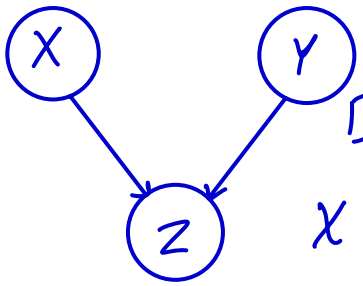
Q1

A) 1) $I(X;Y|Z) < I(X;Y)$

Let X and Y be random variables such that $I(X;Y)>0$. And X and Y deterministic functions of Z, therefore we have $I(X;Y|Z) = 0$.

2) $I(X;Y|Z) > I(X;Y)$

consider the following bayesian network. X and Y are independent, giving $I(X;Y) = 0$.



But given Z, we have either X or Y. If X doesn't happen, the Y must happend and vice versa. Hence ther is a conditional dependance between X and Y, given Z. Therefore $I(X;Y|Z) > 0 = I(X;Y)$

B) 1) $H(X,Y,Z) - H(X,Y) \leq H(X,Z) - H(X)$

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

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

$\Rightarrow$ LHS := $H(Z|X,Y)$

RHS := $H(Z|X)$

conditioning reduces entropy:

$$\mathcal{H}(z|X,Y) \leq \mathcal{H}(z|X)$$

Equality holds when $z|X$ is independent of Y.

2) $I(X;z|Y) \geq I(z;Y|X) - I(z;Y) + I(X;z)$

$$I(z;Y|X) - I(z;Y) + I(X;z)$$
$$= \mathcal{H}(z|X) - \mathcal{H}(z|Y,X) - \mathcal{H}(z) + \mathcal{H}(z|Y)$$
$$+ \mathcal{H}(z) - \mathcal{H}(z|X)$$
$$= \mathcal{H}(z|Y) - \mathcal{H}(z|X,Y)$$
$$= I(X,z|Y)$$

C) 1) First, we have to find the distribution of Z. Let Y be a uniform discrete random variable with values $\{4,5,6\}$. Thus

$$P(Z=A) = \sum_y P(Y=y)\,P(Z=A|Y=y)$$

$$= \frac{1}{3}\left(\frac{1}{1+e^{-0.5\times(-1)}} + \frac{1}{1+e^{-0.5\times 0}} + \frac{1}{1+e^{-0.5\times 1}}\right)$$

$$= 1/2$$

$$P(Z=B) = \sum_y P(Y=y)\,P(Z=B|Y=y) = \sum_y P(Y=y)\frac{1-P(Z=A|Y=y)}{2}$$

$$= \frac{1-P(Z=A)}{2}$$

$$= \frac{1 - \frac{1}{2}}{2} = \frac{1}{4}$$

$$P(Z = C) = \frac{1 - \frac{1}{2}}{6} = \frac{1}{12}$$

$$P(Z = D) = \frac{1 - \frac{1}{2}}{3} = \frac{1}{6}$$

$$I(Z;Y) = H(Z) - H(Z|Y)$$

$$H(Z) = -\sum_{z \in \{A,B,C,D\}} P(Z = z) \log P(Z = z) = 1.73$$

$$H(Z|Y) = -\sum_{y} P(Y = y) H(Z|Y = z)$$

$$= 1.70$$

$$\Rightarrow I(Z;Y) = 0.03$$

$I(Z;Y)$ is a measure of how much information we would have about Z, given we know Y. i.e. how much information does the reward give us about the action taken. If $I(Z;Y)$ is high, it means they are highly related an dependent and vice versa.

D) $D_{KL}(p(x,y,z) \| p(x)p(y)p(z)) = -\mathcal{H}(X,Y,Z) + \mathcal{H}(X) + \mathcal{H}(Y) + \mathcal{H}(Z)$

$$D_{KL}(p(x,y,z)\|p(x)p(y)p(z)) = \sum p(x,y,z) \log \frac{p(x,y,z)}{p(x)p(y)p(z)}$$

$$= \sum p(x,y,z)\log p(x,y,z)$$

$$- \sum p(x,y,z)\log[p(x)p(y)p(z)]$$

$$= -\mathcal{H}(X,Y,Z) - \sum p(x,y,z)\log p(x)$$

$$- \sum p(x,y,z)\log p(y) - \sum p(x,y,z)\log p(z)$$

$$= -\mathcal{H}(X,Y,Z) + \mathcal{H}(X) + \mathcal{H}(Y) + \mathcal{H}(Z)$$

E) $I(X1;X3) + I(X2;X4) \leq I(X1;X4) + I(X2;X3)$

Data processing inequality: $X \to Y \to Z \Rightarrow I(X;Z) \leq I(X,Y)$

$I(X_1;X_3) + I(X_2;X_4) = \mathcal{H}(X_1) - \mathcal{H}(X_1|X_3) + \mathcal{H}(X_2) - \mathcal{H}(X_2|X_4)$

$I(X_1;X_4) + I(X_2;X_3) = \mathcal{H}(X_1) - \mathcal{H}(X_1|X_4) + \mathcal{H}(X_2) - \mathcal{H}(X_2|X_3)$

So we have to prove:

$$\mathcal{H}(X_1|X_3) + \mathcal{H}(X_2|X_4) \geq \mathcal{H}(X_1|X_4) + \mathcal{H}(X_2|X_3)$$

$\mathcal{H}(X_1,X_2|X_3) = \mathcal{H}(X_1|X_3) + \mathcal{H}(X_2|X_1,X_3)$

$\Rightarrow \mathcal{H}(X_1|X_3) = \mathcal{H}(X_1,X_2|X_3) - \mathcal{H}(X_2|X_1,X_3)$

$\mathcal{H}(X_2|X_4) = \mathcal{H}(X_2,X_1|X_4) - \mathcal{H}(X_1|X_2,X_4)$

$$H(X_1|X_4) = H(X_1, X_2|X_4) - H(X_2|X_1, X_4)$$

$$H(X_2|X_3) = H(X_2, X_1|X_3) - H(X_1|X_2, X_3)$$

$$H(X_1, X_2|X_3) - H(X_2|X_1, X_3) + H(X_2, X_1|X_4) - H(X_1|X_2, X_4)$$

$$H(X_1, X_2|X_4) - H(X_2|X_1, X_4) + H(X_2, X_1|X_3) - H(X_1|X_2, X_3)$$

We need to show:

$$H(X_2|X_1, X_3) \leq H(X_2|X_1, X_4)$$

We know:

$$H(X_2|X_1, X_3) = H(X_2|X_1, X_3|X_4) \quad \text{(Markov property)}$$

And conditioning reduces entropy, thus:

$$H(X_2|X_1, X_4) \geq H(X_2|X_1, X_4, X_3)$$

So we are done.

Q2

A) By first order condition, f is convex iff

$\text{dom } f$ is convex , $f(y) \geq f(x) + \nabla f(x)^T (y-x)$

Let for some $x_0 \in \text{dom } f$; $\nabla f(x_0) = 0$. Then by FOC we have :

$\forall y \in \text{dom } f$; $f(y) \geq f(x_0) + \nabla f(x)^T (y-x) = f(x_0)$. Thus x_0 is a minimum point.

B)

$$\min \quad x^2 + y^2 + z^2$$

$$\text{s.t.} \quad z^2 = x^2 + y^2$$

$$2 = x + y + 1$$

$$z^2 = x^2 + y^2 \Rightarrow x^2 + y^2 + 1 + 2xy + 2x + 2y = x^2 + y^2$$

$$\Rightarrow y = -\frac{x + 1/2}{x+1}$$

$$\Rightarrow \frac{1}{2(x+1)^2} (4x^4 + 8x^3 + 8x^2 + 4x + 1)$$

$$\xRightarrow{d/dx} \frac{1}{(x+1)^3} (4x^4 + 12x^3 + 12x^2 + 6x + 1) = 0$$

$\Rightarrow x = 0, -1 \pm \frac{\sqrt{2}}{2}, \cancel{-\frac{1}{2} \pm \frac{i}{2}}$

$\Rightarrow \min = 6 - 4\sqrt{2}$ for $x = \frac{\sqrt{2}}{2} - 1$

$\max = +\infty$

C) prove $\mathcal{L}(\mu, \lambda) \leq \min_x f(x)$

primal :

$$\min \quad f(x)$$
$$\text{s.t.} \quad g_i(x) \leq 0 \quad i = 1, \dots, m$$
$$h_i(x) = 0 \quad i = 1, \dots, p$$

dual :

$$\max \quad \mathcal{L}(\mu, \lambda)$$
$$\text{s.t} \quad \lambda \geq 0$$

$$\mathcal{L}(\mu, \lambda) = \inf_x \mathcal{L}(x, \mu, \lambda)$$

$$\mathcal{L}(x, \mu, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{p} \mu_i h_i(x)$$

consider any feasible point in the primal problem and the dual problem.

e.g: x_i $g_i(x) \leq 0 \quad \forall i = 1, \dots, m$ and $h_i(x) = 0 \quad \forall i = 1, \dots, p$

(λ, μ) : $\lambda \geq 0$

$$\Rightarrow \mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{p} \mu_i h_i(x)$$
$$\leq f(x)$$

$$\Rightarrow L(\lambda, \mu) = \inf_x \mathcal{L}(x, \lambda, \mu) \leq f(x) \qquad \forall x \text{ in dom primal}$$

$$\Rightarrow L(\lambda, \mu) \leq \min_x f(x)$$

Consider the primal problem

$$\begin{aligned} \min \quad & e^{-x} \\ \text{s.t} \quad & x^2/y \leq 0 \end{aligned}$$

$$D = \{(x, y),\ y > 0\}$$

thus $p^* = 1$. The lagrangin is $L(x, y, \lambda) = e^{-x} + \lambda x^2/y$ thus the dual function is:

$$g(\lambda) = \inf_{x,y} (e^{-x} + \lambda x^2/y) = 0$$

so we have

$$\begin{aligned} d^* = \max \quad & 0 \\ \text{s.t} \quad & \lambda \geq 0. \end{aligned}$$

Thus $d^* = 0$. so $p^* - d^* = 1$ and strong duality does not hold

D)

$$\min \quad -\sum \log(\alpha_i + x_i)$$

$$\text{s.t.} \quad x \succeq 0$$

$$1^T x = 1$$

$$\mathcal{L}(x, \lambda, v) = -\sum \log(\alpha_i + x_i) - \lambda^T x + v 1^T x - v$$

to solve the problem, we find the variables such that they satisfy KKT conditions.

$$x \succeq 0, \quad 1^T x = 1, \quad \lambda \succeq 0, \quad \lambda_i x_i = 0 \quad i = 1, \dots, n$$

$$\frac{\partial}{\partial x_i} \mathcal{L}(x, \lambda, v) = -\frac{1}{\alpha_i + x_i} - \lambda_i + v = 0$$

$$\Rightarrow \lambda_i = v - \frac{1}{\alpha_i + x_i} \quad \Rightarrow x_i \left(v - \frac{1}{\alpha_i + x_i} \right) = 0$$

$$, \quad v \geq \frac{1}{\alpha_i + x_i}$$

$$\Rightarrow \text{if} \quad v < \frac{1}{\alpha_i} \quad \Rightarrow x_i > 0 \Rightarrow \lambda_i = 0$$

$$\Rightarrow v = \frac{1}{\alpha_i + x_i} \quad \Rightarrow x_i = \frac{1}{v} - \alpha_i$$

if $v \geq \frac{1}{\alpha_i}$, then if $x > 0$, we have $v > \frac{1}{\alpha_i + x_i}$. But by the complementry slackness we have $x_i (v - \frac{1}{\alpha_i + x_i}) = 0$, which is contradiction. Therfore we have

$$x_i = \begin{cases} \frac{1}{v} - \alpha_i & , \text{ if } v < \frac{1}{\alpha_i} \\ 0 & , \text{ o.w.} \end{cases} \quad \text{or} \quad \max\{0, \frac{1}{v} - \alpha_i\}$$

and we have $\sum \max\{0, \frac{1}{v} - \alpha_i\} = 1$, which gives the value of v.

E)
$$E[X] = \int_{-\infty}^{+\infty} x\, P(x)\, dx = \mu$$

$$Var(X) = E[(X-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 P(x)\, dx = \sigma^2$$

$$F(-\infty \le X \le +\infty) = \int_{-\infty}^{+\infty} P(x)\, dx = 1$$

Since we are integrating over all of the real numbers, the mean of the distribution has no effect on the differential entropy. Thus we only take th variance and the probabilty sum constraints into account.

$$\text{minimize} \quad \int_{-\infty}^{+\infty} P(x) \log P(x)\, dx$$

$$\text{subject to} \quad \int_{-\infty}^{+\infty} (x-\mu)^2 P(x)\, dx = \sigma^2$$

$$\int_{-\infty}^{+\infty} P(x)\, dx = 1$$

$$\Rightarrow \mathcal{L}(P(x), v_1, v_2) = -\int_{-\infty}^{+\infty} P(x)\log P(x)\,dx + v_1\left(1 - \int_{-\infty}^{+\infty} P(x)\,dx\right)$$
$$+ v_2\left(\sigma^2 - \int_{-\infty}^{+\infty} (x-\mu)^2 P(x)\,dx\right)$$

By the KKT conditions, for the optimal y, the small variation of $\mathcal{L}$, i.e. $\delta\mathcal{L}$ must be zero.

$$\delta\mathcal{L} = \int_{-\infty}^{+\infty} \delta P(x)\left(-\frac{\partial P(x)\log P(x)}{\partial P(x)} + v_1\frac{\partial P(x)}{\partial P(x)} + v_2(x-\mu)^2\frac{\partial P(x)}{\partial P(x)}\right)$$
$$= \int_{-\infty}^{+\infty} \delta P(x)\left(-\log P(x) - 1 + v_1 + v_2(x-\mu)^2\right) = 0$$

By the Euler-Lagrange equation (or since the above equality has to hold for any small $\delta P(x)$), regardless of the sign):

$$-\log P(x) - 1 + v_1 + v_2(x-\mu)^2 = 0 \Longrightarrow P(x) = e^{-v_1 - 1 - v_2(x-\mu)^2}$$

using the constraints we get:

$$\int_{-\infty}^{+\infty} e^{-v_1 - 1 - v_2(x-\mu)^2}\,dx = e^{-v_1-1}\left[\frac{\sqrt{\pi}}{2\sqrt{v_2}}\operatorname{erf}\left(\sqrt{v_2}(x-\mu)^2\right)\right]_{-\infty}^{+\infty}$$
$$= e^{-v_1-1} \times \frac{\sqrt{\pi}}{2\sqrt{v_2}} \times 2 = 1$$
$$e^{-v_1-1} = \frac{\sqrt{v_2}}{\sqrt{\pi}}$$

$$\frac{\sqrt{v_2}}{\sqrt{\pi}} \int_{-\infty}^{+\infty} (x-\mu)^2 e^{-v_2 (x-\mu)^2} dx = \sigma^2 = \left(\frac{(x-\mu) e^{-v_2 (x-\mu)^2}}{2a} \Big|_{-\infty}^{+\infty} - \frac{\sqrt{\pi}}{4 v_2^{3/2}} \operatorname{erf}\left(\sqrt{v_2} (x-\mu)^2\right) \Big|_{-\infty}^{+\infty} \right) \frac{\sqrt{v_2}}{\sqrt{\pi}}$$

$$= \frac{\sqrt{\pi}}{4\sqrt{v_2}} \times 2 \times \frac{\sqrt{v_2}}{\sqrt{\pi}} = \sigma^2$$

$$\Rightarrow \frac{1}{2v_2} = \sigma^2 \Rightarrow v_2 = \frac{1}{2\sigma^2}$$

$$\Rightarrow p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The objective functional and the constraints of the problem are all convex, hence the KKT conditions give the optimal solution, sine we have strong duality.

Q3

A) To prove the stationary distribution of the given Markov chain is μ, we must show that if the transition probability matrix is P, then $P\mu = \mu$. Now we show that the SD is a uniform distribution.

For an arbitrary node x we have

$$\mu(x) \times \left(1 - \frac{|N(x)|}{M}\right) + \sum_{y \in N(x)} \mu(y) \frac{1}{M} = \mu(x) - \frac{\mu(x)\,N(x)}{M} + \frac{\sum_{y \in N(x)} \mu(y)}{M}$$

$$= \mu(x)$$

Now consider the node with the highest probability. It's probability is the average of the probability of the nodes connected to it. Therefore the probabilities of this node and it's neighbours are equal. We apply the same argument to the neighbours of the so called node. Since the Markov chain is irreducible, it is connected. Hene the probability of all node is equal. Therefor the stationary distribution is uniform.

(if the graph is directed, the in-degree and out-degree of each node has to be equal, and the we can reason as above for this case.)

B) The problem with this method is that each valid vector in the sample contians more than one valid state in itself. But we count only on of them. (The same problem arises with FVMC algorithm in estimating Q values.) Therefore this method has high variance.

C) 1) In the described Markov process, the trasition is only possible adjacent valid state. since we have $\sum_{i=1}^{n} a_i < b$, every state is valid. Thus this Markov chain is irreducible. Therefore by part (A), this Markov chain has a uniform stationary disribution.

2) The trasition probability of going to each valid state from one is $\frac{1}{n}$. For a given state x we have:

$$P_{x,y} = \begin{cases} \frac{1}{n} & y \neq x,\ y \in N(x) \\ 1 - \frac{|N(x)|}{n} & y = x \\ 0 & y \neq x,\ y \notin N(x) \end{cases}$$

where $N(x)$ is the set of states which are reachable from x. If we reduce the whole process to the valid states, the the Markove process is irreducible, thus by part A, it has a uniform distribution. Thus if we run the process for infinite steps, the probability of each state would be equal, and so we should visit all valid states. so we only need to count the newly visited states.

Q4

A) s is the reward. If the images are the same, the higher the value of s, the higher it's probability (given that $\lambda_s > 0$). And if the images are different, it's vice versa. Which it seems sensible. ($\lambda_d > 0$) The α_s and α_d coefficients are for normalizations, so the given functions are probabilistic.

B) We would like to compute $P(same_j | s_1, \dots, s_n)$, which by the bayes rule we have:

$$P(same_j | s_1, \dots, s_n) = \frac{P(s_1, \dots, s_n | same_j) P(same_j)}{P(s_1, \dots, s_n)}$$

$$= \frac{P(s_1, \dots, s_n | same_j)}{\sum_i P(s_1, \dots, s_n | same_i) P(same_i)} \times P(same_j)$$

since the prior for all the images is equal, then $\forall 1 \le i \le n \; P(same_i) = 1/n$ and $P(different_j) = \frac{n-1}{n}$. Thus we have:

$$P(same_j | s_1, \dots, s_n) = \frac{P(s_1, \dots, s_n | same_j)}{\sum_i P(s_1, \dots, s_n | same_i)}$$

$$P(s_1, \dots, s_n | same_i) = \prod_k P(s_k | same_i) = P(s_i | same_i) \prod_{k \neq i} P(s_k | different_k)$$

$$= \alpha_{ss} \times \alpha_{ds}^{n-1} e^{\lambda_s s_i} \prod_{k \neq i} e^{-\lambda_d s_k}$$

$$\Rightarrow P(same_j | s_1, \dots, s_n) = \frac{e^{(\lambda_s + \lambda_d) s_j}}{\sum_i e^{(\lambda_s + \lambda_d) s_i}}$$

C) We want the probability of the case which the maximum reward is for a correct recognition. In other words, if M_s and M_d are the maximum values

of the scores for the same and different cases, we want to compute $P(M_s > M_d)$. By the equality of priors, we expect that N/s images belong to the "same" category, and $N - N/s$ to the other one. So if we suppose that the scores for each category are sorted, by the order stats, if $X_1 \leq \dots \leq X_K$ are K i.i.d random variables with CDF of F_X, then

$$F_{X_r}(x) = \sum_{j=r}^{K} \binom{K}{j} F_X(x)^j (1 - F_X(x))^{n-j}$$

$$f_{X_r}(x) = \frac{K!}{(r-1)!(K-r)!} f_X(x) F_X(x)^{r-1} (1 - F_X(x))^{n-r}$$

Therefore we have

$$F_D \doteq F_{d_{(N-N/s)}} \quad \text{and} \quad F_D(x) = F_{diff}(x)^{N - N/n}$$

$$f_D(x) = (N - N/n) f_D(x) F_{diff}(x)^{N - N/n - 1}, \quad f_d(x) \doteq P(x \mid \text{different})$$

$$F_S \doteq F_{s_{N/n}} \quad \text{and} \quad F_D(x) = F_{same}(x)^{N/n}$$

$$f_S(x) = N/n \, f_S(x) F_{same}(x)^{N/n - 1}, \quad f_S(x) \doteq P(x \mid \text{same})$$

Now we need to compute $P(M_S \geq M_d)$, which is

$$F_D(M_d)\left(1 - \int_0^{M_d} f_S(x)\, dx\right)$$

Q5

A) $\theta = (\mu, \sigma^2)$, $p_\theta(x) = \mathcal{N}(x; \mu, \sigma^2)$

Let the sequence of samples be $(x_1, x_2, \ldots, x_n)$. Then and x_is ar i.i.d.

$$p_\theta(X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

For simplicity, we use the log-likelihood fontion.

$$L_\theta(X) = \log p_\theta(X) = \frac{-n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

To find the MLE, we derivate L_θ w.r.t θ and set it to zero.

$\theta = (\mu, \sigma^2)$

$$\frac{\partial}{\partial\mu} L_\theta(X) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\Longrightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{\partial}{\partial\sigma^2} L_\theta(X) = \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\Longrightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

For $\hat{\mu}$ we have:

$$E[\hat{\mu}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Thus MLE for μ is unbiased.

For $\hat{\sigma}^2$ we have:

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}((x_i-\mu)-(\hat{\mu}-\mu))^2\right]$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n}((x_i-\mu)^2 - 2(x_i-\mu)(\hat{\mu}-\mu) + (\hat{\mu}-\mu)^2)\right]$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{2}{n}(\hat{\mu}-\mu)\sum_{i=1}^{n}(x_i-\mu) + (\hat{\mu}-\mu)^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[(x_i-\mu)^2] - E[2(\hat{\mu}-\mu)^2] + E[(\hat{\mu}-\mu)^2]$$

$$= \sigma^2 - E[(\hat{\mu}-\mu)^2] = (1-\frac{1}{n})\sigma^2 < \sigma^2$$

$$E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right)^2\right] = E\left[\frac{1}{n^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] = \frac{1}{n^2}\sum_{i=1}^{n} E[(x_i-\mu)^2] = \frac{\sigma^2}{n}$$

Thus MLE for variance is biased.

B) Let L be the likelihood function. To prove the invariance of MLE, we first define the induced likelihood function. For the function $\tau(\theta)$, the induced likelihood function L^* is defined as

$$L^*(\eta|X) = \sup_{\theta:\tau(\theta)=\eta} L(\theta|X)$$

and by definition, the value $\hat{\eta}$ that maximizes L^* is the

MLE for $\eta = \tau(\theta)$. We need to prove $L^*(\hat{\eta}|X) = L^*(\tau(\hat{\theta})|X)$.

we have:

$$L^*(\hat{\eta}|X) = \sup_{\eta} L^*(\eta|X) = \sup_{\eta} \sup_{\theta:\tau(\theta)=\eta} L(\theta|X)$$

$$= \sup_{\theta} L(\theta|X)$$

$$= L(\hat{\theta}|X)$$

And

$$L(\hat{\theta}|X) = \sup_{\theta;\tau(\theta)=\tau(\hat{\theta})} L(\hat{\theta}) = L^*(\tau(\hat{\theta})|X)$$

thus

$$L^*(\hat{\eta}|X) = L^*(\tau(\hat{\theta})|X)$$

An we have:

$$\tau_1(\mu) = \frac{1}{\mu^2+1} \Longrightarrow \tau(\hat{\mu}) = \frac{1}{\frac{1}{n^2}\sum x_i + 1} = \frac{n^2}{\sum x_i + n^2}$$

$$\tau_2(\sigma^2) = \sqrt{\sigma^2} \Longrightarrow \tau(\hat{\sigma}^2) = \sqrt{\frac{1}{n}\sum (x_i - \hat{\mu})^2}$$

c) 1) Consider the estimator $\hat{\mu}_{IS} = \frac{1}{n}\sum_{i=1}^{n} w_i f(x_i)$. So

$$E_q[\hat{\mu}] = E_q\left[\frac{1}{n}\sum_{i=1}^{n} w(x_i) f(x_i)\right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} E_q[w(x_i) f(x_i)]$$

For the random variable x we have

$$E_q[w(x) f(x)] = \int w(x) f(x) q(x) \, dx$$

$$= \int \frac{p(x)}{q(x)} q(x) f(x) \, dx$$

$$= \int p(x) f(x) \, dx$$

$$= E_p[f(x)]$$

Therefore

$$E_q[\hat{\mu}_{IS}] = \frac{1}{n} \sum_{i=1}^{n} E_q[w(x_i) f(x_i)] = E_p[f(x)]$$

And thus th IS estimator is unbiased.

2) For the variance we hav

$$Var(\hat{\mu}_{IS}) = Var\left(\frac{1}{n} \sum_{i=1}^{n} w(x_i) f(x_i)\right)$$

$$= \frac{1}{n} Var(w(X) f(X)) \qquad x_i\text{s are i.i.d}$$

$$= \frac{1}{n} \left(\int \frac{f(x)^2 p(x)^2}{q(x)} \, dx - \mu^2 \right)$$

Let $\bar{q}(x) = \frac{|f(x)| p(x)}{c}$, where c is a constant making $\bar{q}$ a distribution

i.e. $c = \int |f(x)| p(x)$. Then for any arbitrarcy distribution we have:

$$\sigma_{\bar{q}}^2 = \int \frac{f(x)^2 p(x)^2}{\bar{q}(x)} \, dx - \mu^2$$

$$= c \int |f(x)| p(x) \, dx - \mu^2$$

$$= \left(\int |f(x)| p(x)\, dx\right)^2 - \mu^2$$

$$= \left(\int \frac{|f(x)| p(x)}{q(x)}\, q(x)\, dx\right)^2 - \mu^2$$

$$\leq \int \frac{f^2(x)\, p^2(x)}{q^2(x)}\, q(x)\, dx - \mu^2 \qquad \text{(Cauchy-Schwarz)}$$

$$= \sigma_q^2$$

If the function f is non-negative, then the variance whould be zero, otherwise it has no certain value.

Now let $p(x)$ be a distribution concentrated in region A of the domain and $q(x)$ have almost zero probability there. Now let $f(x)$ have high values at A. Then the variance of $\hat{\mu}_{IS}$ can be arbitrarily large.

3) $$\hat{\mu}_{N\text{-}IS} = \frac{\sum_{i=1}^{n} w_i f(x_i)}{\sum_{i=1}^{n} w_i}$$

Let $C = \sum_{i=1}^{n} w(x_i)$

$$\Rightarrow E_q[\hat{\mu}_{N\text{-}IS}] = E_q\left[\frac{\frac{1}{n}\sum_{i=1}^{n} w(x_i) f(x_i)}{\frac{1}{n}\sum_{i=1}^{n} w(x_i)}\right] \neq \frac{E_p[f(X)]}{E_q[w(X)]} = E_p[f(X)] = \mu$$

Thus $\hat{\mu}_{N\text{-}IS}$ is biased

$$E_q[w(X)] = \int \frac{p(x)}{q(x)}\, q(x)\, dx = 1$$

But by the strong law of large numbers we have:

$$\lim_{n\to\infty} \hat{\mu}_{N\text{-}IS} = \lim_{n\to\infty} \frac{\frac{1}{n}\sum_{i=1}^{n} w(x_i) f(x_i)}{\frac{1}{n}\sum_{i=1}^{n} w(x_i)} = \frac{E_q[w(X)f(X)]}{E[w(X)]}$$

$$= \mu$$

4) We have $\hat{\mu}_{N\text{-}IS} = \frac{\sum w_i f(x_i)}{\sum w_i}$. so it is a weighted average of $f(x)$. Let's denote $\hat{\mu}_{N\text{-}IS}$ by $\hat{\mu}$. Then $m \le \hat{\mu} \le M$. Now let's define $g(t)$ as:

$$g(t) = E[(\hat{\mu} - t)^2]$$

$$\Rightarrow g'(t) = \frac{d}{dt} E[(\hat{\mu} - t)^2) = E[\frac{d}{dt}(\hat{\mu} - t)^2]$$

$$= E[-2\hat{\mu} + 2t] = 2E[t] - 2E[\hat{\mu}]$$

$$= 2t - 2E[\hat{\mu}] = 0 \Rightarrow t = E[\hat{\mu}]$$

And $g''(t) = 2 > 0 \Rightarrow g$ is convex $\Rightarrow E[\hat{\mu}] = \arg\min_t g(t)$.

Therfore for any point, e.g $t = \frac{M+m}{2}$ we have

$$var(\hat{\mu}) = g(E[\hat{\mu}]) \le g(\frac{M+m}{2})$$

Now we have:

$$g(\frac{M+m}{2}) = E[(\hat{\mu} - \frac{M+m}{2})] = \frac{1}{4} E[((\hat{\mu} - M) + (\hat{\mu} - m))^2]$$

And $m \le \hat{\mu} \le M$, thus $\hat{\mu} - M \le 0$ and $\hat{\mu} - m \ge 0$, so

$$((\hat{\mu} - M) + (\hat{\mu} - m))^2 \le ((\hat{\mu} - m) - (\hat{\mu} - M))^2 = (M - m)^2$$

And since E is linear, then

$$\text{Var}(\hat{\mu}_{N\text{-}IS}) \le g\left(\frac{M+m}{2}\right) = \frac{1}{4} E\left[\left((\hat{\mu}-M)+(\hat{\mu}-m)\right)^2\right] \le \frac{(M-m)^2}{4}$$

5) 1. $\int_2^{\infty} e^{-\frac{x^2}{2}} dx \longleftrightarrow \int_0^{\frac{1}{2}} \frac{1}{t^2} e^{-\frac{1}{2t^2}} dt$

Now we have to filter the samples through the interval $[0, \frac{1}{2}]$.

$S_{filtered} = \{0.26, 0.41, 0.40, 0.10\}$

$$H = \left(\tfrac{1}{2} - 0\right) \sum_{t \in S_{filtered}} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} \times \frac{1}{t^2} e^{-\frac{1}{2t^2}} = 0.108$$

2. Now we apply IS. Take the sample from $q = N(3,1)$, by just shifting the S sample by 3. Then the shifet filtered sample is:

$S' = \{3.26, 2.73, 3.41, 3.4, 2.98, 3.1, 2.72\}$

$$\Rightarrow H = \sum_{x \in S'} \frac{\frac{1}{|S'|}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-3)^2}{2}}} = 0.03$$

The actual value of H is 0.05. Thus this estimator is better. It is better, since it gives more weigh to the samples in the interval and reduces the probailty of outlier samples.

Q6

A) $-D_{KL}(p\|q) = -\int p(x) \log \frac{p(x)}{q(x)} dx$

$= \int p(x) \log \frac{q(x)}{p(x)} dx$

since log is a concave function, by the Jensen's inequality we have $\leq \log \int p(x) \frac{q(x)}{p(x)} dx$

$= \log \int q(x) dx$

$= \log 1$

$= 0 \quad \Longrightarrow D_{KL}(p\|q) \geq 0$

Again by the Jensen, the equalit holds iff $\frac{p(x)}{q(x)}$ is constant, and since $\int p(x)dx = \int q(x) dx = 1$, the ratio is 1. Thus $D_{KL}(p\|q) = 0 \Longleftrightarrow p = q$.

B) $D_{KL}(q(z)\|P(z|x)) = E_q[\log \frac{q(z)}{P(z|x)}]$

$= E_q[\log q(z) - \log P(z|x)]$

$= E_q[\log q(z)] - E_q[\log P(z|x)]$

$= E_q[\log q(z)] - E_q[\log P(z,x)] + E[\log P(x)]$

$= \log P(x) - (E_q[\log P(x,z)] - E[\log q(z)])$

(I have problem with this. It has to be x, not X, i.e. a random variable)

C) We have $L(q) = E_q[\log P(x,z)] - E_q[\log q(z)]$

Let's denote $E_{q_j, j\neq i}$ by E_{-i}. We can write

$$q(z) = \prod_{i=1}^{n} q(z_i) \Longrightarrow \log q(z) = \sum_{i=1}^{n} \log q(z_i)$$

And we have

$$\begin{aligned}
L(q) &= \int q(z) \log P(X,z) dz - \int q(z) \log q(z) dz \\
&= \int \prod_{i=1}^{n} q(z_i) \log(X,z) dz - \int \prod_{i=1}^{n} q(z_i) \sum_{i=1}^{n} \log q(z_i) dz \\
&= \int \prod_{i=1}^{n} q(z_i) \left(\log(X,z) - \sum_{i=1}^{n} \log q(z_i) \right) dz \\
&= \int \left(q(z_i) \prod_{j\neq i} q(z_j) \right) \left(\log(X,z) - \left(\log q(z_i) + \sum_{j\neq i} \log q(z_j) \right) \right) dz \\
&= \int_{z_i} q(z_i) \int_{z_{j\neq i}} \prod_{j\neq i} q(z_j) \left(\log P(X,z) - \left(\log q(z_i) + \sum_{j\neq i} \log(q_i) \right) \right) d_{z_{j\neq i}} dz_i \\
&= \int_{z_i} q(z_i) \int_{z_{j\neq i}} \prod_{j\neq i} q(z_j) \log P(X,z) \, d_{z_{j\neq i}} dz_i \\
&\quad - \int_{z_i} q(z_i) \int_{z_{j\neq i}} \prod_{j\neq i} q(z_j) \left(\log q(z_i) + \sum_{j\neq i} \log(q_j) \right) d_{z_{j\neq i}} dz_i \\
&= \int_{z_i} q(z_i) E_{-i}[\log P(X,z)] dz_i - \int_{z_i} q(z_i) \log q(z_i) dz_i + C
\end{aligned}$$

where $C = \int_{z_{j\neq i}} \prod q(z_j) \sum_{j=} \log(q_j) \, dz_{j\neq i}$. The Lagrangian of this optimization problem is:

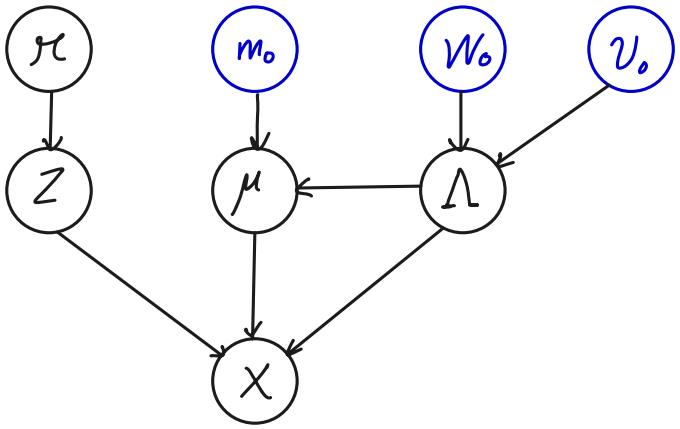
$$L(q) - \sum_{j=1}^{n} \lambda_j \int_{z_j} q(z_j) dz_j$$

$$\frac{\delta L(q)}{\delta q(z_i)} = \frac{\partial}{\partial q(z_i)} \left[q(\theta_i) \left[E_{-i}[\log P(X,z)] - \log q(z_i) \right] - \lambda_i q(z_i) \right]$$

$$= E_{-i}[\log P(X,Z)] - \log q(z_i) - 1 - \lambda_i = 0$$

$$\Longrightarrow \log q(z_i) = E_{-i}[\log P(X,Z) + \text{const} \qquad \forall\ 1 \le i \le n$$

D) The diagram of this GMM can be represented as a PGM like :



In this setting, the posterior we want to estimate is q, and the actual joint distribution is $P(X, Z, \pi, \mu, \Lambda)$. Therefore the ELBO of interest is

$$L(q) = E_q\left[\log \frac{P(X,Z,\pi,\mu,\Lambda)}{q(Z,\pi,\mu,\Lambda)}\right]$$

$$= E_q[\log P(X,Z,\pi,\mu,\Lambda) - \log q(Z,\pi,\mu,\Lambda)]$$

Now we have to compute $P(X,Z,\pi,\mu,\Lambda)$. By the represente PGM, which is a Bayes Net, we have

$$P(X, Z, \pi, \mu, \Lambda) = P(X|Z, \mu, \Lambda)\, P(Z|\pi)\, P(\pi)\, P(\mu, \Lambda)$$

Thus we have

$$L(q) = E_q[\log P(X|Z,\mu,\Lambda)] + E_q[\log P(Z|\pi)] + E_q[\log P(\pi)] + E_q[\log P(\mu,\Lambda)] - E_q[\log q(Z)] - E_q[\log q(\pi,\mu,\Lambda)]$$

(We have $q(Z,\pi,\mu,\Lambda) = q(Z)\, q(\pi,\mu,\Lambda)$)

by the previous part we have,

$$\log q^*(Z) = E_{q(\pi,\mu,\Lambda)}[\log P(X, Z, \pi, \mu, \Lambda)] + \text{const}$$

$$= E_{q(\pi,\mu,\Lambda)}[\log(P(X|Z,\mu,\Lambda) P(Z|\pi) P(\pi)\, P(\mu,\Lambda))] + \text{const}$$

$$= E_{q(\pi,\mu,\Lambda)}[\log P(X|Z,\mu,\Lambda)] + E_{q(\pi,\mu,\Lambda)}[\log P(Z|\pi)] + \underbrace{E_{q(\pi,\mu,\Lambda)}[\log P(\pi)] + E_{q(\pi,\mu,\Lambda)}[\log P(\mu,\Lambda)] + \text{const}}_{\text{const w.r.t } Z}$$

$$= E_{q(\pi,\mu,\Lambda)}[\log P(X|Z,\mu,\Lambda) + \log P(Z|\pi)] + \text{const}$$

$$E_{q(\pi)}[\log P(Z|\pi)] = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} E_{q(\pi)}[\log \pi_k]$$

$$E_{q(\mu,\Lambda)}[\log P(X|Z,\mu,\Lambda) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\, E_{q(\mu,\Lambda)}[\log \mathcal{N}(x_n|\mu_k,\Lambda^{-1})]$$

Thus:

$$\log q^*(Z) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \gamma_{nk} + \text{const}$$

where

$$\gamma_{nk} = E_{q(\pi)}[\log \pi_k] + \frac{1}{2} E_{q(\mu,\Lambda)}[\log |\Lambda_k|] - \frac{D}{2}\log(2\pi) - \frac{1}{2} E_{q(\mu,\Lambda)}[(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)]$$

(Because $E[\log \mathcal{N}(x_n | \mu_k, \Lambda_k^{-1})] = E[\frac{D}{2}\log(2\pi) + \frac{1}{2}\log|\Lambda_k| - \frac{1}{2}(x_k - \mu_k)^T \Lambda_k (x - \mu_k)]$

$= -\frac{D}{2}\log(2\pi) + \frac{1}{2} E[\log|\Lambda_k|]$

$- \frac{1}{2} E[(x_k - \mu_k)^T \Lambda_k (x_k - \mu_k)]$

Therefore $q^*(Z) = C \prod_{n=1}^{N} \prod_{k=1}^{K} e^{z_{nk}\gamma_{nk}}$ where C is a normalizing constant.

Now we need to compute $q^*(\pi)$, $q^*(\mu)$ and $q^*(\Lambda)$. Similar as before,

$$\log q^*(\pi) = E_{q(Z,\mu,\Lambda)}[\log P(X, Z, \pi, \mu, \Lambda)] + \text{const}$$

$$= E_{q(Z)}[\log P(Z|\pi) + \log P(\pi)] + \text{const} \quad \text{(Based on the Bayes Net)}$$

$$= \log P(\pi) + E_{q(Z)}[\log P(Z|\pi)] + \text{const}$$

$$= (\alpha_0 - 1)\log \sum_{k=1}^{K} \pi_k + C \sum_{n=1}^{N} \sum_{k=1}^{K} e^{\gamma_{nk}} \log \pi_k + \text{Const}$$

And

$$\log q^*(\mu, \Lambda) = E_{q(Z,\pi)}[\log P(X, Z, \pi, \mu, \Lambda)] + \text{const}$$

$$= E_{q(Z,\pi)}[\log P(X \mid Z, \mu, \Lambda) + \log P(\mu, \Lambda)] + \text{const}$$

$$= \log P(\mu, \Lambda) + E_{q(Z)}[\log P(X \mid Z, \mu, \Lambda)] + \text{const}$$

to normalize the proposed expression for each latent variable to form a distribution, since each vector is K-of-1, the sum of z_{nk}s over all values of k is 1, we have

$$q^*(Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}} \quad \text{where} \quad r_{nk} = \frac{e^{\gamma_{nk}}}{\sum_{i=1}^{K} e^{\gamma_{ni}}} = E[z_{nk}]$$

$$q^*(\pi) = (\alpha_0 - 1) \log \sum_{k=1}^{K} \pi_k + \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \log \pi_k + \text{const}$$

And

$$\log q^*(\mu, \Lambda) = \sum_{k=1}^{K} \log(\mathcal{N}(\mu_k \mid m_0, (\beta_0 \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k \mid W_0, \nu_0))$$

$$+ \sum_{n=1}^{N} \sum_{k=1}^{K} E_{q(Z)}[z_{nk}] \left(\frac{1}{2} \log|\Lambda_k| - \frac{D}{2} \log(2\pi) - (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)\right)$$

$$\Rightarrow \log q^*(\mu_k, \Lambda_k) = \log \mathcal{N}(\mu_k \mid m_0, (\beta_0 \Lambda_k)^{-1}) + \log \mathcal{W}(\Lambda_k \mid W_0, \nu_0)$$

$$+ \sum_{n=1}^{N} \left(\frac{1}{2} \log|\Lambda_k| - \frac{D}{2} \log 2\pi - \frac{D}{2}\right) E[z_{nk}]$$

$$= \log\left(\mathcal{N}(\mu_k \mid m_k, (\beta_k \Lambda_k)^{-1})\, \mathcal{W}(\Lambda_k \mid W_k, \upsilon_k)\right)$$

where

$$\beta_k = \beta_0 + N_k \quad , \quad N_k = \sum_{n=1}^{N} r_{nk}$$

$$m_k = \frac{\beta_0}{\beta_k} m_0 + \frac{N_k}{\beta_k} \bar{x}_k \quad , \quad \bar{x}_k = \frac{\sum_{n=1}^{N} r_{nk} x_n}{N_k}$$

$$W_k^{-1} = W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{x}_k - m_0)(\bar{x}_k - m_0)^T$$

$$\upsilon_k = \upsilon_0 + N_k$$

The analogy of EM algoritith to the process above is

1) The E step : computes the value of r_{nk} using the current values.

2) The M step : uses the new r_{nk} values to update the parameters.

I have used the Bishop's Pattern Recognition book for the update rules And deriving $q^*(\mu_k, \Lambda_k)$