ECE 8803-GGDL Homework 2

Arati Ganesh

October 6, 2023

1. Kullback-Leibler divergence (15 points). In class, we learned about KL diver- gence. Here, we like to understand why KL is not a true distance measure due to its asymmetric nature. Understanding the asymmetric nature of KL is critical in understanding the loss function in VAEs.

   a. Generally, $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$. Provide an example of univariate distributions P and Q where $D_{KL}(P \parallel Q) \neq \infty$ and $D_{KL}(Q \parallel P) = \infty$.

   Answer : Let us assume -

   P Distribution: Uniform random variable on [0, 1]

$$P(x) = \begin{cases} 1 & \text{for } x \in [0, 1] \\ 0 & \text{elsewhere} \end{cases}$$

   Q Distribution: Uniform random variable on [-2, 2]

$$Q(x) = \begin{cases} \frac{1}{4} & \text{for } x \in [-2, 2] \\ 0 & \text{elsewhere} \end{cases}$$

   1. Forward KL Divergence :

   To calculate the forward KL divergence, we use the formula:

$$DKL(P \parallel Q) = \int_{-\infty}^{+\infty} P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx$$

   For our distributions, we have:

$$DKL(P \parallel Q) = \int_{0}^{1} 1 \cdot \log\left(\frac{1}{1/4}\right) dx = \int_{0}^{1} \log(4)\, dx = \log(4)$$

   2. Reverse KL Divergence :

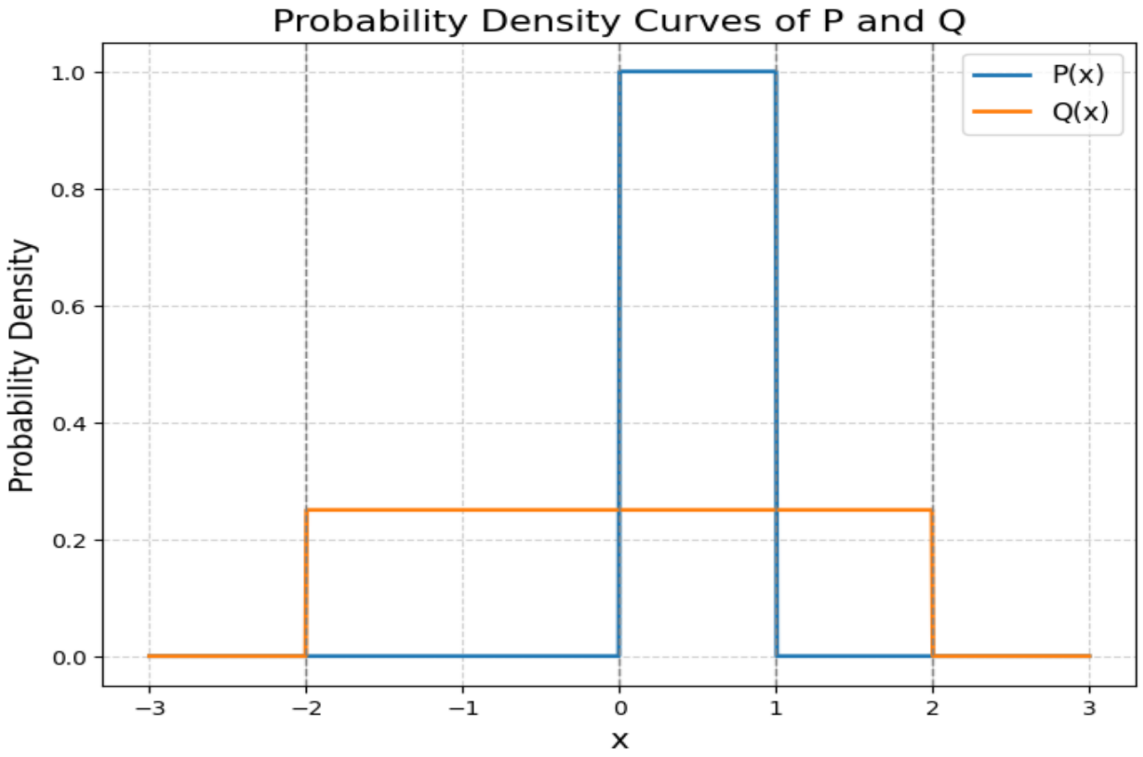
   To calculate the reverse KL divergence, we use the formula:

$$DKL(Q \parallel P) = \int_{-\infty}^{+\infty} Q(x) \log\left(\frac{Q(x)}{P(x)}\right) dx$$

   For our distributions, we have:

$$DKL(Q \parallel P) = \int_{0}^{1} 0 \cdot \log\left(\frac{0}{1}\right) dx + \int_{-2}^{2} \frac{1}{4} \cdot \log\left(\frac{1/4}{0}\right) dx = \int_{-2}^{2} \frac{1}{4} \cdot \log(\infty)\, dx$$

The second term in the integral becomes $\log(\infty)$, which implies that Reverse KL is infinite. This divergence can be infinite when P(x) is zero somewhere in its support where Q(x) is non-zero.



---

b. For a fixed target distribution P, we call $DKL(P \parallel Q)$ the forward-KL and $DKL(Q \parallel P)$ the reverse-KL. Due to the asymmetric nature of KL, distributions Q that minimize $DKL(P \parallel Q)$ can be different from those minimizing $DKL(Q \parallel P)$. From the following plots, identify which of (A, B) corresponds to minimizing forward and reverse KL. Here, only the mean and standard deviation of Q is allowed to vary during the minimization. Give a brief reasoning.

Answer :

(a) **(A) corresponds to minimizing the Forward KL:**

- The forward-KL measures how well Q approximates P, with an emphasis on areas where P has non-zero probability density.
- The penalty function contributes loss to the total KL wherever $P(Z) > 0$. In such regions, $\lim_{Q(Z)\to 0} \log\left(\frac{P(Z)}{Q(Z)}\right) \to \infty$. This means that the forward-KL will be large wherever $Q(Z)$ fails to overlap $P(Z)$.
- To minimize the forward-KL, we need to ensure that $Q(Z) > 0$ wherever $P(Z) > 0$. The optimized variational distribution $Q(Z)$ is known as "zero-avoiding. That is, it avoids having zero values in regions where $P(Z)$ is non-zero.

(b) **(B) corresponds to minimizing the Reverse KL:**

- The reverse-KL measures how well P approximates Q, with an emphasis on areas where Q has non-zero probability density.

- When $P(Z) = 0$, we must ensure that the weighting function $Q(Z) = 0$ as well, as otherwise, the reverse-KL diverges.
- To minimize the reverse-KL, we use a "zero-forcing" approach, where we make sure that $Q(Z)$ is forced to be zero in regions where $P(Z)$ is zero. This effectively "squeezes" the $Q(Z)$ under $P(Z)$.

---

c. What are the implications of this asymmetry in Variational Autoencoders (VAE) while maximizing the evidence lower bound (ELBO):

$$L(x; \theta, \phi) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)\|p(z))$$

Answer :
Minimizing Forward KL ($D_{KL}(q_\phi(z|x)\|p(z))$): Minimizing the forward KL divergence encourages the encoder (represented by $q_\phi(z|x)$) to match the approximate posterior distribution ($q_\phi(z|x)$) to the prior distribution ($p(z)$). This regularization is beneficial for controlling the variability of the latent space. It helps ensure that the latent representations are well-behaved and adhere to the prior assumptions. However, excessive minimization of the forward KL divergence can lead to over-regularization. This results in an overly smooth latent space that lacks the necessary diversity to accurately represent complex data distributions.

Minimizing Reverse KL ($D_{KL}(p(z)\|q_\phi(z|x))$): Overemphasizing the reverse KL divergence focuses on fitting the prior distribution ($p(z)$) to the approximate posterior ($q_\phi(z|x)$). This approach can lead to overfitting, where the prior distribution tries to capture noise present in the training data. As a result, the latent space becomes overly flexible but may not generalize effectively to unseen data.

The implication of this asymmetry is that while maximising the ELBO, the forward and reverse KL have opposing effects. Balancing these divergences is key to striking the right equilibrium between compression and flexibility, avoiding overfitting and over-regularization, and generating high-quality sample.

---

2. EM algorithm for a Mixture of Bernoullis (25 points). You will derive an expectation-maximization (EM) algorithm to cluster black and white images. The inputs x (i) can be thought of as vectors of binary values corresponding to black and white pixel values, and the goal is to cluster the images into groups. You will be using a mixture of Bernoullis model to tackle this problem.

   a. **Mixture of Bernoullis:**

   i) Bernoulli Parameters: Consider a vector of binary random variables $\mathbf{x} \in \{0, 1\}^D$. Assume each variable x_d is drawn from a Bernoulli(p_d) distribution, where $P(x_d = 1) = p_d$. Let $\mathbf{p} \in (0, 1)^D$ be the resulting vector of Bernoulli parameters. Write an expression for $P(\mathbf{x}|\mathbf{p})$.
   Answer :

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

   ii) Mixture of K Bernoulli Distributions: Now suppose we have a mixture of K Bernoulli distributions: each vector $\mathbf{x}^{(i)}$ is drawn from some vector of Bernoulli random variables with parameters

$\mathbf{p}^{(k)}$, which we will call Bernoulli($\mathbf{p}^{(k)}$). Let $\{\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(K)}\} = \mathbf{p}$. Assume a distribution $\pi(k)$ over the selection of which set of Bernoulli parameters $\mathbf{p}^{(k)}$ is chosen. Write an expression for $P(\mathbf{x}^{(i)}|\mathbf{p}, \pi)$.
Answer :

$$P(x^{(i)}|\mathbf{p}, \pi) = \sum_{k=1}^{K} \pi(k) \prod_{d=1}^{D} \left(p_d^{(k)}\right)^{x_d^{(i)}} \left(1 - p_d^{(k)}\right)^{1-x_d^{(i)}}$$

iii) Log Likelihood of Data: Finally, suppose we have inputs $X = \{\mathbf{x}^{(i)}\}_{i=1}^{n}$. Using the above, write an expression for the log likelihood of the data X, $\log P(X|\pi, \mathbf{p})$.
Answer :

$$\log P(X|\pi, \mathbf{p}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi(k) \prod_{d=1}^{D} \left(p_d^{(k)}\right)^{x_d^{(i)}} \left(1 - p_d^{(k)}\right)^{1-x_d^{(i)}} \right)$$

---

b. **Expectation step:**

i. Now, we introduce the latent variables for the EM algorithm. Let $z^{(i)} \in \{0, 1\}^K$ be an indicator vector, such that $z_k^{(i)} = 1$ if $x^{(i)}$ was drawn from a Bernoulli($p^{(k)}$), and 0 otherwise. Let $Z = \{z^{(i)}\}_{i=1}^{n}$. What is $P(z^{(i)}|\pi)$? What is $P(x^{(i)}|z^{(i)}, p, \pi)$?
Answer :

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

$$P(x^{(i)}|z^{(i)}; p, \pi) = \sum_{k=1}^{K} P(x^{(i)}|p_k)^{z_k^{(i)}}$$

ii. Using the above two quantities, derive the likelihood of the data and the latent variables, $P(Z, X|\pi, p)$.
Answer :

$$\begin{aligned} P(Z, X|\pi, p) &= \prod_{i=1}^{N} P(x^{(i)}, z^{(i)}|\pi, p) \\ &= \prod_{i=1}^{N} P(x^{(i)}, |z^{(i)}, \pi, p) P(z^{(i)}|\pi) \\ &= \prod_{i=1}^{N} \left(\prod_{k=1}^{K} P(x^{(i)}|p^{(k)})^{z_k^{(i)}} \right) \left(\prod_{k=1}^{K} \pi_k^{z_k^{(i)}} \right) \end{aligned}$$

iii. Let $\eta(z_k^{(i)}) = E[z_k^{(i)}|x^{(i)}, \pi, p]$. Show that

$$\eta(z_k^{(i)}) = \frac{\pi_k \prod_{d=1}^{D} \left(p_d^{(k)}\right)^{x_d^{(i)}} \left(1 - p_d^{(k)}\right)^{1-x_d^{(i)}}}{\sum_j \pi_j \prod_{d=1}^{D} \left(p_d^{(j)}\right)^{x_d^{(i)}} \left(1 - p_d^{(j)}\right)^{1-x_d^{(i)}}}$$

Answer :
Let $\hat{p}, \hat{\pi}$ be the new parameters that we'd like to maximize, so p, π are from the previous iteration. Use this to derive the following final expression for the E step in the expectation-maximization algorithm:

$$E[\log P(Z, X|\hat{p}, \hat{\pi})|X, p, \pi] = \sum_{i=1}^{N}\sum_{k=1}^{K}\eta(z_k^{(i)}) \times \left[\log \hat{\pi}_k + \sum_{d=1}^{D}\left(x_d^{(i)} \log \hat{p}_d^{(k)} + (1 - x_d^{(i)}) \log(1 - \hat{p}_d^{(k)})\right)\right]$$

$$\begin{aligned}\eta(z_k^{(i)}) &= E[z_k^{(i)}|x^{(i)}, \pi, p] \\ &= P[z_k^{(i)} = 1|x^{(i)}, \pi, p] \\ &= \frac{P(x^{(i)}|z_k^{(i)} = 1, p, \pi)P(z_k^{(i)} = 1|\pi)}{\sum_{k'} P(x^{(i)}|z_{k'}^{(i)} = 1, p, \pi)P(z_{k'}^{(i)} = 1|\pi)} \\ &= \frac{\pi_k \prod_{d=1}^{D}(p_d^{(k)})^{x_d^{(i)}}(1 - p_d^{(k)})^{1-x_d^{(i)}}}{\sum_{k'} \pi_{k'} \prod_{d=1}^{D}(p_d^{(k')})^{x_d^{(i)}}(1 - p_d^{(k')})^{1-x_d^{(i)}}}\end{aligned}$$

$$\begin{aligned}\log P(Z, D|\pi, p) &= \sum_{i=1}^{N}\left(\sum_{k=1}^{K} z_k^{(i)} \log\left(P(x^{(i)}|p^{(k)})\right) + \sum_{k=1}^{K} z_k^{(i)} \log \pi_k\right) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} z_k^{(i)}\left(\log \pi_k + \sum_{d=1}^{D}\left[x_d^{(i)} \log(p_d^{(k)}) + (1 - x_d^{(i)}) \log(1 - p_d^{(k)})\right]\right)\end{aligned}$$

c. Maximization Step

i. To maximize the above expression with respect to $\hat{\pi}$ and $\hat{p}$, first, show that the value of $\hat{p}(k)$ that maximizes the E step is:

$$\hat{p}(k) = \frac{\sum_{i=1}^{N} \eta(z_k^{(i)})x^{(i)}}{N_k}$$

where $N_k = \sum_{i=1}^{N} \eta(z_k^{(i)})$.
Answer :
Setting the derivative to 0:

$$\frac{d}{dp_d^{(k)}} E[\log P(Z, D|\theta, p)] = \sum_{i=1}^{N} \eta(z_k^{(i)})\left[x_d^{(i)} \frac{d}{d(p_d^{(k)})}(\log p_d^{(k)}) + (1 - x_d^{(i)}) \frac{d}{d(p_d^{(k)})}\left(\log(1 - p_d^{(k)})\right)\right] = 0$$

Multiply by the denominators:

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

Solving for $p_d^{(k)}$ results in:

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

ii. Show that the value of $\hat{\pi}_k$ that maximizes the E step is:
Answer :

$$\hat{\pi}_k = \frac{N_k}{\sum_{k'} N_{k'}}$$

The exponential families notation may be useful. Alternatively, you can use Lagrange multipliers.

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

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

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

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

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

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

---

3. You will now cluster the images on the MNIST dataset by implementing the algorithm derived above. Each input is a binary number corresponding to black and white pixels, and is a flattened version of the 28x28 pixel image. These are the conventions we will use:

   (a) $N = 1000$ is the number of datapoints (sampled at random from the full MNIST dataset).

   (b) D is the dimension of each input.

   (c) K is the number of clusters.

   (d) Xs is an $N \times D$ matrix of the input data, where row i is the pixel data for picture i.

   (e) p is a $K \times D$ matrix of Bernoulli parameters, where row k is the vector of parameters for the kth mixture of Bernoullis.

   (f) mix_p is a $K \times 1$ vector containing the distribution over the various mixtures.

   (g) η is an $N \times K$ matrix containing the results of the E step, so $\eta[i, k] = \eta(z_k^{(i)})$.

Answer :
Part a - c in ipynb notebook

d. Run the EM algorithm and plot the resulting Bernoulli parameters p. In order to do so, you need to reshape each row into a 28x28 matrix and print the resulting grayscale image. What do you see? Repeat the experiment for K = 5 and K = 20 and explain how the results differ.
Answer :

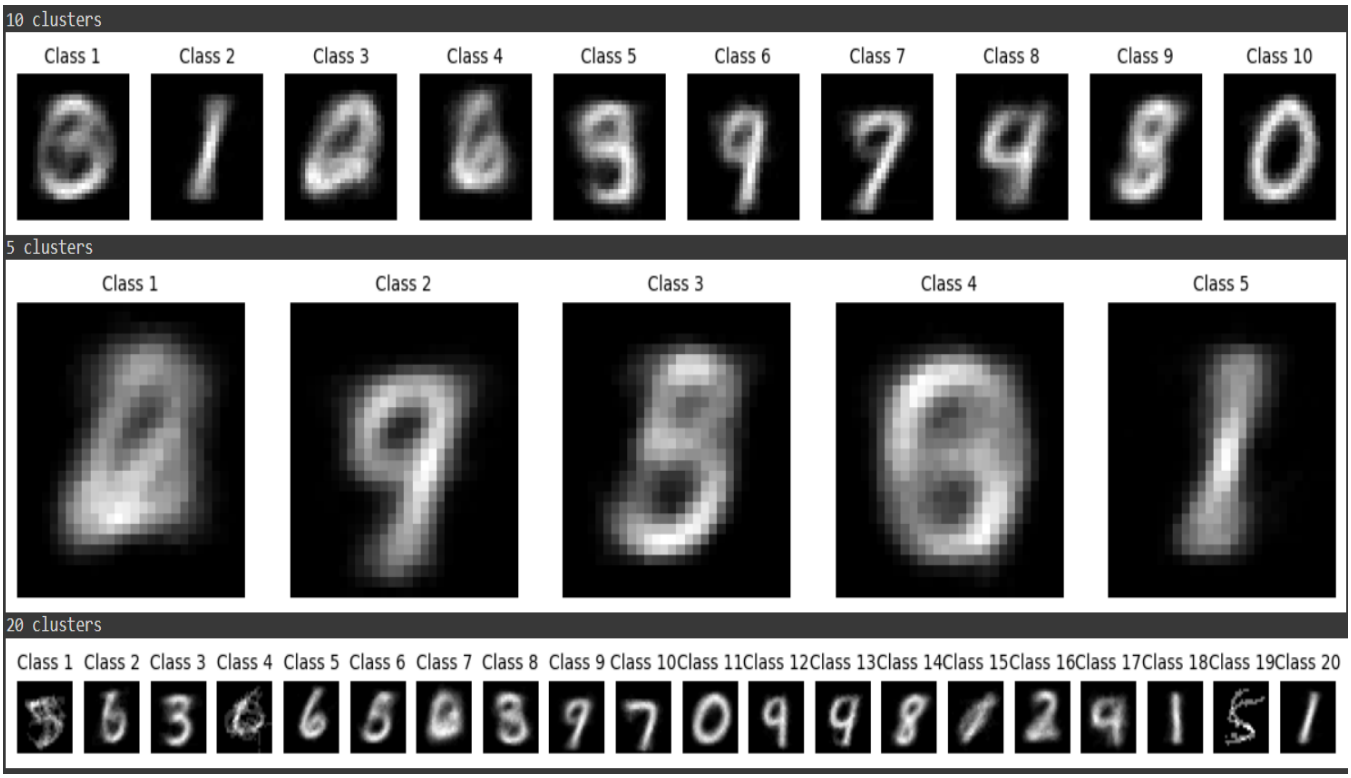


Figure 1: EM Cluster wise Result

The results differ in the sense that smaller the cluster, many more different digits are clustered together so there is more variation. More clusters usually leads to a cluster being associated with a single digit only.

---

e. Explain the process you would follow to generate new images using the parameters obtained through the above algorithm. Based on your observations from the three cases in the previous part, explain why (or why not) the generated images would be plausible samples from the 'original' distribution.

Answer :
To create new samples, the following steps are followed:

Step 1: Selecting a Cluster/Component Utilize the cluster weights denoted as mix_p to choose a specific component or cluster. This selection process involves sampling from a categorical distribution, with probabilities determined by the mix_p values.

Mathematically, this step can be represented as:

$$\text{Cluster\_selected} \sim \text{Categorical}(\text{mix\_p})$$

Where:

- Cluster_selected represents the chosen cluster or component.
- Categorical(mix_p) is the categorical distribution based on the probabilities defined in mix_p.

Step 2: Sampling from the Cluster Within the selected cluster, every pixel follows a Bernoulli distribution. Sampling from this distribution is done to generate new images.

Mathematically, this process can be represented as:

$$\text{Pixel\_sampled} \sim \text{Bernoulli}(\text{p})$$

Where:

- Pixel_sampled represents the sampled pixel value.
- Bernoulli(p) is the Bernoulli distribution with probability p.

The quality of the generated samples can be significantly impacted by the choice of hyperparameters, such as the number of clusters. Properly selecting these hyperparameters is essential to ensure that the generated images closely resemble the original distribution.

---

4. NICE vs VAE on MNIST

   In this problem, we will compare the performance of a NICE (Non-linear Independent Component Estimation) model and a VAE (Variational Autoencoder) implementation on the MNIST dataset. To conduct this comparison, you will need to clone the repository located at https://github.com/EugenHotaj/pytorch-generative into your working directory.

   a. Train the NICE model for 50 epochs and 256 batch size (keep the other parameters at default). Visualize a few samples from the model and comment on the result.
   Answer :

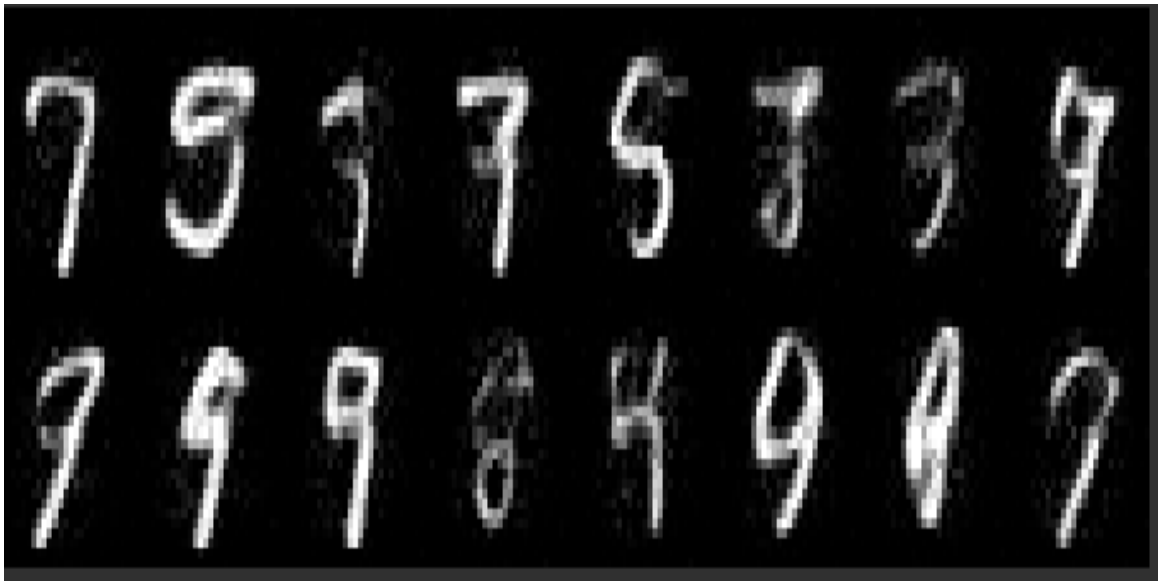


Figure 2: NICE Result

The results generated by NICE are pretty clear but only some of the digits are well formed.

---

b. Repeat the process with the VAE model with the same parameters as above (keep in mind that the setups for the models and loss functions are different). Does this models generate more plausible samples? Explain
Answer :

Figure 3: VAE Result

The results generated by VAE are clear and digits well formed. Yes, there is more variation in the

generated digits due to the low dimension latent representation. In VAE

---

c. Considering these two models as well as the EM-based generation you described in P3.e, which one would be more suitable for this type of data? Your answer may consider runtime, generative expressiveness/power, simplicity or any other factors you consider relevant.
Answer :
Generative Expressiveness/Power:

- NICE: Captures data structure but less expressive for detailed image generation.
- VAE: Explicitly designed for generative modeling with success in complex image generation.
- EM-Based: Expressiveness depends on the chosen model's accuracy in representing data.

Simplicity:

- NICE: Conceptually simple with focus on invertible transformations.
- VAE: Involves probabilistic framework and variational inference, requiring careful tuning.
- EM-Based: Conceptually simpler, especially with established models like Gaussian Mixtures.

Runtime Efficiency:

- NICE: More efficient at runtime due to simple transformations.
- VAE: Can be computationally expensive during training but faster in inference.
- EM-Based: Can be efficient, but convergence of many iterations may take time.

---

5. Simple flow models (15 points).
a. Let X be a Cauchy random variable with the probability density function (pdf):

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

Answer :
To find the distribution of the random variable $Y = \frac{1}{X}$, where X is a Cauchy random variable with the probability density function (pdf) $f_X(x) = \frac{1}{\pi(1+x^2)}$:

PDF of X:

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

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$$

$$f_Y(y) = f_X\left(\frac{1}{y}\right) \cdot \left| -\frac{1}{y^2} \right|$$

Simplifying :

$$f_Y(y) = \frac{1}{y^2} \cdot \left| f_X\left(\frac{1}{y}\right) \right|$$

Now, find the pdf of Y using the method of transformations:

$$f_Y(y) = f_X\left(\frac{1}{y}\right) \cdot \frac{1}{y^2}$$

Substitute $\frac{1}{y}$ for x in the pdf of X:

$$f_X\left(\frac{1}{y}\right) = \frac{1}{\pi\left(y^{-2}+1\right)}$$

The pdf of Y is also a Cauchy distribution with the form:

$$f_Y(y) = \frac{1}{\pi y^2(y^{-2}+1)}$$

---

b. Let Y be a normal random variable with parameters μ and σ. Find the probability density function of e^Y.

Answer :
Pdf of Y is $f_Y(y)$, and the pdf of e^Y is $f_{e^Y}(z)$, where $z = e^y$. We want to find $f_{e^Y}(z)$.

The transformation equation is as follows:

$$f_{e^Y}(z) = f_Y(\ln(z)) \cdot \left| \frac{d}{dz} \ln(z) \right|$$

Now, for a normal random variable Y with parameters μ and σ, the pdf $f_Y(y)$ is given by:

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

Substitute $\ln(z)$ for y in the above expression:

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

Derivative of $\ln(z)$ with respect to z:

$$\frac{d}{dz} \ln(z) = \frac{1}{z}$$

Substitute:

$$f_{e^Y}(z) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(\ln(z)-\mu)^2}{2\sigma^2}} \cdot \frac{1}{z}$$

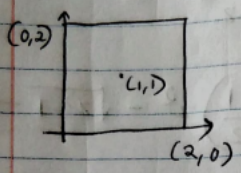
Simplify:

$$f_{e^Y}(z) = \frac{1}{\sigma\sqrt{2\pi}z} \cdot e^{-\frac{(\ln(z)-\mu)^2}{2\sigma^2}}$$

---

c. We choose a point P uniformly on the interval $[0, 2]^2$, and we denote the distance of P from the point $(1, 1)$ by Z. Find the probability density function of Z.
Answer :

c) $P \sim \text{uniform}([0,2])$



Let Z = Distance from $(1,1)$ to point $P(x,y)$

$$= \sqrt{(x-1)^2+(y-1)^2}$$

$\therefore$ cdf Z: $F_Z(z) = P_Z(z \le Z)$

Let $A = \{(x,y) \mid \sqrt{(x-1)^2+(y-1)^2} \le z\}$

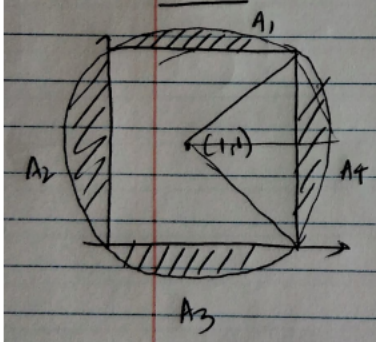
$$P_{xy}(x,y) = \begin{cases} \frac{1}{4} & 0<x<2 \text{ and } 0<y<2 \\ 0 & \text{otherwise} \end{cases}$$

<u>case 1:</u> $0 \le z \le 1$

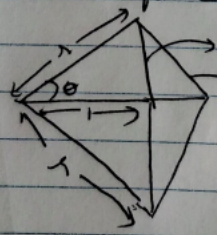
↳ set A is a circle inside support of $\{x,y\}$

$$\therefore F_Z(z) = \iint_A P_{xy}(x,y)\,dx\,dy = \frac{\pi z^2}{4}$$

<u>case 2:</u> $1 \le z \le 2$



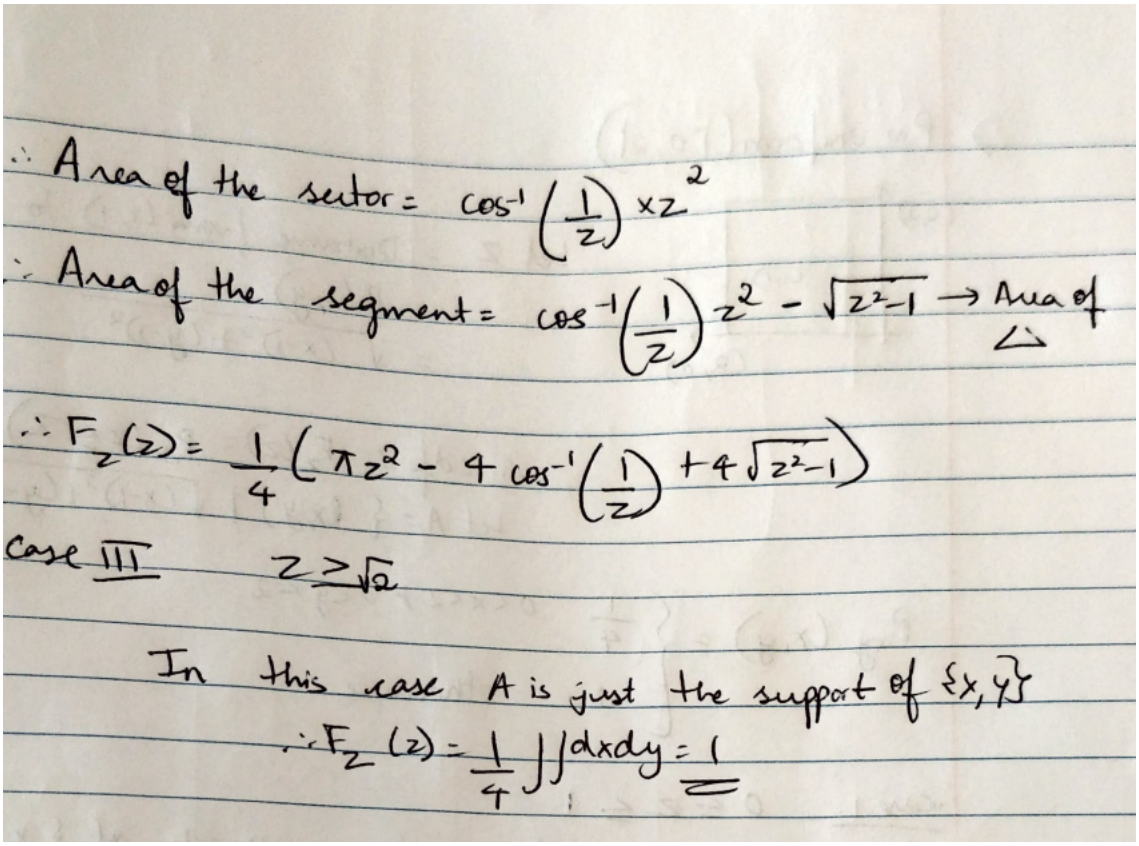
In this case A is the circle minus regions $A1, A2, A3, A4$ that are outside the square



$$\sqrt{z^2-1+(z-1)^2} = \sqrt{2z^2-2z}$$

$\therefore$ using cosine rule $\cos(\theta) = \frac{z^2+z^2-(2z^2-2z)}{2z^2}$

$$\theta = \cos^{-1}\left(\frac{1}{2}\right) \quad \boxed{z = 1}$$



---

6. Jacobian of the Leaky Flow (15 points)

   The Leaky ReLU is defined as:

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

   Write an expression for the inverse of the Leaky ReLU. Write an expression for the inverse absolute determinant of the Jacobian $|\frac{\partial f[z]}{\partial z}|^{-1}$ for an elementwise transformation $x = f[z]$ of the multivariate variable z where:

$$f[z] = \begin{bmatrix} LReLU[z_1] \\ LReLU[z_2] \\ \vdots \\ LReLU[z_n] \end{bmatrix}$$

   Answer :

   For i-th element of x:

   - If $LReLU[z_i] = 0.1z_i$ (for $z_i < 0$), then $D_{i,i} = 0.1$. - If $LReLU[z_i] = z_i$ (for z_i0), then $D_{i,i} = 1$.

   The Jacobian matrix D is a diagonal matrix with these conditional elements:

$$D = \begin{bmatrix} 0.1 & 0 & 0 & \cdots & 0 \\ 0 & 0.1 & 0 & \cdots & 0 \\ 0 & 0 & 0.1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

The inverse determinant of the Jacobian matrix D is:

$$\text{Inverse Determinant} = \frac{1}{\det(D)} = \frac{1}{(0.1^{N_{<0}}) \cdot (1^{N-N_{<0}})}$$

Compute the Inverse Absolute Determinant of the Jacobian:

Therefore, $|(\det \frac{\partial f[z]}{\partial z})|^{-1} = 10^{N_{<0}}$
