Problem 1

(a). - for $D_{KL}(Q\|P) = \infty$, we can have $P(x) = 0$ where $Q(x) \neq 0$.

- for example we can have P to be uniform random variable on $[0\ 1]$ and Q be a uniform random variable on $[-1, 1]$.

- for $x \in [-1, 0)$, $Q(x)$ is divided by 0 hence $D_{KL}(Q\|P) = \infty$.

$$- D_{KL}(P\|Q) = \underset{x \sim P(x)}{E}\left[\log \frac{P(x)}{Q(x)}\right]$$

$$= \int_0^1 \log \frac{P(x)}{Q(x)}\, dx$$

$$= \int_0^1 \log \frac{1}{(1/2)}\, dx$$

$$= \log 2 \neq \infty$$

(b) Forward KL => $D_{KL}(P||Q) = \underset{x \sim P(x)}{E}\left[\log \frac{P(x)}{Q(x)}\right]$

Reverse KL => $D_{KL}(Q||P) = \underset{x \sim Q(x)}{E}\left[\log \frac{Q(x)}{P(x)}\right]$

For Plot A:-

- Q avoids being tiny where P is large, which is not very costly for forward KL.
- Q establishes "average" balance between the modes of P where P is small. This is costly in reverse KL
- Hence we can conclude that 'A' corresponds to minimizing forward KL

For Plot B:-

- where P is large, Q is extremely small, causing high forward KL
- where Q is large, P is also notably large, leading to not very costly reverse KL
- we can conclude that B minimizes reverse KL.

(c) For the ELBO

$$\mathcal{L}(x;\theta,\phi) = E_{q_\phi(z|x)}[\log p(x|z;\theta)] - D_{KL}(q_\phi(z|x) \| p(z))$$

- made up of two terms
  - reconstruction & regularization (KL)

- The regularization / KL term measures the divergence b/w the learned latent distribution $q(z|x)$ and prior distribution $p(z)$ which is typically gaussian.

- The assymmetry here means the VAE is more penalized when it propes a distribution $q_\phi(z|x)$ that places probability mass in regions where $p(z)$ does not. This means "penalizing" for proposing unlikely / unexpected latent space and is less concerned about "missing out" on regions prior considers likely

- Implications :- ① regularization pushes posterior to match prior, preventing overfitting & ensuring well structured / latent space

- Assymmetry ensures the encoder does not produce representations that are too unusual. even if it helps reconstruction

Problem 2

(a) Mixture of Bernoullis

(i) $x \in \{0,1\}^D \longrightarrow \begin{bmatrix} 0 \\ 1 \\ 1 \\ \vdots \end{bmatrix}$

$P(x_d = 1) = p_d \quad ; \quad p \in (0,1)^D \rightarrow \begin{bmatrix} P(x_0=1) \\ P(x_1=1) \\ \vdots \end{bmatrix}$

$$P(x|p) = \prod_{d=1}^{D} p_d^{x_d} (1-p_d)^{(1-x_d)}$$

(ii) mixture of K Bernoullis

$x^{(i)}$ is drawn from some vetor of Bernoulli random variables with parameters $p^{(k)}$ distributed $\pi(k)$

$x^{(i)}$ drawn from $\begin{bmatrix} x_0 \\ \vdots \\ x_n \end{bmatrix}$ — $\begin{bmatrix} x_0 \\ \vdots \\ x_n \end{bmatrix}$ mixture with distribution $\pi(k)$

Let A_k be an event of drawing/sampling $x = x^{(i)}$ from the mixture, Then,

$$P(x|p,\pi) = \sum P(x, A_k | p, \pi)$$
$$= \sum P(x | A_k, p, \pi) \times P(A_k | p, \pi)$$

(1) (2)

(1) prob of observing $x = x^{(i)}$ given it drawn from $p^k \rightarrow P(x|p^{(k)})$ because knowing x was sampled from p^k makes π irrelevant

② $P(A_k|p,\pi)$ – measures prob of sampling $x(i)$ nothing but $\pi(k)$

Hence

$$P(x^{(i)}|p,\pi) = P(x, A_k|p,\pi) = \sum_k \pi_k P(x|p^{(k)})$$

→ (iii) $X = \{x^{(i)}\}_{i=1\ldots n}$ → drawing from entire dataset

$x^{(i)}$ is individual observation in X

log likelihood is calculated by summing the log prob of each individual observation $x^{(i)}$ assuming the observations are independent

$$\log \mathcal{L}(p,\pi) = \sum_{i=1}^{N} \log P(x^{(i)}|p,\pi)$$

(b) Expectation step

(i) Let $A_k^{(i)}$ be the event that $x^{(i)}$ was drawn from $p^{(k)}$

$$P(z^{(i)}|\pi) = \prod_{k=1}^{K} \pi_k^{z_k^{(i)}}$$

$$P(x^{(i)}|z^{(i)}, p, \pi) = \prod_{k} P(x^{(i)}|z^{(i)}, p, \pi, A_k^{(i)})^{z_k^{(i)}}$$

$$= \prod_{k=1}^{K} \left[P(x^{(i)}|p^{(k)})\right]^{z_k^{(i)}}$$

(ii) Likelihood of data and latent variable

$$P(Z, X|\pi, p) = \prod_{i=1}^{N} P(x^{(i)}, z^{(i)}|\pi, p)$$

$$= \prod_{i=1}^{N} P(x^{(i)}|z^{(i)}, \pi, p)\, P(z^{(i)}|\pi)$$

$$= \prod_{i=1}^{N} \left[\prod_{k=1}^{K} \left[P(x^{(i)}|p^{(k)})\right]^{z_k^{(i)}} \right] \left[\prod_{k=1}^{K} \pi_k^{z_k^{(i)}} \right]$$

$$= \prod_{i=1}^{N} \left[\prod_{k=1}^{K} \left[\pi_k P(x^{(i)}|p^{(k)})\right]^{z_k^{(i)}} \right]$$

(c)

$$\eta(z_k^{(i)}) = E\left[z_k^{(i)} \,\middle|\, x^{(i)}, \pi, p\right]$$

$$= P\left[z_k^{(i)} = 1 \,\middle|\, x^{(i)}, \pi, p\right]$$

$$= \frac{P(x^{(i)} \mid z_k^{(i)} = 1, \pi, p)\, P(z_k^{(i)} = 1 \mid \pi, p)}{\sum_j \pi_j \prod_{d=1}^{D} (p_d^j)^{x_d^{(i)}} (1 - p_d^j)^{(1 - x_d^{(i)})}}$$

Computing log likelihoods:-

$$\log P(Z, D \mid \pi, p) = \sum_{i=1}^{N} \left[\sum_{k=1}^{K} z_k^{(i)} \left(\log P(x^{(i)} \mid p^{(k)}) + \log \pi_k \right) \right]$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} z_k^{(i)} \left[\log \pi_k + \sum_{d=1}^{D} \left(x_d^{(i)} \log p_d^k + (1 - x_d^{(i)}) \log(1 - p_d^{(k)}) \right) \right]$$

Taking the expected value & replacing

$E[z_k^{(i)}] = \eta(z_k^{(i)})$ gives the solution

(c) (i) setting the derivative of

$$\frac{d}{dp_d^{(k)}} E[\log P(z, D | \pi, p)] = 0.$$

$$\sum_{i=1}^{N} \eta(z_k^{(i)}) \left[\frac{x_d^{(i)}}{p_d^{(k)}} - \frac{1 - x_d^{(i)}}{1 - p_d^{(k)}} \right] = 0$$

$$\sum_{i=1}^{N} \eta(z_k^{(i)}) \left[x_d^{(i)} (1 - p_d^{(k)}) - (1 - x_d^{(i)}) p_d^{(k)} \right] = 0$$

$$\Rightarrow p_d^{(k)} = \frac{\sum_{i=1}^{N} \eta(z_k^{(i)}) x_d^{(i)}}{\sum_{i=1}^{N} \eta(z_k^{(i)})} = \sum_{i=1}^{N} \frac{\eta(z_k^{(i)}) x_d^{(i)}}{N_k}.$$

(ii) We need to only minimize $\sum_{i=1}^{N} \sum_{k=1}^{K} \eta(z_k^{(i)}) \log \pi_k$ since the rest are not a function of π

In order to keep π a distribution, we require $\sum_k \pi_k = 1$. Let λ be the dual variable for this constraint.

$$\mathcal{L}(\pi, \lambda) = -\sum_{i=1}^{N} \sum_{k=1}^{K} \eta(z_k^{(i)}) \log \pi_k + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

Taking the derivative w.r.t π_k :-

$$\frac{d}{d\pi_k} \mathcal{L}(\pi, \lambda) = -\sum_{i=1}^{N} \frac{\eta(z_k^{(i)})}{\pi_k} + \lambda = 0$$

$$\Rightarrow \pi_k = \frac{\sum_{i=1}^{N} \eta(z_k^{(i)})}{\lambda} = \frac{N_k}{\lambda}$$

Solving for λ

$$\mathcal{L}(\lambda) = -\sum_{i=1}^{N} \sum_{k=1}^{K} \eta(z_k^{(i)}) (\log N_k - \log \lambda) + \sum_{k=1}^{K} (N_k - \lambda)$$

taking derivative w.r.t λ

$$\frac{1}{\lambda} \sum_{i=1}^{N} \sum_{k=1}^{K} \left(\eta(z_k^{(i)}) - 1 \right) = 0.$$

$$\lambda = \sum_{i=1}^{N} \sum_{k=1}^{K} \eta(z_k^{(i)}) = \sum_{k=1}^{K} N_k.$$

Optimal $\boxed{\hat{\pi}_k = \frac{N_k}{\sum_{k=1}^{K} N_k}}$

Problem 3

→ To generate samples,

- we start by sampling a mixture component/cluster from the distribution mix_pi, means choosing a cluster 'k' with prob mix_pi[k]

- after choosing cluster, we generate a new image by sampling each pixel based on Bernoulli distribution parameterized by p[k].

- once all pixels are sampled, we reshape the flattened binary image data.

→ Plausibility

- largely depends on how well the EM algo has clustered the data & how accurately the Bernoulli params p represent each cluster's typical image.

For k=5; EM algo might group diverse digits into same cluster, leading to muddled representations shown in the image attached.

For k=20; more clusters → can better segregate different digits & variations of digits. This results in more recognizable digits. looking more like MNIST samples as shown in images attached.

Problem 4

→ NICE:-

→ volume-preserving transformations & invertible mappings, tries to learn a more structured latent space.

→ visualised samples do not show clear digit shapes, there are occasional unusual digit forms, as shown in the image.

→ VAE

- has ability to produce more smoother & continuous latent space, helpful in generative tasks.
- the images generated are more consistent with MNIST data compared to NICE
- VAE's regularize latent space using probabilistic framework, ensuring that samples generated from any point in latent space produces valid outputs as shown in images

→ GM generation quality is largely depends on the number of clusters. Given the nature of MNIST where the goal is to generate clear & recognizable digits, VAE is most suitable because of the ability to generate images from smooth & continuous latent space.

Problem 5

(a) Cauchy distribution:-

$$f_X(u) = \frac{1}{\pi(1+u^2)} \quad \text{for } -\infty < u < \infty$$

Let $Y = 1/X$,

Consider the CDF of Y, for $v > 0$.

$$F_Y(v) = P\{Y \le v\} = P\left\{\frac{1}{X} \le v\right\}$$

$$= P\left\{\frac{1}{v} \le X, X > 0\right\} + P\left\{\frac{1}{v} \ge X, X < 0\right\}$$

$$= P\left\{\frac{1}{v} \le X\right\} + P\{X < 0\}$$

$$= 1 - P\left\{\frac{1}{v} \ge X\right\} + P\{X < 0\}$$

$$= 1 - F_X\left(\frac{1}{v}\right) + F_X(0)$$

Taking the derivative w.r.t v.

$$f_Y(v) = \frac{d}{dv} F_Y(v) = -\frac{-1}{v^2} f_X\left(\frac{1}{v}\right)$$

$$= \frac{1}{v^2} \frac{1}{\pi\left(1+\left(\frac{1}{v}\right)^2\right)} = \frac{1}{\pi(1+v^2)}$$

Similarly for $v<0$,

$$F_Y(v) = P(Y \le v) = P\left\{\frac{1}{X} \le v\right\}$$

$$= P\left\{\frac{1}{v} \ge X, X>0\right\} + P\left\{\frac{1}{v} \le X, X<0\right\}$$

$$= 0 + P\left\{\frac{1}{v} \le X < 0\right\} = F_X(0) - F_X\left(\frac{1}{v}\right)$$

Taking the derivative w.r.t. v.

$$f_Y(v) = \frac{d}{dv} F_Y(v) = -\left(\frac{-1}{v^2}\right) f_X\left(\frac{1}{v}\right)$$

$$= \frac{1}{\pi(1+v^2)}$$

Therefore $Y = \frac{1}{X}$ also has cauchy distribution

(b) $Z = e^Y$ where Y is normally distributed with parameters μ & σ.

CDF of Z, in terms of CDF of Y

For $z > 0$,

$$P(Z \leq z) = P(e^Y \leq z) = P(Y \leq \ln(z))$$

This is the CDF of Y evaluated at $\ln(z)$, so:

$$f_Z(z) = F_Y(\ln(z))$$

PDF of Z by differentiating

$$f_Z(z) = \frac{d}{dz} F_Y(\ln(z))$$

$$= f_Y \ln(z) \left(\frac{1}{z}\right)$$

using the PDF of Y

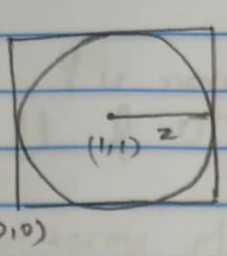
$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y-\mu)^2}{2\sigma^2}}$$

we get

$$f_Z(z) = \frac{1}{\sigma\sqrt{2\pi}\, z} e^{-\frac{(\ln(z)-\mu)^2}{2\sigma^2}}$$

↳ this is the density of e^Y where $z > 0$

(c)

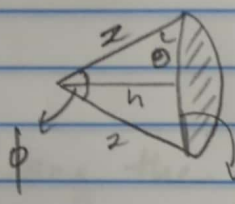


(i) for $0 \leq z \leq 1$, CDF $= \frac{1}{4}$ area of circle $= \frac{1}{4} \pi z^2$

$1/4 \rightarrow$ normalising factor.

(ii) for $1 \leq z \leq \sqrt{2}$

CDF $= \frac{1}{4}$ (area of outer circle − shaded region)



area of $\Delta^{le} = \frac{1}{2} \cdot b \cdot h$ | $\left(\frac{b}{2}\right)^2 + h^2 = z^2$

$= \frac{1}{2} \left(2\sqrt{z^2-1}\right) \cdot (1)$ | $b = 2\sqrt{z^2-h^2}$

$= \sqrt{z^2-1}$ | $b = 2\sqrt{z^2-1}$

area of the sector $= \frac{\phi}{2\pi} \cdot (\pi z^2)$ | $\tan\left(\frac{\phi}{2}\right) = \frac{b/2}{h}$

$= \frac{\phi z^2}{2}$ | $= \sqrt{z^2-1}$

$= z^2 \tan^{-1} \sqrt{z^2-1}$ | $\phi = 2\left(\tan^{-1}\sqrt{z^2-1}\right)$

∴ area of shaded region

$= \left(z^2 \tan^{-1} \sqrt{z^2-1}\right) - \left(\sqrt{z^2-1}\right)$

CDF :-

$$F_Z(z) = \frac{1}{4}\left(\pi z^2 - 4\left(z^2 \tan^{-1}\left(\sqrt{z^2-1}\right) - \sqrt{z^2-1}\right)\right)$$

Differentiating to find PDF :-

(i) for $0 \leq z \leq 1$

$$f_Z(z) = \frac{\pi z}{2}$$

for $1 < z \leq \sqrt{2}$

$$f_Z(z) = \frac{1}{4}\left[2\pi z - 4\left(2z \tan^{-1}\sqrt{z^2-1} + z^2 \cdot \frac{1}{1+\left(\sqrt{z^2-1}\right)^2} \cdot \frac{2z}{2\sqrt{z^2-1}} - \frac{2z}{2\sqrt{z^2-1}}\right)\right]$$

$$= \frac{1}{4}\left(2\pi z - 8z \tan^{-1}\sqrt{z^2-1}\right)$$

$$= \frac{\pi z}{2} - 2z \tan^{-1}\sqrt{z^2-1}$$

Problem 6. Jacobian of Leaky ReLU

Given $f(z) = [LReLU[z_1], LReLU[z_2] \dots LReLU[z_n]]^T$

The Jacobian matrix J is an n×n matrix where entry J_{ij} is $\frac{\partial f_i}{\partial z_j}$

For Leaky ReLU

$$LReLU[z_i] = \begin{cases} 0.1 z_i & \text{for } z_i < 0 \\ z_i & \text{for } z_i \geq 0 \end{cases}$$

So partial derivatives are:-

$$\frac{\partial LReLU(z_i)}{\partial z_i} = \begin{cases} 0.1 & \text{for } z_i < 0 \\ 1 & \text{for } z_i \geq 0. \end{cases}$$

The Jacobian matrix J is diagonal for elementwise operation because it depends only on z_i and not on other components of z

$$J = \text{diag}\left(\frac{\partial LReLU(z_1)}{\partial z_1}, \dots \frac{\partial LReLU(z_n)}{\partial z_n} \right)$$

Determinate of the Jacobian:-

$$\det(J) = \prod_{i=1}^{n} \frac{\partial LReLU[z_i]}{\partial z_i}$$

Inverse of the Jacobian

Inverse of diagonal matrix is also diagonal with inverse of each diag element

$$J^{-1} = \text{diag}\left(\frac{1}{\frac{\partial LReLU[z_1]}{\partial z_1}}, \dots, \frac{1}{\frac{\partial LReLU[z_n]}{\partial z_n}} \right)$$

Compute the inverse absolute determinate of Jacobian

Inverse absolute of det of Jacobian

$$\det(J^{-1}) = \frac{1}{\left| \prod_{i=1}^{n} \frac{\partial LReLU(z_i)}{\partial z_i} \right|}$$

Problem 7

Difference :-

- Continuous flows operate over continuous state space, where any small change in parameters can cause change in model o/p. Discrete flows operate over finite set of states making modeling difficult

- In continuous flows, the determinant of the Jacobian gives a measure of change in volume (density). whereas, in discrete space there is no volume/density change in conventional sense, hence Jacobian is not directly applicable.

- In discrete domain, a simple invertible function is a bitwise NOT operation. A discrete flow model could consist of a sequence of such invertible operations. For instance, given a binary input, applying a series of bitwise not operation & bitwise shifts could constitute the transformations of a flow model

Each transformation in the sequence must have a computable inverse to adhere to the principles of flow-based models. This simplistic model demonstrates discrete flow models' essence - applying a series of invertible transformations to discrete data while ensuring each step's invertibility & tractability