In The Name of God

Sharif University of Technology
Electrical Engineering Department

Deep Generative Models

Assignment 2

Fall 2024

Instructor: Dr. S. Amini

Due on Aban 10, 1403 at 23:55

1 D_{kl} Intuitions

1.1 Which One?

Consider you have a *target distribution* $p(x \mid \theta)$ and an *approximate distribution* $q(x \mid \phi)$. Our objective is to optimize ϕ to "fit" $q(x \mid \phi)$ such that it becomes close to $p(x \mid \theta)$. A common dilemma arises: Should we solve

$$\arg\min_{\phi} \mathrm{KL}(q(x \mid \phi), p(x \mid \theta)) \quad \text{or} \quad \arg\min_{\phi} \mathrm{KL}(p(x \mid \theta), q(x \mid \phi))?$$

There is no one-size-fits-all answer to this question, as the optimal approach depends on the specific properties of $p(x \mid \theta)$ and $q(x \mid \phi)$.

In this problem, you will be provided with several scenarios. For each one, choose one of the following options and provide a concise one-to-two-sentence justification for your choice, e.g. why a particular objective allows us to optimize ϕ and how you plan to perform the optimization:

A. We can minimize $\mathrm{KL}(q(x \mid \phi), p(x \mid \theta))$

B. We can minimize $\mathrm{KL}(p(x \mid \theta), q(x \mid \phi))$

C. We can minimize both $\mathrm{KL}(q(x \mid \phi), p(x \mid \theta))$ and $\mathrm{KL}(p(x \mid \theta), q(x \mid \phi))$

D. None of the above, i.e. neither direction of KL divergence can be minimized.

The scenarios are listed below:

- **Scenario 1:** We can access samples and evaluate the exact density for both $p(x \mid \theta)$ and $q(x \mid \phi)$.
- **Scenario 2:** We can access samples from $p(x \mid \theta)$ but we do not have access to the density function. For $q(x \mid \phi)$, we do not have access to samples but we can evaluate the density.
- **Scenario 3:** We do not have access samples from $p(x \mid \theta)$ and we only know its *unnormalized* density. However, we have access to both samples and (normalized) density function from $q(x \mid \phi)$.

Hint 1: You may not assume any closed-form solution for integral or expectation therefore you would need Monte Carlo estimation to estimate any expectations you come across.

Hint 2: You may use the log-derivative trick :

$$\nabla_{\phi} \mathbb{E}_{x \sim q(x|\phi)}[f(x; \phi)] = \mathbb{E}_{x \sim q(x|\phi)}[\nabla_{\phi} \log q(x \mid \phi) f(x)]$$

or you may use reparameterization trick.

Answer:

Scenario 1: if you write the difinitaion of KL you can see that We can minimize both $KL(q(x|\phi), p(x|\theta))$ and $KL(p(x|\theta), q(x|\phi))$

Scenario 2:

Minimizing $KL(p \parallel q)$:

$$D_{KL}(p \parallel q) = \mathbb{E}_{x \sim p(x|\theta)} \left[\ln \left(\frac{p(x|\theta)}{q(x|\phi)} \right) \right] = \mathbb{E}_{x \sim p(x|\theta)} \left[\ln \left(\frac{1}{q(x|\phi)} \right) \right] - \mathcal{H}(p(x|\theta))$$

When we optimize with respect to ϕ, the term $\log p(x \mid \theta)$ remains constant because it does not depend on ϕ, which means it does not impact the gradient. This allows us to use samples from $p(x \mid \theta)$ to compute $\log q(x \mid \phi)$ and minimize $KL(p \parallel q)$ effectively with respect to ϕ.

also not that this is the same as minimization cross-entropy

$$\arg\min_{\phi} D_{KL}(p(x|\theta), q(x|\phi)) = \arg\min_{\phi} \mathcal{H}(p(x|\theta), q(x|\phi)) \approx \arg\min_{\phi} \sum_{x_i \sim p(x|\theta)}^{N} \ln \left(\frac{1}{q(x_i|\phi)} \right)$$

Minimizing $KL(q \parallel p)$:

the KL divergence from q to p, defined as:

$$D_{KL}(q(x|\phi), p(x|\theta)) = \mathbb{E}_{x \sim q(x|\phi)} \left[\ln \left(\frac{q(x|\phi)}{p(x|\theta)} \right) \right] = \mathbb{E}_{x \sim p(x|\theta)} \left[\frac{q(x|\phi)}{p(x|\theta)} \ln \left(\frac{q(x|\phi)}{p(x|\theta)} \right) \right]$$

However, minimizing this divergence is not possible because we cannot sample directly from $q(x \mid \phi)$ or calculate $\log p(x \mid \theta)$. Due to these limitations, only $KL(p \parallel q)$ can be minimized effectively.

Scenario 3: Assume that the unnormalized density function for $p(x|\theta)$ is $f(x|\theta)$, with the intractable partition function $Z(\theta)$:

$$p(x|\theta) = \frac{f(x|\theta)}{Z(\theta)}$$

For $D_{KL}(q(x|\phi), p(x|\theta))$, we can see:

$$D_{KL}(q(x|\phi), p(x|\theta)) = \mathbb{E}_{x \sim q(x|\phi)} \left[\ln \left(\frac{q(x|\phi)}{p(x|\theta)} \right) \right] = \mathbb{E}_{x \sim q(x|\phi)} \left[\ln \left(\frac{q(x|\phi)}{f(x|\theta)} \right) + \ln(Z(\theta)) \right]$$
$$= \mathbb{E}_{x \sim q(x|\phi)} \left[\ln \left(\frac{q(x|\phi)}{f(x|\theta)} \right) \right] + \ln(Z(\theta)) \quad \text{(constant)}$$

Therefore, we can calculate $\nabla_{\phi} D_{KL}(q(x|\phi), p(x|\theta))$ and optimize it using REINFORCE or by reparameterization tricks if applicable.

Now considering $D_{KL}(p(x|\theta), q(x|\phi))$, the minimization can be done by:

$$D_{KL}(p \parallel q) = \mathbb{E}_{x \sim p(x|\theta)} \left[\ln \left(\frac{p(x|\theta)}{q(x|\phi)} \right) \right] = \mathbb{E}_{x \sim p(x|\theta)} \left[\ln \left(\frac{1}{q(x|\phi)} \right) \right] - \mathcal{H}(p(x|\theta)) \quad \text{(constant)}$$
$$= \mathbb{E}_{x \sim q(x|\phi)} \left[\frac{f(x|\theta)}{Z(\theta) q(x|\phi)} \ln \left(\frac{1}{q(x|\phi)} \right) \right] + c = \frac{1}{Z(\theta)} \mathbb{E}_{x \sim q(x|\phi)} \left[f(x|\theta) \ln \left(\frac{1}{q(x|\phi)} \right) \right] + c$$

$$\Rightarrow \arg\min_{\phi} D_{KL}(p(x|\theta), q(x|\phi)) = \arg\min_{\phi} \frac{1}{Z(\theta)} \mathbb{E}_{x \sim q(x|\phi)} \left[f(x|\theta) \ln \left(\frac{1}{q(x|\phi)} \right) \right]$$

$$= \arg\min_{\phi} \mathbb{E}_{x \sim q(x|\phi)} \left[f(x|\theta) \ln \left(\frac{1}{q(x|\phi)} \right) \right]$$

The minimization can again be done using REINFORCE or by reparameterization tricks.

We can minimize both $KL(q(x|\phi), p(x|\theta))$ and $KL(p(x|\theta), q(x|\phi))$

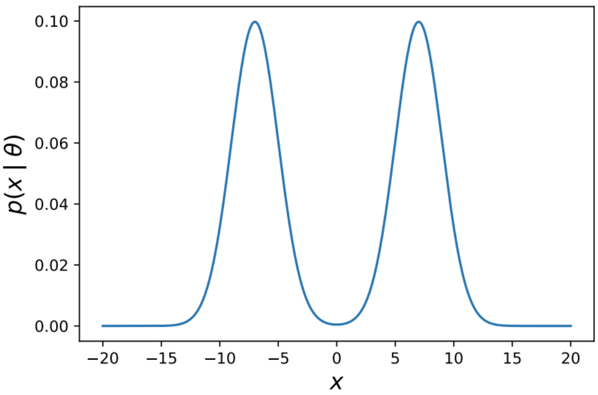

Figure 1: A two-mode Gaussian mixture

1.2 Mode-Seeking or Mode-Coverage

Recall the setting in previous problems, where you have a target distribution $p(x \mid \theta)$ and an approximate distribution $q(x \mid \phi)$. In this problem, we will study the "mode-seeking" behavior of $\mathrm{KL}(q(x \mid \phi) \parallel p(x \mid \theta))$ and the "mode-coverage" behavior of $\mathrm{KL}(p(x \mid \theta) \parallel q(x \mid \phi))$.

Now, assume we have $p(x \mid \theta)$ as a bi-mode symmetric Gaussian mixture, e.g., the one illustrated in Fig. 1. And we have $q(x \mid \phi)$ as a uni-variate Gaussian with learnable mean and variance. Now, please think about

(a) If we choose to minimize $\mathrm{KL}(q(x \mid \phi), p(x \mid \theta))$, is there a global minima? If not, how many local minima would there exist? Where would the minima(s) be located?

(b) If we choose to minimize $\mathrm{KL}(p(x \mid \theta), q(x \mid \phi))$, is there a global minima? If not, how many local minima would there exist? Where would the minima(s) be located?

(c) Based on your arguments from the (A) and (B), could you try to explain the meaning behind "mode-seeking" and "mode-coverage"?

Answer:

A:Mode seeking, two local minima near modes, no global minima This KL divergence is:

$$KL(q(x|\phi) \parallel p(x|\theta)) = \int q(x|\phi) \log \frac{q(x|\phi)}{p(x|\theta)} \, dx = \mathbb{E}_{x \sim q(x|\phi)} \left[\log \frac{q(x|\phi)}{p(x|\theta)} \right]$$

In this setup, minimizing $KL(q \parallel p)$ encourages $q(x|\phi)$ to cover only the most probable region of $p(x|\theta)$ rather than all of its support. Since if it covers one of the modes, the samples from the other mode are improbable, therefore we wouldn't really care to regress towards it. This behavior is **mode-seeking** because it tends to make $q(x|\phi)$ center around one of the modes of $p(x|\theta)$, ignoring other modes if q can't cover both simultaneously.

Local minima: This minimization will yield two local minima, where $q(x|\phi)$ aligns with either of the two modes of $p(x|\theta)$. Both positions are local minima because each one captures a high-probability region of $p(x|\theta)$.

Global minima: There is no single global minimum here since $q(x|\phi)$, being uni-modal, cannot capture both modes of $p(x|\theta)$ simultaneously. The modes are symmetric; therefore, neither is the preferred mode.

Thus, minimizing $KL(q \parallel p)$ results in two local minima where $q(x|\phi)$ is centered around either one of the modes of $p(x|\theta)$.

B:Mode coverage, single global minima between modes This KL divergence is:

$$KL(p(x|\theta) \parallel q(x|\phi)) = \int p(x|\theta) \log \frac{p(x|\theta)}{q(x|\phi)} \, dx = \mathbb{E}_{x \sim p(x|\theta)} \left[\log \frac{p(x|\theta)}{q(x|\phi)} \right]$$

Minimizing $KL(p \parallel q)$ encourages $q(x|\phi)$ to assign probability mass to all regions where $p(x|\theta)$ has probability mass. Since the samples are from p, q tries to be highly probable wherever p is probable. This behavior is **mode-coverage** because q is encouraged to "cover" both modes of p rather than focusing on one.

Since $p(x|\theta)$ is bi-modal and symmetric, $q(x|\phi)$, which is uni-modal, will place its mean between the two modes of $p(x|\theta)$ to minimize the divergence as much as possible.

Global minimum: There is a single global minimum, where $q(x|\phi)$ is centered between the two modes of $p(x|\theta)$, attempting to cover both modes symmetrically. Any divergence will increase the overall divergence; therefore, there are no local minima.

Thus, minimizing $KL(p \parallel q)$ results in one global minimum where $q(x|\phi)$ is centered between the two modes of $p(x|\theta)$.

C:

- Mode-Seeking (minimizing $KL(q \parallel p)$): Here, $q(x|\phi)$ seeks to place its probability mass on one of the modes of $p(x|\theta)$ rather than covering the entire distribution. This happens because the penalty for q not covering low-probability regions of p is minimal. Therefore, q only "seeks" one of the modes of p that gives the lowest divergence.
- Mode-Coverage (minimizing $KL(p \parallel q)$): Here, $q(x|\phi)$ is encouraged to cover all high-probability regions of $p(x|\theta)$. Any area where $p(x|\theta)$ has significant probability mass that $q(x|\phi)$ doesn't cover contributes a large penalty, so q tries to "cover" both modes by positioning itself between them.

2 General Covariance Matrix for VAE

In class, the VAE we considered having $q(\mathcal{Z}|\phi)$ as a multivariate Gaussian with diagonal covariance, this is also known as mean-field Gaussian. Now, we will consider extending it to general covariance matrix. In particular, consider a random vector drawn from a Gaussian

$$\epsilon \sim \mathcal{N}(0, I),$$

Which is then linearly transformed using the relation

$$z = \mu + L\epsilon$$

Where L is a lower or (upper) triangular matrix. The off-diagonal elements define the correlations (covariances) of the elements in z. Show that z has a distribution $\mathcal{N}(\mu, \Sigma)$ and write down an expression for Σ in terms of L. Explain why the diagonal elements of L must be positive. Describe how μ and L can be expressed as the outputs of a neural network and discuss suitable choices of output-unit activation functions.

Answer:

Using the change of variables formula, we have that $p(z(\epsilon)) = p(\epsilon)\det|L|^{-1}$. Furthermore, we have that $\epsilon = L^{-1}(z-\mu)$. So the log likelihood of z is

$$\begin{aligned}
\log p(z) &= -\frac{1}{2}\epsilon^T\epsilon - \frac{d}{2}\log(2\pi) - \log\det|L| && (3)\\
&= -\frac{1}{2}(z-\mu)^T L^{-T} L^{-1}(z-\mu) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det|LL^T| && (4)\\
&= -\frac{1}{2}(z-\mu)^T (LL^T)^{-1}(z-\mu) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det|LL^T| && (5)\\
&= \log N(z|\mu, LL^T) && (6)
\end{aligned}$$

The diagonal elements of L must be positive(Cholesky decomposition) in order for LL^T to be positive definite. note that theoretically we can have PSD covariance matrices but they are not very applicable

μ can be the output of an unconstrained neural network and the lower left triangular part of L can be parametrized using a neural network with $\frac{1}{2}d(d-1)$ parameters and we can pass the entries corresponding to the diagonal into a nonlinearity like softplus that will ensure that the values are strictly positive.

3 Gaussian Mixture Latent Model

In this question we will investigate how can we adapt VAE to have a Gaussian mixture embedding. The architrave of the model is in fig 2

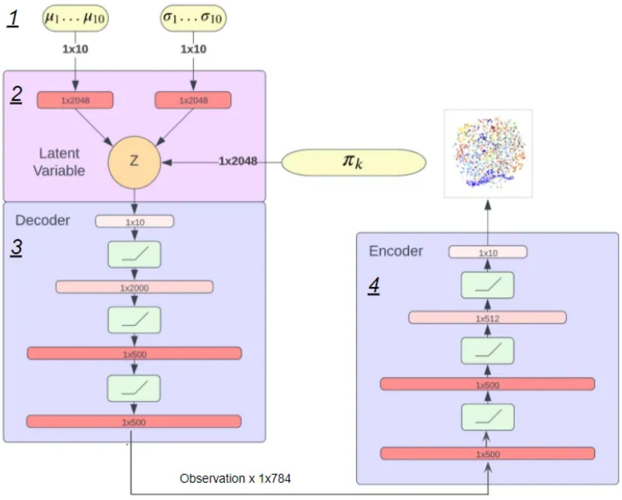

Figure 2: model architecture

This model generalizes the VAE model by using a Gaussian Mixture Model prior to replace the single Gaussian prior. This makes this network much more suitable for clustering tasks.

Below is an overview of how the network operates. Refer to the numbers in the figure 2.

1. A Cluster is selected from a Gaussian Mixture Model, and its mean and log variance is fed into the network
2. A latent embedding is generated based on the picked cluster
3. A DNN decodes the latent embedding into an observable x
4. An Encoder network is used to maximize the ELBO

More precisly:

1. Choose a cluster $c \sim \text{Cat}(\pi)$

2. Choose a latent vector $z \sim \mathcal{N}(\mu_c, \sigma_c^2 \mathbf{I})$

3. Choose a sample x:

   (a) If x is binary:

      i. Compute the expectation vector μ_x:

      $$\mu_x = f(z; \theta)$$

      ii. Choose a sample $x \sim \text{Ber}(\mu_x)$

   (b) If x is real-valued:

      i. Compute μ_x and σ_x^2:

      $$[\mu_x; \log \sigma_x^2] = f(z; \theta)$$

      ii. Choose a sample $x \sim \mathcal{N}(\mu_x, \sigma_x^2 \mathbf{I})$

- π_k is the prior probability for cluster k
- $\text{Cat}(\pi)$ is the categorical distribution parameterized by π
- μ_c and σ_c^2 are the mean and variance of the Gaussian Distribution corresponding to cluster c

- $f(z;\theta)$ is a neural network whose input is z and is parameterized by θ
- $\text{Ber}(\mu_x)$ and $\mathcal{N}(\mu_x, \sigma_c^2)$ are multivariate Bernoulli and Gaussian Distributions.

3.1 Factorization

Using the processes above factorize the joint probability $P(x, z, c)$ and explain how the network can compute each part of the factorized probability

Answer:

According to the generative process above, the joint probability $P(x, z, c)$ can be factorized as:

$$p(x, z, c) = p(x|z)p(z|c)p(c) = p(x|z) \cdot \mathcal{N}(z|\mu_c, \sigma_c^2 I) \cdot \text{Cat}(c|\pi)$$

And that $p(x|z) = \text{Ber}(x|\mu_x)$ or $\mathcal{N}(x|\mu_x, \sigma_x^2 I)$, each part of the factorized version is easily computable for the neural network.

3.2 ELBO

By maximizing $p(x)$ prove that ELBO for this model will be

$$\mathcal{L}_{\text{ELBO}}(\mathbf{x}) = \mathbb{E}_{q(z,c|x)} \left[\log p(\mathbf{x}|z) + \log p(z|c) + \log p(c) - \log q(z|x) - \log q(c|x)\right]$$

Note that we assumed a meanfield distribution for $q(c, z|x)$

Answer:

$$\log p(x) = \log \int_z \sum_c p(x, z, c)\, dz = \log \mathbb{E}_{q(z,c|x)} \left[\frac{p(x, z, c)}{q(z, c|x)}\right]$$

$$\geq \mathbb{E}_{q(z,c|x)} \left[\log \frac{p(x, z, c)}{q(z, c|x)}\right] = \mathcal{L}_{\text{ELBO}}(x) = \mathbb{E}_{q(z,c|x)} \left[\log p(x, z, c) - \log q(z, c|x)\right]$$

$$= \mathbb{E}_{q(z,c|x)} \left[\log p(x|z) + \log p(z|c) + \log p(c) - \log q(z|x) - \log q(c|x)\right]$$

Where in the last equality, we used the factorization from the last part, and another factorization in beacasue of a assumed meanfield distribution: $q(z, c|x) = q(z|x)q(c|x)$.

3.3 On derivation of $q(c_i \mid x)$

Prove that ELBO can be re expressed as below

$$\mathcal{L}_{\text{ELBO}}(\mathbf{x}) = \int_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}\, d\mathbf{z} - \int_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}) D_{KL}\left(q(\mathbf{c}|\mathbf{x}) \| p(\mathbf{c}|\mathbf{z})\right)\, d\mathbf{z}$$

Now explain how can we estimate $q(c|x)$, using the equation you derived?

Answer:

$$\mathcal{L}_{\text{ELBO}}(x) = \mathbb{E}_{q(z,c|x)} \left[\log \frac{p(x, z, c)}{q(z, c|x)}\right]$$

$$= \int_z \sum_c q(z, c|x) \log \frac{p(x, z, c)}{q(z, c|x)}\, dz = \int_z \sum_c q(z|x)q(c|x) \log \frac{p(x|z)p(c|z)p(z)}{q(z|x)q(c|x)}\, dz$$

$$= \int_z \sum_c q(z|x)q(c|x) \left[\log \frac{p(x|z)p(z)}{q(z|x)} + \log \frac{p(c|z)}{q(c|x)}\right]\, dz$$

$$= \int_z q(z|x) \log \frac{p(x|z)p(z)}{q(z|x)} \sum_c q(c|x)\, dz - \int_z q(z|x) \sum_c q(c|x) \log \frac{q(c|x)}{p(c|z)}\, dz$$

$$= \int_z q(z|x) \log \frac{p(x|z)p(z)}{q(z|x)}\, dz - \int_z q(z|x) D_{KL}(q(c|x) \parallel p(c|z))\, dz$$

Hence, to maximize ELBO, $D_{KL}(q(c|x) \parallel p(c|z)) = 0$ should be satisfied. As a result, we use the following equation to compute $q(c|x)$:

$$q(c|x) = p(c|z) = \frac{p(c)p(z|c)}{p(z)} = \frac{p(c)p(z|c)}{\sum_{c'} p(z|c')p(c')}$$

For SGVB we will be using the following

$$q(z|x) = \mathcal{N}(z; \tilde{\mu}, \tilde{\sigma} I)$$

3.4 A useful lemma for the next part

Prove that Given two multivariate Gaussian distributions $q(\mathbf{z}) = \mathcal{N}\left(\mathbf{z}; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\sigma}}^2\mathbf{I}\right)$ and $p(\mathbf{z}) = \mathcal{N}\left(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I}\right)$, we have:

$$\int q(\mathbf{z}) \log p(\mathbf{z}) d\mathbf{z} = \sum_{j=1}^{J} -\frac{1}{2}\log\left(2\pi\sigma_j^2\right) - \frac{\tilde{\sigma}_j^2}{2\sigma_j^2} - \frac{(\tilde{\mu}_j - \mu_j)^2}{2\sigma_j^2}$$

Where $\mu_j, \sigma_j, \tilde{\mu}_j$ and $\tilde{\sigma}_j$ simply denote the j^{th} element of $\boldsymbol{\mu}, \boldsymbol{\sigma}, \tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\sigma}}$, respectively, and J is the dimensionality of $\mathbf{z}$.

Answer:

$$\int q(z) \log p(z)\, dz = \int q(z) \log\left(\prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(z_j - \mu_j)^2}{2\sigma_j^2}\right)\right) dz =$$

$$= \int q(z) \sum_{j=1}^{J}\left(-\frac{1}{2}\log(2\pi\sigma_j^2) - \frac{(z_j - \mu_j)^2}{2\sigma_j^2}\right) dz = \sum_{j=1}^{J} -\frac{1}{2}\log(2\pi\sigma_j^2) \int q(z)\, dz - \int q(z) \frac{(z_j - \mu_j)^2}{2\sigma_j^2}\, dz$$

$$= \sum_{j=1}^{J} -\frac{1}{2}\log(2\pi\sigma_j^2) - \frac{1}{2\sigma_j^2} \int q(z) \left(z_j^2 + \mu_j^2 - 2\mu_j z_j\right) dz$$

$$= \sum_{j=1}^{J} -\frac{1}{2}\log(2\pi\sigma_j^2) - \frac{1}{2\sigma_j^2} \int q(z_j) \left(z_j^2 + \mu_j^2 - 2\mu_j z_j\right)\, dz_j$$

$$= \sum_{j=1}^{J} -\frac{1}{2}\log(2\pi\sigma_j^2) - \frac{1}{2\sigma_j^2}\mathbb{E}_q\left[z_j^2 - 2\mu_j z_j + \mu_j^2\right] = \sum_{j=1}^{J} -\frac{1}{2}\log(2\pi\sigma_j^2) - \frac{1}{2\sigma_j^2}(\tilde{\sigma}_j^2 + \tilde{\mu}_j^2 + \mu_j^2 - 2\mu_j\tilde{\mu}_j)$$

$$= \sum_{j=1}^{J} -\frac{1}{2}\log(2\pi\sigma_j^2) - \frac{\tilde{\sigma}_j^2}{2\sigma_j^2} - \frac{(\tilde{\mu}_j - \mu_j)^2}{2\sigma_j^2}$$

3.5 Putting everything together

Prove that using the SGVB estimator and the reparameterization trick, along with the result of the privous parts the ELBO(x) for a bernouli x can be rewritten as

$$\begin{aligned}\mathcal{L}_{\text{ELBO}}(\mathbf{x}) = &\frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{D}\left(x_i \log \mu_x^{(l)}|_i + (1 - x_i)\log(1 - \mu_x^{(l)}|_i)\right)\\ &- \frac{1}{2}\sum_{c=1}^{K}\gamma_c \sum_{j=1}^{J}\left(\log \sigma_c^2|_j + \frac{\tilde{\sigma}_j^2}{\sigma_c^2|_j} + \frac{(\tilde{\mu}_j - \mu_c|_j)^2}{\sigma_c^2|_j}\right)\\ &+ \sum_{c=1}^{K}\gamma_c \log\frac{\pi_c}{\gamma_c} + \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log\tilde{\sigma}_j^2\right)\end{aligned}$$

- L is the number of Monte Carlo Samples in SGVB estimator

- D is the dimensionality of $\mathbf{x}$ and μ_x
- x_i is the i^{th} element of $\mathbf{x}$
- J is the dimensionality of μ_c
- $\sigma_j^2, \mu_j, \tilde{\mu}_j, \tilde{\sigma}_j^2$ denote the j^{th} element of each variable
- K is the number of clusters
- π_c is the prior probability of cluster c
- γ_c denotes $q(c|\mathbf{x})$

Answer:

$$
\begin{aligned}
\mathcal{L}_{\text{ELBO}}(x) &= \mathbb{E}_{q(z,c|x)}\left[\log p(x|z) + \log p(z|c) + \log p(c) - \log q(z|x) - \log q(c|x)\right] \\
&= \mathbb{E}_{q(z,c|x)}\left[\log p(x|z)\right] + \mathbb{E}_{q(z,c|x)}\left[\log \frac{p(c)}{q(c|x)}\right] + \mathbb{E}_{q(z,c|x)}\left[\log p(z|c)\right] - \mathbb{E}_{q(z,c|x)}\left[\log q(z|x)\right] \\
&= \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] + \sum_{c=1}^{K} \gamma_c \log \frac{\pi_c}{\gamma_c} + \mathbb{E}_{q(z|x)}\left[\log p(z|c)\right] + \mathcal{H}(q(z|x)) \\
&= \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] + \sum_{c=1}^{K} \gamma_c \log \frac{\pi_c}{\gamma_c} + \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log(2\pi\sigma_j^2)\right) \\
&+ \mathbb{E}_{q(c|x)}\left[\sum_{j=1}^{J} -\frac{1}{2}\log(2\pi\sigma_{c,j}^2) - \frac{\tilde{\sigma}_j^2}{2\sigma_{c,j}^2} - \frac{(\tilde{\mu}_j - \mu_{c,j})^2}{2\sigma_{c,j}^2}\right] \\
&= \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] + \sum_{c=1}^{K} \gamma_c \log \frac{\pi_c}{\gamma_c} + \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log(\sigma_j^2)\right) \\
&- \frac{1}{2}\sum_{c=1}^{K}\sum_{j=1}^{J} \gamma_c \log(\sigma_{c,j}^2) + \frac{\tilde{\sigma}_j^2}{\sigma_{c,j}^2} + \frac{(\tilde{\mu}_j - \mu_{c,j})^2}{\sigma_{c,j}^2}
\end{aligned}
$$

And then we will calculate $\mathbb{E}_{q(z|x)}\left[\log p(x|z)\right]$ using Monte Carlo methods.

$$
\begin{aligned}
\mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] &\approx \frac{1}{L}\sum_{l=1}^{L} \log p(x|z_i) = \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{D} \log \mathrm{Ber}(x_i|\mu_{x,i}^{(l)}) \\
&= \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{D}\left(x_i \log \mu_{x,i}^{(l)} + (1 - x_i)\log(1 - \mu_{x,i}^{(l)})\right)
\end{aligned}
$$

4 Vector Quantized Variational Autoencoders (VQ-VAE) *(Bonus)

Answer the below questions.

1. Why VQ-VAE uses a discrete latent model instead of a continues one? Give an example of a senorio where discrete latent is better that continues one?
2. How a discrete latent variable is modeled in VQ-VAE?
3. How gradient can be propagated to the input through the discrete latent variable?
4. How the dictionary of VQ-VAE updates?