the random variables below, what is the correct conditional to sample from? **Note:** If there are multiple correct answers, select the one that conditions upon the fewest number of random variables.



Problem 19. [1 points] Sample A.

A. $p(A)$ (sample from the prior)

B. $p(A|B, C, D, F)$

C. $p(A|B, E)$

D. $p(A|B, D, E)$

E. $p(A|B, C, E)$

Solution: $p(A|B, C, E)$

Problem 20. [1 points] Sample B.

A. $p(B|A, D)$

B. $p(B|A, E)$

C. $p(B|A, D, G)$

D. $p(B|D, G, F, H, I, J)$

E. $p(B|A)$

Solution: $p(B|A, D)$

Problem 21. [1 points] Sample C.

A. $p(C|A, H, F)$

B. $p(C|A, H)$

C. $p(C|A, C, D)$

D. $p(C)$

E. $p(C|H, E)$

Solution: $p(C|A, H, F)$

Problem 22. [1 points] Sample D.

A. $p(D|B, G, F)$

B. $p(D|A, B, F, G)$

C. $p(D)$

D. $p(D|C, F)$

E. $p(D|F, I, J)$

Solution: $p(D|B, G, F)$

Problem 23. [1 points] Sample E.

A. $p(E|A)$

B. $p(E|A, B)$

C. $p(E)$

D. $p(E|C)$

E. $p(E|H)$

Solution: $p(E|A)$

Problem 24. [1 points] Sample F.

A. $p(F|C, D, G, H, I)$

B. $p(F|A, B)$

C. $p(F|D, H, I)$

D. $p(F|C, D, G)$

E. $p(F|C, D, G, H, I, J)$

Solution: $p(F|C, D, G, H, I)$

Problem 25. [1 points] Sample J.

A. $p(J|I)$

B. $p(J|A, B, C, D, E, F, G, H, I)$

C. $p(J|A, B, C, D, E, F, G, H)$

D. $p(J|G, F, I)$

E. $p(J)$

Solution: $p(J|I)$

5 A Mixture Model

For the following questions, consider the model shown in the figure below where $\mathbf{X}_n$ are K-dimensional binary random variables (having 1-of-K representation[1]), while y_n and z_n are Gaussian random variables. We first sample $\mathbf{x}_n$, which has distribution,

$$p(\mathbf{x}_n) = \prod_{k=1}^{K} \pi_k^{x_{n,k}}$$

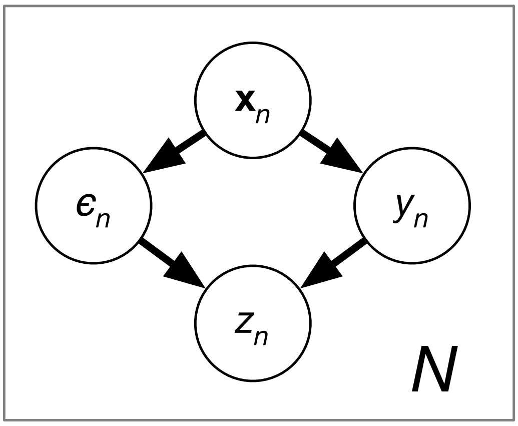
and then sample y_n from the conditional distribution,

$$p(y_n|\mathbf{x}_n) = \prod_{k=1}^{K} \mathcal{N}(\mu_k, \sigma_k^2)^{x_{n,k}}$$

where each (μ_k, σ_k^2) is associated with the component indicated by the variable $\mathbf{x}_n$. The variable Z_n is linearly related to y_n:

$$Z_n = \alpha y_n + \epsilon$$

where $\epsilon|(\mathbf{X}_n = \mathbf{x}_n) \sim \prod_{k=1}^{K} \mathcal{N}(0, v_k^2)^{x_{n,k}}$. Note that the variance of the noise term ϵ depends on $\mathbf{x}_n$. The parameters of the model are $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \alpha, \mathbf{v}^2\}$, where $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_K\}$, $\boldsymbol{\mu} = \{\mu_1, \mu_2, \ldots, \mu_K\}$, $\boldsymbol{\sigma}^2 = \{\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2\}$ and $\mathbf{v}^2 = \{v_1^2, v_2^2, \ldots, v_K^2\}$.



[1] where a particular element x_k is 1 and the rest are 0.

Problem 26. [2 points] Consider that the variables $\{\mathbf{x}_n, y_n, z_n\}_{n=1}^N$ are observed. Your friend Mickey suggests that to learn the model parameters $\boldsymbol{\theta}$ via MLE, we should use Expectation Maximization (EM). Is Mickey correct?

A. Yes, EM should be used since the above is essentially a mixture model.

B. Yes, EM should be used since the parameters are unknown.

C. Yes, EM should be used since $\mathbf{X}_n$ is categorical.

D. Yes, EM should be used since Z_n is a Gaussian.

E. No, EM is not necessary since there are no latent variables.

Solution: No, EM is not necessary since there are no latent variables. We can simply maximize the log likelihood.

Problem 27. [2 points] Consider that the variables $\{y_n, z_n\}_{n=1}^N$ are observed and the remaining variables are latent. We wish to learn the model parameters $\boldsymbol{\theta}$ using EM. Which of the following is correct Q-function to maximize during the M-step? Here, $\mathbb{E}[x_{n,k}]$ refers to the expectation of $x_{n,k}$ under the posterior found in the E-step.

A. $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{n=1}^N \sum_{k=1}^K \mathbb{E}[x_{n,k}] \Big[\log \pi_k + \log \mathcal{N}(z_n|\alpha y_n, v_k^2) + \log \mathcal{N}(y_n|\mu_k, \sigma_k^2) \Big]$

B. $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \prod_{n=1}^N \prod_{k=1}^K \mathbb{E}[x_{n,k}] \Big[\log \pi_k + \log \mathcal{N}(z_n|\alpha y_n, v_k^2) + \log \mathcal{N}(y_n|\mu_k, \sigma_k^2) \Big]$

C. $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{n=1}^N \sum_{k=1}^K \mathbb{E}[z_{n,k}] \Big[\log \pi_k + \log \mathcal{N}(x_n|\alpha y_n, v_k^2) + \log \mathcal{N}(y_n|\mu_k, \sigma_k^2) \Big]$

D. $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \prod_{n=1}^N \sum_{k=1}^K \mathbb{E}[z_{n,k}] \Big[\log \pi_k + \log \mathcal{N}(x_n|\alpha y_n, v_k^2) + \log \mathcal{N}(y_n|\mu_k, \sigma_k^2) \Big]$

E. None of the above.

Solution: Applying the definition of the Q function and the distributions above leads to $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{n=1}^N \sum_{k=1}^K \mathbb{E}[x_{n,k}] \Big[\log \pi_k + \log \mathcal{N}(z_n|\alpha y_n, v_k^2) + \log \mathcal{N}(y_n|\mu_k, \sigma_k^2) \Big]$

Problem 28. [2 points] Consider that the variables $\{z_n\}_{n=1}^{N}$ are observed and the remaining variables are latent. Specifically, consider the following six data points are observed:

$$\mathcal{D} = \{z_1 = 2.12, z_2 = 1.23, z_3 = 4.22, z_4 = 1.52, z_5 = 2.19, z_6 = 1.01\}$$

Which is following is true regarding the MLE parameters $\boldsymbol{\theta}_{\text{MLE}}$?

A. There is a unique global optimum, i.e., there is a unique $\boldsymbol{\theta}_{\text{MLE}}$ that maximizes the log-likelihood of the data.

B. There are an infinite number of possible $\boldsymbol{\theta}_{\text{MLE}}$ that maximize the log-likelihood of the data, i.e., the set of parameters that attain the maximum value of the log-likelihood is infinitely large.

C. There is no set of parameters $\boldsymbol{\theta}$ that maximizes the data log-likelihood due to aliasing.

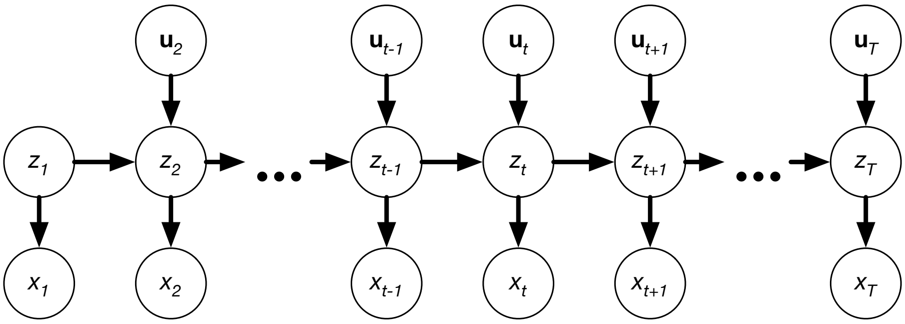
D. All of the above.

E. None of the above.

Solution: There are an infinite number of possible $\boldsymbol{\theta}$ that maximize the log-likelihood. Because y_n is not observed, there can be combinations of parameters (e.g., α and μ_k) that give rise to the exact same data.

6 A Sequential Model with Controls

Consider the following dynamic Bayesian network:



It is similar to the HMM and linear dynamical system (LDS) that we have studied, except that there are additional *binary* control variables $\mathbf{u}_t$. We represent $\mathbf{u}_t$ in a 1-of-K representation. In other words, $\mathbf{u}_t = [1, 0]$ or $\mathbf{u}_t = [0, 1]$. We can refer to each component of $\mathbf{u}_t$ using subscripts, i.e., $\mathbf{u}_t = [u_{1,t}, u_{2,t}]$.

The initial distribution and transition and emission probabilities have the form:

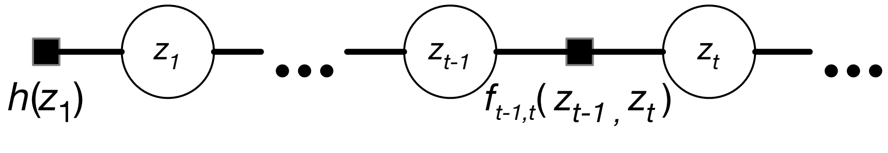
$$p(z_1) = \mathcal{N}(0, 1) \tag{1}$$

$$p(z_t|z_{t-1}, \mathbf{u}_t) = \mathcal{N}(z_t|a_{u_{k,t}} z_{t-1} + b, v^2) \tag{2}$$

$$p(x_t|z_t) = \mathcal{N}(x_t|cz_t, \sigma^2) \tag{3}$$

Note that the transitions are now controlled by $\mathbf{u}_t$. If $\mathbf{u}_t = [1, 0]$, then parameter $a_{u_{1,t}}$ is used in the transition. Else, if $\mathbf{u}_t = [0, 1]$ then $a_{u_{2,t}}$ is used. The parameters of the model are $\boldsymbol{\theta} = \{\mathbf{a}, b, c, v^2, \sigma^2\}$ where $\mathbf{a} = \{a_1, a_2\}$.

We assume that the controls $\mathbf{u}_t$ and emissions x_t are observed. Then we can consider a simplified factor tree shown below (similar to the HMM and LDS).



where the factors are given by $h(z_1) = p(z_1)p(x_1|z_1)$ and $f(z_{t-1}, z_t) = p(z_t|z_{t-1}, \mathbf{u}_t)p(x_t|z_t)$.

Problem 29. [2 points] What is the first message sent from $h(z_1) \to z_1$?

A. $p(z_1)$

B. $p(x_1|z_1)$

C. $\mathcal{N}(x_1|cz_1, \sigma^2)\mathcal{N}(z_1|0,1)$

D. All of the above.

E. None of the above.

Solution: $\mathcal{N}(x_t|cz_1, \sigma^2)\mathcal{N}(z_1|0,1)$.

Problem 30. [2 points] Consider the forward message

$$\alpha(z_t) = p(x_1, \dots, x_t, z_t|\mathbf{u}_1, \dots, \mathbf{u}_t,) = \mathcal{N}(z_t|\mu_t, s_t^2)$$

Which of the following corresponds to the forward message?

A. $p(x_t|z_t)\int_{z_{t-1}} p(z_t|z_{t-1}, \mathbf{u}_t)\alpha(z_{t-1})dz_{t-1}$

B. $\mathcal{N}(x_t|cz_t, \sigma^2)\int_{z_{t-1}} \mathcal{N}(z_t|a_{u_{k,t}}z_{t-1} + b, v^2)\alpha(z_{t-1})dz_{t-1}$

C. $\mathcal{N}(x_t|cz_t, \sigma^2)\int_{z_{t-1}} \mathcal{N}(z_t|a_{u_{k,t}}z_{t-1} + b, v^2)\mathcal{N}(z_{t-1}|\mu_{t-1}, s_{t-1}^2)dz_{t-1}$

D. All of the above.

E. None of the above.

Solution: All of the above.

Problem 31. [2 points] Consider the backward message

$$\beta(z_t) = p(x_{t+1}, \dots, x_T|z_t, \mathbf{u}_{t+1}, \dots, \mathbf{u}_T)$$

Which of the following corresponds to the backward messages?

A. $p(x_t|z_t)\int_{z_{t-1}} p(z_t|z_{t-1}, \mathbf{u}_t)\beta(z_{t-1})dz_{t-1}$

B. $\mathcal{N}(x_t|cz_t, \sigma^2)\int_{z_{t+1}} \mathcal{N}(z_t|a_{u_{k,t}}z_{t-1} + b, v^2)\beta(z_{t+1})dz_{t+1}$

C. $\int_{z_{t+1}} \mathcal{N}(x_{t+1}|cz_{t+1}, \sigma^2)\mathcal{N}(z_{t+1}|a_{u_{k,t}}z_t + b, v^2)\beta(z_{t+1})dz_{t+1}$

D. All of the above.

E. None of the above.

Solution: $\int_{z_{t+1}} \mathcal{N}(x_{t+1}|cz_{t+1}, \sigma^2)\mathcal{N}(z_{t+1}|a_{u_{k,t}}z_t + b, v^2)\beta(z_{t+1})dz_{t+1}$

Problem 32. [2 points] Which of the following corresponds to the first backward message $\beta(z_T)$?

A. 0

B. 1

C. $p(x_T|z_T)$

D. All of the above.

E. None of the above.

Solution: 1