1)

RHS

$$\underset{\theta \in \Theta}{\arg\min}\ E_{\hat{p}(x)} \left[D_{KL} \left(\hat{p}(y|x) \,||\, p_\theta (y|x) \right) \right]$$

$$= \underset{\theta \in \Theta}{\arg\min}\ E_{\hat{p}(x)} \left[\log \hat{p}(y|x) \right] - E_{\hat{p}(x)} \left[\log p_\theta (y|x) \right]$$

Since first term does not depend on θ, it is equivalent to

$$\underset{\theta \in \Theta}{\arg\max}\ E_{\hat{p}(x)} \left[\log p_\theta (y|x) \right]$$

LHS

$$\underset{\theta \in \Theta}{\arg\max}\ E_{\hat{p}(x,y)} \left[\log p_\theta (y|x) \right]$$

$$= \underset{\theta \in \Theta}{\arg\max} \sum_x \hat{p}(x,y) \left[\log p_\theta (y|x) \right]$$

$$= \underset{\theta \in \Theta}{\arg\max} \sum_x \hat{p}(x)\, \hat{p}(y|x) \left[\log p_\theta (y|x) \right]$$

Since $\hat{p}(y|x)$ is not dependent on θ, it can be taken out, So, the above expression is equivalent to

$$= \underset{\theta \in \Theta}{\arg\max} \sum_x \hat{p}(x) \left[\log p_\theta (y|x) \right]$$

$$= \underset{\theta \in \Theta}{\arg\max}\ E_{\hat{p}(x)} \left[\log (p_\theta (y|x)) \right]$$

which is what our RHS is equivalent to. Hence, both are equivalent problems

2) LHS

$$P_\theta(y|x) = \frac{P_\theta(x,y)}{P_\theta(x)}$$

$$= \frac{\pi_y \exp\left(-\frac{1}{2\sigma^2}(x-\mu_y)^T(x-\mu_y)\right) . Z^{-1}(\sigma)}{\sum_i^k \pi_i \exp\left(-\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i)\right) . Z^{-1}(\sigma)}$$

Since, $Z^{-1}(\sigma)$ in the denominator is not dependent of i, it can be taken out of summation and cancel with $Z^{-1}(\sigma)$ in numerator.

$$= \frac{\pi_y \exp\left(-\frac{1}{2\sigma^2}(x-\mu_y)^T(x-\mu_y)\right)}{\sum_i^k \pi_i \exp\left(-\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i)\right)} \quad — (1)$$

Simplifying the numerator of (1)

$$-\frac{1}{2\sigma^2}(x-\mu_y)^T(x-\mu_y) =$$

$$= -\frac{1}{2\sigma^2}(x^T-\mu_y^T)(x-\mu_y)$$

$$= -\frac{1}{2\sigma^2}\left(x^T(x-\mu_y) + (\mu_y^T x - \mu_y^T \mu_y)\right)$$

$$= x^T \frac{(x-\mu_y)}{-2\sigma^2} + \left(\frac{\mu_y^T x - \mu_y^T \mu_y}{-2\sigma^2}\right)$$

In this $\frac{(x-\mu_y)}{-2\sigma^2} \in R^n$

$$\left(\frac{\mu_y^T x - \mu_y^T \mu_y}{-2\sigma^2}\right) \in R_{1\times 1}$$

$$= x^T s^y + t^y$$

where $s^y = \frac{(x-\mu_y)}{-2\sigma^2}$

$$t^y = \frac{\mu_y^T x - \mu_y^T \mu_y}{-2\sigma^2}$$

So ① can be written as

$$= \frac{\pi_y \exp(x^T s^y + t^y)}{\sum_i^k \pi_i \exp(x^T s^i + t^i)}$$

$$= \frac{\exp(\log \pi_y) \exp(x^T s^y + t^y)}{\sum_i^k \exp(\log \pi_i) \exp(x^T s^i + t^i)}$$

$$= \frac{\exp(x^T s^y + (t^y + \log \pi_y))}{\sum_{i=1}^k \exp(x^T s^i + (t^i + \log \pi_i))}$$

which is equivalent to RHS.

$$Pr(y|x) = \frac{\exp(x^T w_y + b_y)}{\sum_i \exp(x^T w_i + b_i)}$$

3(a)

$$P(x_1, x_2 \ldots x_n) = P(x_1)\, P(x_2|x_1)\, P(x_3|x_1, x_2) \ldots P(x_n | x_1 \ldots x_{n-1})$$

- $P(x_1)$ requires $k_1 - 1$ parameters
- $P(x_2|x_1)$ requires $k_2 \cdot (k_1 - 1)$ parameters
- $P(x_3|x_1, x_2)$ requires $k_3 \cdot k_2 \cdot (k_1 - 1)$

$\vdots$

- $P(x_n | x_1 \ldots x_{n-1})$ requires $(k_1 - 1) \cdot k_2 \ldots k_{n-1}$ parameters

So, total number of params required are

$$= (k_1 - 1)(1 + k_2 + k_2 . k_3 + \ldots + k_2 \ldots k_{n-1})$$

(b) If each variable X_i is independent of others, then

$$P(X_1, X_2 \dots X_n) = P(X_1)\, P(X_2) \dots P(X_n)$$

- For $P(X_1)$, we require $k_1 - 1$ parameters
- For $P(X_n)$, we require $k_n - 1$ parameters

So, total we require $\sum_{i=1}^{n} k_i - 1$ parameters

(c)

$$P(x_1, \dots x_n) = P(x_1) \cdot P(x_2 | x_1) \cdot P(x_3 | x_1, x_2) \dots \dots$$
$$P(x_{m+1} | x_1, x_2 \dots x_m) \cdot$$
$$P(x_{m+2} | x_2 \dots \dots x_{m+1})$$
$$\vdots$$
$$P(x_n | x_{n-m} \dots \dots x_{n-1})$$

As discussed earlier, for first "m" terms, the number of parameters required are

$$\boxed{total_{first\_m} = (k_1 - 1)(1 + k_2 + k_2 . k_3 + . \quad + \quad k_2 \dots k_{m-1})}$$

For each term 'i' in the above equation (There will be "n-m" such terms)

$$P_i = (k_{i-m} - 1)(k_{i-m+1}) \dots (k_{i-1})$$
$$= (k_{i-m} - 1) \prod_{j=i-m+1}^{i-1} k_j$$

So, total number of params for last n-m terms are

$$total_{last\_i\_m} = \sum_{t=m+1}^{n} P_m$$

So, overall $\text{total} = \text{total}_{\text{first\_m}} + \text{total}_{\text{last\_i\_m}}$

4) Let's take the case where $n=2$ i.e there are only 2 variables X_1, X_2.

where

$$P_f(X_1) = N(0,1)$$

$$P_f(x_2 \mid x_1) = N(\mu_\theta(x_1), \Sigma_\theta(x_1))$$

$$P_f(x_2) = \int P_\theta(x_2, x_1)\, dx_1$$

Essentially we need to find all the possible values of x_1 and for each get $P(x_2 \mid x_1)$ to find $p(x_2)$. That is a ~~not~~ mixture of infinite gaussians which is not a gaussian. Whereas,

$$P_r(x_2) = N(0,1)$$

· An intuitive example of this is large language models which are used to generate english sentences. A model trained in reverse order may not be able to used for Q&A use cases because in that "Q" comes before "A" and "A" is dependent on "Q" which is not modelled in reverse order.

5 (a)

$$E(A) = E\left[\frac{1}{K}\sum_{i=1}^{k} p(x|z^{(i)})\right]$$

$$= \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{K}\sum_{i=1}^{k} p(x|z^{i})\right)$$

$$= \frac{1}{K}\sum_{i=1}^{k}\left(\frac{1}{n}\sum_{j=1}^{n} p(x|z^{(j)})\right)$$

Let's say "n" is a very large number, then

$$= \frac{1}{K}\sum_{i=1}^{k}\left(\sum p(z)\, p(x|z^{(j)})\right)$$

$$= \frac{1}{K}\sum_{i=1}^{K} p(x)$$

$$= p(x)$$

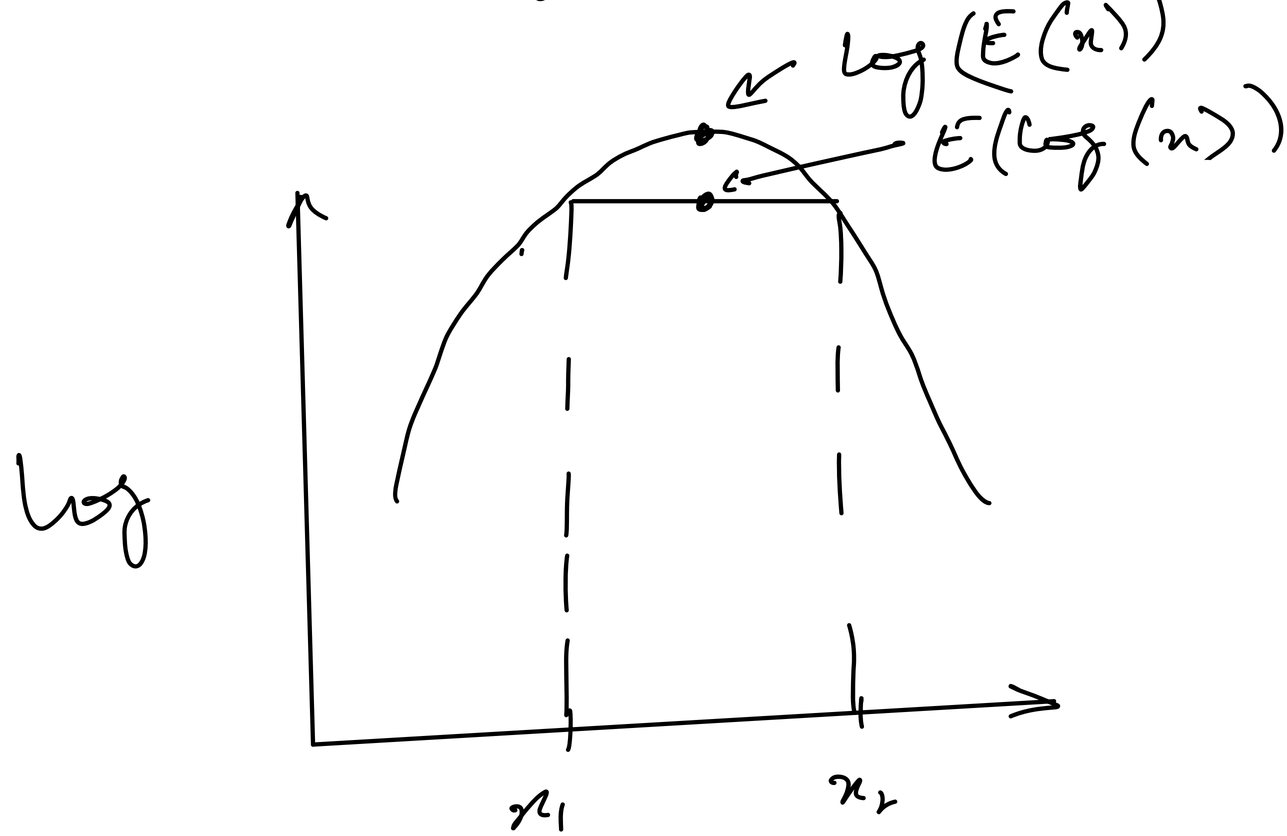
Hence: A is an unbiased estimator of $p(x)$.

5 (b)

$$E(\log A) \leq \log E(A) \quad \text{(Jensen's Eq.)}$$

$$\leq \log p(x) \qquad \because E(A) = p(x) \text{ as per 5(a)}$$

So, by definition of unbiased estimator, log A is not an unbiased estimator. For that, LHS had to be strictly equal to RHS



6 a) $n = 16$ $(\lceil \log_2 50257 \rceil)$ (ceil ($\log_2$ (50257)))

6 b) Weight matrix of Fully connected layer will increase from (768×50257) to (768×60000).

No. of bias terms will increase from 50275 to 60000

So the increase will be

$768(60000 - 50257) + (60000 - 50257) =$

7,492,367