NOTE:
collaborators for this assignment:
1) John Gatopoulos
2) Victor Zuanazzi

ML-2

Homework Assignment - 3

Shantanu Chandra
(12048461)

Q1.

(1) Below is the derivation for : $H(X,Y) = H(X) + H(Y|X)$. The derivation for $H(Y) + H(X|Y)$ can be done in similar fashion:

$$H(x,y) = -\iint p(x,y) \log p(x,y)\, dx\, dy \qquad \text{(from the definition of } \mathbb{E}[x]_{p(x)}\text{)}$$

$$= -\iint p(x|y)\cdot p(y)\, \underbrace{\log[p(x|y)\, p(y)]}\, dx\, dy \qquad \text{(Expanding the joint prob.)}$$

$$= -\iint \underbrace{p(x|y)\cdot p(y)} \log p(x|y)\, dx\, dy - \iint p(x|y)\cdot p(y) \log p(y)\, dx\, dy$$

$$= -\iint \underbrace{p(x,y)} \log p(x|y)\, dx\, dy - \int p(y) \log p(y) \underbrace{\int p(x|y)\, dx}_{=1}\, dy$$

$$= \underbrace{\mathbb{E}_{p(x,y)}[-\log p(x|y)]}_{\text{definition of conditional entropy}} + \underbrace{\mathbb{E}_{p(y)}[-\log p(y)]}_{\text{definition of entropy}}$$

$$= H(X|Y) + H(Y)$$

// Hence proved //

NOTE: for $H(Y) + H(X|Y)$ we expand $p(x,y)$ as $p(y|x)$ and follow the same steps.

(2)

$$I(x:y|z) = \mathbb{E}_{p(z)}\left[KL\left(p(x,y|z) \,\|\, p(x|z)p(y|z)\right)\right]$$

Expanding the above term using definition of $E_{p(x)}[x]$ and $KL(p\|q)$ we get :

$$= \iiint p(z)\cdot p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)\,p(y|z)}\, dx\, dy\, dz$$

We can expand this term using : $p(x,y|z) = p(x|y,z)\cdot p(y|z)$:

$$= \iiint \underbrace{p(z)}\cdot p(x|y,z)\cdot\underbrace{p(y|z)} \log \frac{p(x|y,z)}{p(x|z)}\, dx\, dy\, dz$$

$$= \iiint \underbrace{p(x|y,z)\cdot p(y,z)} \cdot \log \frac{p(x|y,z)}{p(x|z)}\, dx\, dy\, dz$$

$$= \iiint p(x,y,z)\cdot \log \frac{p(x|y,z)}{p(x|z)}\, dx\, dy\, dz$$

$$= \underbrace{\iiint p(x,y,z)\cdot \log p(x|y,z)\, dx\, dy\, dz} - \iiint p(x,y,z) \log p(x|z)\, dx\, dy\, dz$$

$$= \mathbb{E}_{p(x,y,z)}[-\log(x|y,z)] - \iint \underbrace{\left(\int p(x,y,z)\, dy\right)}_{\text{factorize out } y} \log p(x|z)\, dx\, dz$$

(using (1) & (2) from given equations of this exercise)

$$= -H(X|Y,Z) - \iint p(x,z) \log p(x|z)\, dx\, dz$$

$$= -H(X|Y,Z) + \mathbb{E}_{p(x,y,z)}[-\log(x|z)]$$

$$= H(X|Z) - H(X|Y,Z)$$ (using ① & ② from given equations of this exercise)

Proved

By expanding $p(x,y|z)$ as $p(y|x,z) \cdot p(x|z)$ we can prove in similar fashion for $I(X:Y|Z) = H(Y|Z) - H(Y|X,Z)$

Q 2.

(i) $$\text{Mult}(x|\pi) = \frac{M!}{x_1! \cdots x_k!} \pi_1^{x_1} \cdots \pi_k^{x_k}$$

general form of exponential family (Bishop 2.195):

$$p(x|\eta) = h(x)\, g(\eta) \exp\{\eta^T u(x)\}$$, sufficient stats $= \frac{1}{N} \sum_{i=1}^{N} u(x)$

or, $$p(x|\eta) = h(x) \exp\{\eta^T u(x) - A(\eta)\}$$ where, $A(\eta) =$ log partition function. $= -\log g(\eta)$

We will now re-write the $\text{Mult}(x|\pi)$ in this form -

$$\text{Mult}(x|\pi) = \frac{M!}{x_1! \cdots x_k!} e^{\log \prod_{i=1}^{k} \pi_i^{x_i}}$$

$$= \frac{M!}{x_1! \cdots x_k!} e^{\sum_{i=1}^{k} x_i \log \pi_i}$$

Now, it is given that: $\sum_i^k x_i = M \rightarrow \sum_i^K x_i = \left(M - \sum_i^{K-1} x_i\right) + \sum_i^{K-1} x_i$

and, $\sum_i^k \pi_i = 1 \rightarrow \sum_i^{K-1} \pi_i = \left(1 - \sum_i^{K-1} \pi_i\right) + \sum_i^{K-1} \pi_i$

Thus, we can rewrite above expression as:

$$= \frac{M!}{x_1! \cdots x_k!} e^{\sum_{i=1}^{k-1} x_i \log \pi_i + \left(M - \sum_i^{k-1} x_i\right) \log\left(1 - \sum_i^{k-1} \pi_i\right)}$$

$$= \frac{M!}{x_1! \cdots x_k!} e^{\sum_i^{K-1} x_i \log\left(\frac{\pi_i}{1 - \sum_j^{K-1} \pi_j}\right) + M \log\left(1 - \sum_i^{K-1} \pi_i\right)}$$ (collecting x_i & constant terms)

Comparing this with the expression above,

$$h(x) = \frac{M!}{x_1! \cdots x_k!}, \quad \eta = \begin{bmatrix} \log\left(\frac{\pi_1}{1 - \sum_j^{K-1} \pi_j}\right) \\ \vdots \\ \log\left(\frac{\pi_{K-1}}{1 - \sum_j^{K-1} \pi_j}\right) \end{bmatrix}, \quad u(x_j) = \begin{bmatrix} x_1 \\ \vdots \\ x_{K-1} \end{bmatrix}, \quad -A(\eta) = M \log\left(1 - \sum_i^{K-1} \pi_i\right)$$

, sufficient stats $= \frac{1}{N} \sum_i^N \sum_j^{K-1} x_{ij}$

Ans.

To express $A(\eta)$ in terms of η,

$$\eta_i = \log \frac{\pi_i}{1-\sum_j \pi_j} \Rightarrow e^{\eta_i} = \frac{\pi_i}{1-\sum_j \pi_j} \quad \text{— ①}$$

$$\Rightarrow \sum_i^{K-1} e^{\eta_i} = \sum_i^{K-1} \frac{\pi_i}{1-\sum_j^{K-1}\pi_j} = \frac{1}{1-\sum_j^{K-1}\pi_j} - \left(\frac{\cancel{1-\sum_j^{K-1}\pi_j}}{\cancel{1-\sum_j^{K-1}\pi_j}}\right)$$

$$\Rightarrow \boxed{1-\sum_j^{K-1}\pi_j = \frac{1}{1+\sum_i^{K-1} e^{\eta_i}}} \quad \text{— ②}$$

Substituting this in $A(\eta)$:

$$A(\eta) = -M \log \left(1+\sum_i^{K-1} e^{\eta_i}\right)^{-1}$$

$$= M \log\left(1+\sum_i^{K-1} e^{\eta_i}\right)$$

// Ans.

②

$$\mathbb{E}[x] = \frac{\partial A(\eta)}{\partial \eta_i} = \frac{M}{1+\sum_j e^{\eta_j}} \cdot e^{\eta_i} = \underbrace{\frac{M}{\frac{1}{\cancel{1-\sum_j^{K-1}\pi_j}}} \cdot \frac{\pi_i}{\cancel{1-\sum_j^{K-1}\pi_j}}}_{\text{using results from intermediate steps ① \& ② above :}} = M\pi_i \quad \text{(mean)}$$

// Ans.

$$\text{cov.}(x_i, x_j) = \frac{\partial^2 A(\eta)}{\partial \eta_i \eta_j} = \frac{\partial}{\partial \eta_j}\left(\frac{M e^{\eta_i}}{\underbrace{(1+\sum_j e^{\eta_j})}_{\text{only term with } \eta_j}}\right) = \frac{-M e^{\eta_i}}{(1+\sum_j e^{\eta_j})^2} \cdot e^{\eta_j}$$

$$= -M \frac{\frac{\pi_i}{\cancel{1-\sum_j^{K-1}\pi_j}} \, \frac{\pi_j}{\cancel{1-\sum_j^{K-1}\pi_j}}}{\cancel{(1+\sum_j e^{\eta_j})^2}} = -M\pi_i\pi_j \quad \text{(covariance)}$$

// Ans.

③ From Bishop 2.229, the conjugate prior of a distribution belonging to exponential family $p(x|\eta)$, is given by :

$$p(\eta|\chi,\nu) = f(\chi,\nu)\, g(\eta)^{\nu} \exp\{\nu \eta^T \chi\}$$

$$\propto \exp\{\nu \chi \eta^T - \nu A(\eta)\} \quad \text{(without the normalization constant)}$$

In case of multinomial, this is given as :

$$p(\eta|\chi,\nu) = \exp\left\{\sum_i^{K-1} \chi_i \nu \cdot \log\left(\frac{\pi_i}{1-\sum_j^{K-1}\pi_j}\right) + M\nu \log\left(1-\sum_j^{K-1}\pi_j\right)\right\}$$

$$= \exp\left[\log\left(\left(\prod_i^{K-1} \frac{\pi_i}{1-\sum_j \pi_j}\right)^{\chi_i \nu} \cdot \left(1-\sum_j^{K-1}\pi_j\right)^{M\nu}\right)\right]$$

$$= \prod_{i=1}^{K-1} \left(\frac{\pi_i}{\pi_K}\right)^{\chi_i \nu} \cdot (\pi_K)^{M\nu} \quad \left[\because 1-\sum_j^{K-1}\pi_j = \pi_K \Rightarrow \text{from } \sum_i^K \pi_i = 1\right]$$

$$= \prod_{i=1}^{K-1} (\pi_i)^{x_i \nu} \cdot (\pi_K)^{\nu(M - \sum_j^{K-1} x_j)}$$

$$\approx Dir(\pi | \gamma) \quad \text{where, } \gamma_i = x_i \nu + 1 \quad , \quad i = 1, \dots, K-1$$

$$\& \quad \gamma_K = \nu(M - \sum_j^{K-1} x_j) + 1 \quad , \quad i = K$$

Thus, Dirichlet distribution is the conjugate prior of Multinomial.

Ans.

(4) From Bishop 2.230, the posterior distribution of exponential family is given by:

$$p(\eta | X, \chi, \nu) = g(\eta)^{\nu + N} \exp\left\{ \eta^T \left(\sum_{n=1}^{N} u(x_i) + \nu \chi \right) \right\}$$

Thus, we observe that in prior-to-posterior update of hyperparameters,

$$\nu \longrightarrow \nu + N$$

In case of Multinomial-Dirichlet conjugacy it is:

$$\gamma_i \longrightarrow \gamma_i + N$$

$$= \begin{cases} \gamma_i + \sum_{j=1}^{n} x_i^j & \text{for } i = 1, \dots, K \\ \gamma_i + \underbrace{\sum_{j=1}^{N} \left(M^j - \sum_{i=1}^{K-1} x_i^j \right)}_{= \sum_{j=1}^{n} x_K^j} & \text{for } i = K \end{cases}$$

Thus, update is given with very easy computation as:

$$\gamma_i \longrightarrow \gamma_i + \sum_{j=1}^{n} x_i^j$$

Ans.

Q3.

(i) The necessary conditions for an ICA model are:

a) Sources are statistically independent - satisfied
b) Source distributions are non-gaussian - satisfied
c) Measurements/recordings are linear combination of sources - satisfied (with some added noise).
d) there is no time delay. - satisfied
e) datapoints are also independent w.r.t time. - satisfied

Since all the conditions are met, it is an ICA model.

Ans.

(2) $p(s_{1t}, s_{2t}, x_{1t}, x_{2t}, x_{3t})$ for $t = 1, \dots, T$ is:

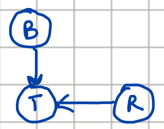
$$= \prod_{t=1}^{T} \underline{p(s_{1t}|\nu_1) \cdot p(s_{2t}|\nu_2)} \cdot \underline{p(x_{1t}|s_{1t}, s_{2t}, A_1, \sigma_1) \cdot p(x_{2t}|s_{1t}, s_{2t}, A_2, \sigma_2) \cdot p(x_{3t}|s_{1t}, s_{2t}, A_1, \sigma_2)}$$

$$= \prod_{t=1}^{T} \left[\underline{\left(\prod_{i=1}^{2} T(0, \nu_i) \right)} \cdot \prod_{i=1}^{3} \left[\underline{N(0, \sigma_i^2) + A_{i1} T(0, \nu_1) + A_{i2} T(0, \nu_2)} \right] \right] \qquad \left(\because \; x = As + \varepsilon \text{ and } A = (A_1, A_2) \right)$$

$$= \prod_{t=1}^{T} \prod_{i=1}^{2} \prod_{j=1}^{3} T(0, \nu_i) \left[N(0, \sigma_j^2) + A_{j1} T(0, \nu_1) + A_{j2} T(0, \nu_2) \right]$$

Ans.

(3) "Explaining away" is a phenomenon that occurs when we have 2 random variables conditioning a third (ie, a collider case). For eg, in the graph below we have B = basketball game, R = raining and T = traffic. We see that B and R are independent events, ie, basketball game & rain can occur independent of the other.



However, when we observe traffic when it is not raining ($P(T=1|R=0)$), the probability of basketball game goes up. Thus, given the value of conditioned variable and one of the causes, the other cause can be "explained away" (to some degree), ie, the causes become dependent.

This is also the case in our ICA model. The signals are the cause variables and the observed mixture is the conditioned variable. When we observe the mixture sound & know one source s_{1t} then we can infer the value of s_{2t} easily (thus s_{1t} "explains away" s_{2t}).

(4) a) $x_1 \perp\!\!\!\perp x_2 \,|\, \phi$: False

Using Bishop section 8.2.2 a) rule, x_1 & x_2 are not d-separated & thus not independent (x_1 & x_2 are connected by tail-to-tail by s_{1t} & s_{2t} & $\{s_{1t}, s_{2t}\} \notin \phi$)

b) $s_1 \perp\!\!\!\perp s_2 \,|\, \phi$: True

Using Bishop section 8.2.2 b) rule, s_1 & s_2 are d-separated and thus independent. (s_1 & s_2 connected head-to-head by x_1, x_2 & x_3 and $\{x_1, x_2, x_3\} \notin \phi$)

c) $x_1 \perp\!\!\!\perp s_1 \,|\, \phi$: False.

They are directly connected by a dependency edge.

d) $x_1 \perp\!\!\!\perp x_2 \,|\, \{s_1, s_2\}$: True

Using Bishop section 8.2.2 b) rule, x_1 & x_2 are d-separated and thus independent. (x_1 & x_2 are connected by tail-to-tail by $\underline{s_{1t} \,\&\, s_{2t}}$ & $\{s_1, s_2\} \in \underline{\{s_1, s_2\}}$)

conditioning set

e) $x_1 \perp\!\!\!\perp x_2 \,|\, \{s_1\}$: False

Using Bishop section 8.2.2 b) rule, x_1 & x_2 are not d-separated and thus not independent (x_1 & x_2 are connected by tail-to-tail by s_1 & s_2 & $s_1 \in \{s_1, s_2\}$)

[but path s_2 is still open]

f) $s_1 \perp\!\!\!\perp s_2 \mid \{x_1, x_2, x_3\}$: False

Using Bishop section 8.2.2 b) rule, s_1 & s_2 are not d-separated & hence not independent (s_1 & s_2 connected head-to-head by x_1, x_2 & x_3 and $\{x_1, x_2, x_3\} \in \{x_1, x_2, x_3\}$).

g) $s_1 \perp\!\!\!\perp s_2 \mid x_1$: False

Using Bishop section 8.2.2 b) rule, s_1 & s_2 are not d-separated & hence not independent (s_1 & s_2 connected head-to-head by x_1, x_2 & x_3 and $\{x_1\} \in \{x_1, x_2, x_3\}$).

h) $x_1 \perp\!\!\!\perp s_1 \mid \{s_2, x_2, x_3\}$: False.

They are directly connected by a dependency edge.

5) Markov blanket of a node 'x' is the minimal set of nodes that isolates 'x' from the graph. In other words it is a set that comprises of parents, children & co-parents of 'x'.

Markov Blanket of s_1 : children = x_1, x_2, x_3
parent = None
co-parent = s_2

$\therefore MB(s_1) = \{x_1, x_2, x_3, s_2\}$ // Ans.

Markov Blanket of x_1 : children = None
parent = s_1, s_2
co-parent = None

$\therefore MB(x_1) = \{s_1, s_2\}$ // Ans.

6) We know,

$$p_x(x) = p_s(s(x)) \cdot |\det(\mathrm{Jac}\; s \to x)|$$

On explicitly writing down the parameters :

$$p(\{x_t\} \mid W, \nu_i) = \prod_{t=1}^{T} p_s(Wx_t \mid \{\nu_i\}) \cdot \left|\det \frac{\partial s}{\partial x}\right|$$

where, $s = Wx$

$$\frac{\partial s}{\partial x} = \frac{\partial}{\partial(x_1 \dots x_k)} \sum_{k=1}^{K} W_k x_{kt} = W$$

Thus,

$$p(\{x_t\} \mid W, \nu_i) = \prod_{t=1}^{T} \prod_{i=1}^{K} p(s_{it}) \cdot |\det W|$$

$$= \prod_{t=1}^{T} |\det W| \prod_{i=1}^{K} \mathcal{T}(s_i \mid 0, \nu_i)$$

// Ans.

7) $$\log p(\{x_t\} \mid W, \nu_i) = \log\left[\prod_{t=1}^{T} |\det W| \cdot \prod_{i=1}^{K} \mathcal{T}(s_i \mid 0, \nu_i)\right]$$

$$= \sum_{t=1}^{T} \log|\det W| + \sum_{i=1}^{K} \log \mathcal{T}(s_i \mid 0, \nu_i)$$

// Ans.

⑧ We use the SGA to optimize the above log-likelihood. We initialize the mixture matrix randomly and then iteratively update W such that the above log-likelihood is maximized. This is done by taking the gradient of the above eq. (ie, differentiating w.r.t to W) and updating the current 'W' in the direction of this gradient. Mathematically this translates to:

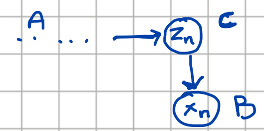
$$W_{new} \leftarrow W_{old} + \eta \nabla_W LL$$

where, $\nabla_W LL$ is gradient of log-likelihood w.r.t W
and η = learning rate (to control magnitude of the updates).

⑨ We would expect $K >> T$ case to cause overfitting. This because in this case we have way more parameters to estimate (K^2, for square matrix W of $K \times K$ size) than we actual datapoints ($K \cdot T$). In such a case there will be MANY functions/solutions for K that will explain these observations perfectly. This is a classic case of overfitting.

Q4.

① If A is d-separated from B given C then $A \perp\!\!\!\perp B \mid C$ holds.

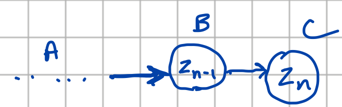


Here, we have:

$A = (x_1, \cdots x_{n-1})$, $B = x_n$, $C = z_n$

Using Bishop section 8.2.2 a) rule we can say that A is d-separated by B given C. (all paths from A to B are blocked as C connects them head-to-tail & is in the conditional set)

② If A is d-separated from C given B then $A \perp\!\!\!\perp C \mid B$ holds.



Here, we have:

$A = (x_1, \cdots x_{n-1})$, $B = z_{n-1}$, $C = z_n$

Using Bishop section 8.2.2 a) rule we can say that A is d-separated by B given C. (all paths from A to C are blocked as B connects them head-to-tail & is in the conditional set)

③ On expressing the given conditional as joint we get:

$$p(x_{n+1} \cdots x_N \mid z_n, z_{n+1}) = \frac{p(x_{n+1} \cdots x_N, z_n, z_{n+1})}{p(z_n, z_{n+1})}$$

$$= \frac{\sum_{z_{n+2}} \cdots \sum_{z_N} p(x_{n+1} \cdots x_N, z_n \cdots z_N)}{p(z_n, z_{n+1})}$$ [Marginalizing the non-essential variables]

Using graph factorization (Bishop 8.5): $p(x) = \prod_{k=1}^{K} p(x_k) \cdot p(x_k \mid pa_k)$

The given graph can be written as:

$$p(x_1 \cdots x_N, z_1, \cdots z_N) = p(z_1) p(x_1 \mid z_1) \prod_{i=2}^{N} p(z_i \mid z_{i-1}) p(x_i \mid z_i)$$

or, $$p(x_{n+1} \cdots x_N \mid z_n, z_{n+1}) = \frac{\cancel{p(z_n)}\, \cancel{p(z_{n+1} \mid z_n)}\, p(x_{n+1} \mid z_{n+1}) \sum_{z_{n+2}} \cdots \sum_{z_N} \prod_{i=n+2}^{N} p(z_i \mid z_{i-1}) p(x_i \mid z_i)}{\cancel{p(z_{n+1} \mid z_n)}\, \cancel{p(z_n)}}$$

$$= p(x_{n+1}|z_{n+1}) \underbrace{\sum_{z_{n+2}} \cdots \sum_{z_{N-1}} \left[\prod_{i=n+2}^{N-1} p(z_i|z_{i-1})\, p(x_i|z_i) \right]} \cdot \underbrace{\sum_{z_N} p(z_N|z_{N-1})\, p(x_N|z_N)}$$

breaking one off

$$= p(x_{n+1}|z_{n+1}) \sum_{z_{n+2}} \cdots \sum_{z_{N-1}} \left[\prod_{i=n+2}^{N-1} p(z_i|z_{i-1})\, p(x_i|z_i) \right] \cdot p(x_N|z_{N-1})$$

[Using: $\sum_B p(A|B) \cdot p(B|C) = p(A|C)$]

$$= p(x_{n+1}|z_{n+1}) \underbrace{\sum_{z_{n+2}} \cdots \sum_{z_{N-2}} \left[\prod_{i=n+2}^{N-2} p(z_i|z_{i-1})\, p(x_i|z_i) \right]} \sum_{z_{N-1}} \underbrace{p(z_{N-1}|z_{N-2})}\, \underbrace{p(x_{N-1}|z_{N-1}) \cdot p(x_N|z_{N-1})}_{= p(x_{N-1}, x_N | z_{N-1})}$$

breaking off another term

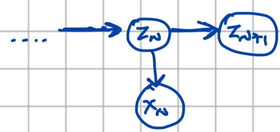
$$= p(x_{n+1}|z_{n+1}) \sum_{z_{n+2}} \cdots \sum_{z_{N-2}} \left[\prod_{i=n+2}^{N-2} p(z_i|z_{i-1})\, p(x_i|z_i) \right] \underbrace{\sum_{z_{N-1}} p(z_{N-1}|z_{N-2})\, p(x_{N-1}, x_N | z_{N-1})}_{= p(x_{N-1}, x_N | z_{N-2})}$$

$$= p(x_{n+1} \cdots x_N \mid z_{N+1})$$

[using the graph factorization of given graph]

ANS.

4 Following the pattern of graph, z_{N+1} is added as a new node & the new graph factorization gives:



$$p(x_1 \cdots x_N, z_1 \cdots z_N, z_{N+1}) = p(z_1)\, p(x_1|z_1) \left[\prod_{i=2}^{N} p(z_i|z_{i-1})\, p(x_i|z_i) \right] \cdot \underbrace{p(z_{N+1}|z_N)}$$

There is no x_{N+1} given hence this term is outside.

Now, we are given:

$$p(z_{N+1}|z_N, x) = \frac{p(x, z_N, z_{N+1})}{p(x, z_N)}$$

$$= \frac{\sum_{z_1} \cdots \sum_{z_{N-1}} p(x, z_1 \cdots z_N, z_{N+1})}{p(x, z_n)}$$

$$= \frac{p(z_{N+1}|z_N)\, p(x_{N+1}|z_{N+1}) \prod_{i=1}^{N-1} \sum_{z_{N-i}} p(z_N|z_{N-1}) \cdot p(x_{N-i}|z_{N-i})}{p(x, z_N)}$$

$$= \frac{p(z_{N+1}|z_N) \cdot \cancel{p(x|z_{N+1})} \cdot \cancel{p(z_N)} \prod_{i=1}^{N-1} p(x_i)}{\cancel{p(x|z_N)}\, \cancel{p(z_N)}}$$

$$= p(z_{N+1}|z_N)$$

Ans.