Q1) To prove,

$$\arg\max_{\theta \in \Theta} E_{\hat{p}(x,y)}[\log p_\theta(y|x)] = \arg\min_{\theta \in \Theta} E_{\hat{p}(x)}[D_{KL}(\hat{p}(y|x) \| p_\theta(y|x))]$$

<u>Proof</u>:

KL divergence for each x:

$$D_{KL}(\hat{p}(y|x) \| p_\theta(y|x)) = \sum_y \hat{p}(y|x) \log \frac{\hat{p}(y|x)}{p_\theta(y|x)}$$

$$= \sum_y \hat{p}(y|x) (\log \hat{p}(y|x) - \log p_\theta(y|x))$$

Taking expectation over x,

$$E_{x \sim \hat{p}(x)}[D_{KL}(\hat{p}(y|x) \| p_\theta(y|x))] = \sum_x \hat{p}(x) D_{KL}(\hat{p}(y|x) \| p_\theta(y|x))$$

$$= \sum_x \hat{p}(x) \sum_y \hat{p}(y|x) [\log \hat{p}(y|x) - \log p_\theta(y|x)]$$

$$= \sum_{x,y} \hat{p}(x,y) [\log \hat{p}(y|x) - \log p_\theta(y|x)]$$

$$= E_{\hat{p}(x,y)}[\log \hat{p}(y|x)] - E_{\hat{p}(x,y)}[\log p_\theta(y|x)]$$

$E_{\hat{p}(x,y)}[\log \hat{p}(y|x)]$ is independent of θ & can be treated as a constant, C.

$$\therefore E_{x \sim \hat{p}(x)}[D_{KL}(\hat{p}(y|x) \| p_\theta(y|x))] = C - E_{\hat{p}(x,y)}[\log p_\theta(y|x)]$$

As C is a constant, minimizing the KL divergence over θ is equivalent to maximizing the expected log-likelihood.

$$\arg\min_{\theta}\left(C - E_{\hat{p}(x,y)}[\log p_\theta(y|x)]\right) = \arg\max_{\theta} E_{\hat{p}(x,y)}[\log p_\theta(y|x)]$$

$$\therefore \arg\max_{\theta\in\Theta} E_{\hat{p}(x,y)}[\log p_\theta(y|x)] = \arg\min_{\theta\in\Theta} E_{\hat{p}(x)}[D_{KL}(\hat{p}(y|x)\,||\,p_\theta(y|x))]$$

Q2) $p_\theta(y) = \pi$, $\sum_{y=1}^{k} \pi_y = 1$, $p_\theta(x|y) = \mathcal{N}(x|\mu_y, \sigma^2 I)$

$$P_\gamma(y|x) = \frac{\exp(x^T w_y + b_y)}{\sum_{i=1}^{k} \exp(x^T w_i + b_i)}$$

To prove - For any choice of θ, there exists γ such that $p_\theta(y|x) = p_\gamma(y|x)$

Proof:

$$p_\theta(x,y) = \pi_y \mathcal{N}(x|\mu_y, \sigma^2 I)$$

$$\mathcal{N}(x|\mu_y, \sigma^2 I) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_2}||x-\mu_y||^2\right)$$

$$p_\theta(y|x) = \frac{p_\theta(x,y)}{p_\theta(x)} = \frac{\pi_y \mathcal{N}(x|\mu_y, \sigma^2 I)}{\sum_{i=1}^{k} \pi_i \mathcal{N}(x|\mu_i, \sigma^2 I)}$$

$$p_\theta(y|x) = \frac{\pi_y \exp(-1/2\sigma_2 ||x-\mu_y||^2)}{\sum_{i=1}^{k} \pi_i \exp(-1/2\sigma_2 ||x-\mu_i||^2)}$$

Expanding the quadratic term :

$$\|x-\mu_y\|^2 = (x-\mu_y)^T(x-\mu_y) = x^Tx - 2x^T\mu_y + \mu_y^T\mu_y$$

$$P_\theta(y|x) = \frac{\pi_y \exp\left(-\frac{x^Tx}{2\sigma^2} + \frac{x^T\mu_y}{\sigma^2} - \frac{1}{2\sigma^2}\mu_y^T\mu_y\right)}{\sum_{i=1}^{K}\pi_i \exp\left(-\frac{x^Tx}{2\sigma^2} + \frac{x^T\mu_i}{\sigma^2} - \frac{\mu_i^T\mu_i}{2\sigma^2}\right)}$$

$$P_\theta(y|x) = \frac{\pi_y \exp\left(\frac{x^T\mu_y}{\sigma^2} - \frac{\mu_y^T\mu_y}{2\sigma^2}\right)\cdot\cancel{\exp\left(\frac{-x^Tx}{2\sigma^2}\right)}}{\cancel{\exp\left(\frac{-x^Tx}{2\sigma^2}\right)}\sum_{i=1}^{K}\pi_i \exp\left(\frac{x^T\mu_i}{\sigma^2} - \frac{\mu_i^T\mu_i}{2\sigma^2}\right)}$$

$$P_\theta(y|x) = \frac{\pi_y \exp\left(\frac{x^T\mu_y}{\sigma^2} - \frac{\mu_y^T\mu_y}{2\sigma^2}\right)}{\sum_{i=1}^{K}\pi_i \exp\left(\frac{x^T\mu_i}{\sigma^2} - \frac{\mu_i^T\mu_i}{2\sigma^2}\right)}$$

we need to express the posterior in the form:

$$P_\gamma(y|x) = \frac{\exp(x^Tw_y + b_y)}{\sum_{i=1}^{K}\exp(x^Tw_i + b_i)}$$

to match the exponents of $P_\theta(y|x)$ with $P_\gamma(y|x)$,

let $w_y = \frac{\mu_y}{\sigma^2}$, $b_y = -\frac{\mu_y^T\mu_y}{2\sigma^2} + \log\pi_y$

Thus $P_\theta(y|x)$ can be written as

$$P_\theta(y|x) = \frac{\exp(x^Tw_y + b_y)}{\sum_{i=1}^{K}\exp(x^Tw_i + b_i)} = P_\gamma(y|x)$$

This is standard form of multi-class logistic regressio

Q3)

n discrete random variables $\{X_i\}_{i=1}^{n}$ each having K_i different outcomes

A) Total no. of parameters needed to express the joint distribution without conditional independence

Total number of joint outcomes $= K_1 \times K_2 \times \dots \times K_n$
$= \prod_i^n K_i$

But since probabilites sum to 1, we would need 1 parameter less to express the joint.

Hence, total parameters $= \prod_{i=1}^{n} K_i - 1$.

B) To use only $\sum_{i=1}^{n} (K_i - 1)$ parameters, the parameters for each variable need to be specified independently. This is achieved when the random variables are mutually independent.

C) For $i <= m$, No conditional independence. X_i depends on all predecessors

For $i > m$, X_i is conditionally independent of all its ancestors except most recent m ones

For a Bayesian network, no. of independent parameters:

(no. of values for $X_i - 1$) $\times$ (No. of parent config)

→ For $i <= m$,

No. of possible parent config: $\prod_{j=1}^{i-1} k_j \quad \forall\, i = 2, 3, \dots m$

For each config, $k_i - 1$ free parameters.

∴ Total parameters $= (k_i - 1) \prod_{j=1}^{i-1} k_j$

Total for $i <= m$, we get → $\sum_{i=1}^{m} \left((k_i - 1) \prod_{j=1}^{i-1} k_j \right)$

→ For $i > m$,

Total no. of parent config: $\prod_{j=i-m}^{i-1} k_j$

Total no. of free parameters $= (k_i - 1) \prod_{j=i-m}^{i-1} k_j$

Total for $i > m$, we get → $\sum_{i=m+1}^{n} (k_i - 1) \prod_{j=i-m}^{i-1} k_j$

Total parameters:

$$\sum_{i=1}^{m} \left((k_i - 1) \prod_{j=1}^{i-1} k_j \right) + \sum_{i=m+1}^{n} (k_i - 1) \prod_{j=i-m}^{i-1} k_j$$

---

5) $p(x) = \int_z p(x, z)\, dz$

A) $A(z^{(1)}, \dots, z^{(K)}) = \frac{1}{K} \sum_{i=1}^{K} p(x \mid z^{(i)}), \quad z^{(i)} \sim p(z)$

To show the above expression is an unbiased estimator, we need to show,

$$E[A(z^{(1)}, \dots, z^{(K)})] = p(x)$$

proof:

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

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

we know, $\underset{z \sim p(z)}{E}[p(x|z)] = \int_z p(x|z)p(z)dz = p(x)$

$$\therefore E[A(z^{(1)}, \dots, z^{(K)})] = \frac{1}{K}\sum_{i=1}^{K} p(x)$$

$$= \frac{1}{\cancel{K}} \cdot \cancel{K} \cdot p(x)$$

$$= p(x)$$

Thus the Monte Carlo estimator A provides an unbiased estimate of $p(x)$.

B) To find if log A is an unbiased estimator or not,

we need to check if $E[\log A] = \log p(x)$

Using Jensen's inequality for concave function

$$f(E[X]) \geq E[f(X)]$$

$$\log(E[A]) \geq E[\log A]$$

we know $E[A] = p(x)$

$\therefore \log(p(x)) \geq E[\log A]$, which shows

$\log A$ is not an unbiased estimator.

---

6)

a) Minimum bits to represent 50257 tokens

$2^n \geq 50257$

$2^{15} = 32768$, $2^{16} = 65536$

Thus minimal $n = 16$ bits

b) Increase in no. of parameters when expanding from 50257 to 60000

Original parameters $= 50257 \times 768$

New parameters $= 60000 \times 768$

Difference $= (60000 - 50257) \times 768 = 7482624$

Total difference = contribution from embedding + Fully connected layer

$= 7482624 \times 2 = 14974848$

---

4) Setup with example, where $n=2$

$P_f(x_1, x_2) = P_f(x_1) P_f(x_2 | x_1)$

- Marginal for $x_1 \rightarrow P_f(x_1) = N(x_1 | 0, 1)$
- conditional for x_2 given x_1:

$p_f(x_2|x_1) = \mathcal{N}(x_2 | \mu_2(x_1), \varepsilon),$

where $\mu_2(x_1) = \begin{cases} 0 & \text{if } x_1 <= 0 \\ 1 & \text{if } x_1 > 0 \end{cases}$

$$p_f(x_2) = \int p_f(x_1)\, p_f(x_2|x_1)\, dx_1$$

$\because$ $x_1 \sim \mathcal{N}(0,1)$:

→ Probability that $x_1 <= 0 \approx 0.5$

→ Probability that $x_1 > 0 \approx 0.5$

For $x_1 <= 0$,

$\mu_2(x_1) = 0$ & $p_f(x_1 <= 0) = 0.5$

$\therefore$ $p_f(x_2) = 0.5\,\mathcal{N}(x_2 | 0, \varepsilon)$

For $x_1 > 0$

$\mu_2(x_1) = 0$ & $p_f(x_1 > 0) = 0.5$

$\therefore$ $p_f(x_2) = 0.5\,\mathcal{N}(x_2 | 1, \varepsilon)$

The combined marginal:

$$p_f(x_2) \approx 0.5\,\mathcal{N}(x_2|0,\varepsilon) + 0.5\,\mathcal{N}(x_2|1,\varepsilon)$$

For the reverse autoregressive model,

$p_r(x_1, x_2) = p_r(x_1|x_2)\, p_r(x_2)$

$p_r(x_2) = \mathcal{N}(x_2 | \mu^2(0), \sigma^2(0))$

→ In the reverse process, x_2 being the first variable has no previous variable to condition on.

→ $p_r(x_2)$ is hence just defined by a single gaussian & is always Unimodal, defined by a single peak.

→ This is in contrast to $p_f(x_2)$ which is a bimodal distribution being a mixture 2 Gaussian with no overlap

→ Since $p_f(x_2) \neq p_r(x_2)$, the foward & reverse auto regressive models do not cover the same hypothesis space.