GEORGIA INSTITUTE OF TECHNOLOGY
SCHOOL of ELECTRICAL and COMPUTER ENGINEERING

ECE 8803-GGDL Fall 2023
Problem Set #2

Assigned: 25 Sep
Due Date: 6 Oct

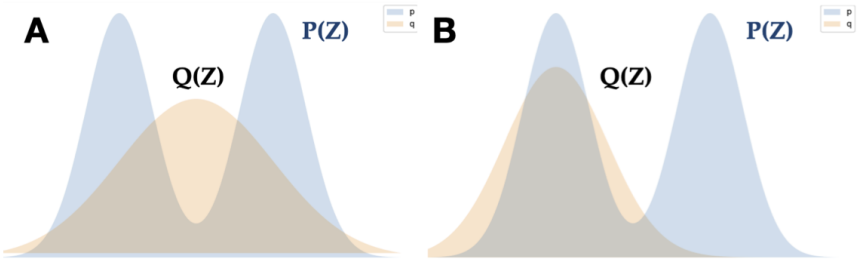
Please contact the TAs for clarification on the instructions in the homework assignments.

Problem 1: Kullback-Leibler divergence (15 points). In class, we learned about KL divergence. 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 $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$. Give an example of univariate distributions P and Q where $D_{\mathrm{KL}}(P\|Q) \neq \infty$ and $D_{\mathrm{KL}}(Q\|P) = \infty$.

b. For a fixed target distribution P, we call $D_{\mathrm{KL}}(P\|Q)$ the forward-KL and $D_{\mathrm{KL}}(Q\|P)$ the reverse-KL. Due to the asymmetric nature of KL, distributions Q that minimize $D_{\mathrm{KL}}(P\|Q)$ can be different from those minimizing $D_{\mathrm{KL}}(Q\|P)$. 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.

c. What are the implications of this asymmetry in VAE while maximizing the evidence lower bound: $\mathcal{L}(x;\theta,\phi) = E_{q_\phi(z|x)}[\log p(x|z;\theta)] - D_{KL}(q_\phi(z|x)\|p(z))$.



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

ii. Now suppose we have a mixture of K Bernoulli distributions: each vector $x^{(i)}$ is drawn from some vector of Bernoulli random variables with parameters $p^{(k)}$, we will call this Bernoulli($p^{(k)}$). Let $\{p^{(1)}, \ldots, p^{(K)}\} = \mathbf{p}$. Assume a distribution $\pi(k)$ over the selection of which set of Bernoulli parameters $p(k)$ is chosen. Write an expression for $P(x^{(i)}|\mathbf{p}, \pi)$.

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

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 $P(x^{(i)}|z^{(i)}, \mathbf{p}, \pi)$?

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

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

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

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

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

c. **Maximization step**

i. We need to maximize the above expression with respect to $\hat{\pi}, \hat{\mathbf{p}}$. First, show that the value of $\hat{\mathbf{p}}$ that maximizes the E step is

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

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

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

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

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

Problem 3: Expectation-Maximization on MNIST (20 points). 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:

- $N = 1000$ is the number of datapoints (sampled at random from the full MNIST dataset), D is the dimension of each input, and K is the number of clusters.
- `Xs` is an $N \times D$ matrix of the input data, where row i is the pixel data for picture i.
- `p` is a $K \times D$ matrix of Bernoulli parameters, where row k is the vector of parameters for the kth mixture of Bernoullis.

- `mix_pi` is a $K \times 1$ vector containing the distribution over the various mixtures.
- `eta` is a $N \times K$ matrix containing the results of the E step, so `eta[i,k]` $= \eta(z_k^{(i)})$

a. Implement the E step of the algorithm within the function `Estep(Xs,p,mix_pi)`, saving your calculated values in the returned variable `eta`.

b. Implement the M step of the algorithm within the function `Mstep(Xs, eta, alpha1, alpha2)`. The function should return the new values `p` and `mix_pi` that maximize the E step. `alpha1, alpha2` are Dirichlet smoothing parameters (explained below).

c. Implement the full EM algorithm in the function `EM(Xs, K, iter)` by alternating between the E and M steps for `iter=`20 iterations. The function should return the final parameters `p` and `mix_pi`.

A few tips:

- Use the log operator to keep your calculations numerically stable (pay special attention to $= \eta(z_k^{(i)})$).
- You will need to avoid zeros in π and $\mathbf{p}$ or else you will take $\log(0) = -\infty$. Use Dirichlet prior smoothing with the parameters $\alpha_1 = \alpha_2 = 10^{-8}$ when updating these variables:

$$\hat{p}^{(k)} = \frac{\sum_{i=1}^{N} \eta(z_k^{(i)}) x^{(i)} + \alpha_1}{N_k + \alpha_1 D}$$
$$\hat{\pi}_k = \frac{N_k + \alpha_2}{\sum_{k'} N_{k'} + \alpha_2 K}$$

- Initialize your parameters $\mathbf{p}$ by randomly sampling from a Uniform(0, 1) distribution and normalizing each $p^{(k)}$ to have unit length, and $\pi_k = 1/k$.

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.

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 òriginal' distribution.

Problem 4: NICE vs VAE on MNIST (10 points). We will now compare the performance of a NICE and VAE implementation on MNIST, again using the repository `https://github.com/EugenHotaj/pytorch-generative` which you will need to clone 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.

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.

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.

Problem 5: Simple flow models (15 points).

a. Let X be a Cauchy random variable with pdf $\frac{1}{\pi(1+x^2)}$. Find the distribution of $\frac{1}{X}$.

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

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

Problem 6: Jacobian of the Leaky Flow (15 points). The Leaky ReLU is defined as: $\text{LReLU}[z] = 0.1z$ for $z < 0$ and $\text{LReLU}[z] = z$ for $z > 0$. Write an expression for the inverse of the leaky ReLU. Write an expression for the inverse absolute determinate of the Jacobian $|\partial \mathbf{f}[\mathbf{z}]/\partial \mathbf{z}|^{-1}$ for an elementwise transformation $\mathbf{x} = \mathbf{f}[z]$ of the multivariate variable $\mathbf{z}$ where:

$$\mathbf{f}[\mathbf{z}] = [\text{LReLU}[z_1], \text{LReLU}[z_2], \ldots, \text{LReLU}[z_n]]^T. \tag{1}$$

Problem 7 (Bonus question): Discrete Flows (15 points). Consider the problem of developing a Normalizing Flow model in the discrete space. Describe how and why the modeling will be different from the continuous Flow model. Do we need to compute the Jacobian? Bring an example of an invertible function in the discrete case and develop a working discrete flow model. Further reading: `https://arxiv.org/pdf/1905.10347.pdf`