Deep Generative Models
Homework Set 2 (Sol)

Mohammad Parsa Dini

November 1, 2024

Problem 1

A. (a) In this context, we first show that we can exactly compute the KL-Divergence for both distributions. However, for any x, if we are lucky enough for $p(x|\theta)$ and $q(x|\Phi)$ to have the same support set $\chi \subset \mathbb{R}$, this KL-Divergence is well-defined and thus finite.
Since we have the samples for both distributions, we can change the expectation to a sum over the observed samples. Furthermore, since we had access to the density of both distributions, we know both $p(x_i|\theta)$ and $q(x_i|\Phi)$ for any x_i.

$$D_{KL}(p(x|\theta) \parallel q(x|\Phi)) = \mathbb{E}_{x\sim p(x|\theta)}\left[\log\left(\frac{p(x|\theta)}{q(x|\Phi)}\right)\right] \approx \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{p(x_i|\theta)}{q(x_i|\Phi)}\right)$$

$$D_{KL}(q(x|\Phi) \parallel p(x|\theta)) = \mathbb{E}_{x\sim q(x|\Phi)}\left[\log\left(\frac{q(x|\phi)}{p(x|\theta)}\right)\right] \approx \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{q(x_i|\Phi)}{p(x_i|\theta)}\right)$$

So in order to optimize the KL-Divergence between these two distributions, we can use reparametrization tricks in order to use gradient descent algorithms and do back propagation in finding the optimum Φ and since the terms above are tractable and differentiable with respect to Φ, we are all good.
So our choice will be: C. We can minimize both $D_{KL}(q(x|\Phi) \parallel p(x|\theta))$ and $D_{KL}(p(x|\theta) \parallel q(x|\Phi))$.

(b) In this context, since we will show that $D_{KL}(q(x|\Phi) \parallel p(x|\theta))$ can't be computed since we don't have the density of $p(x|\theta)$, we have to approximate it using Monte-Carlo estimation. Additionally, we don't have the density of $p(x|\theta$ so we can't evaluate the density for iid observed samples coming from $q(x|\Phi)$.
In contrast to this, we will show that $D_{KL}(p(x|\theta) \parallel q(x|\Phi))$ can be computed and since it is differentiable with respect to Φ, we can minimize it.

$$\begin{aligned}
\arg\min_{\Phi} D_{KL}\left(p(x|\theta) \parallel q(x|\Phi)\right) &= \arg\min_{\Phi} \mathbb{E}_{x\sim p(x|\theta)}\left[\log\left(\frac{p(x|\theta)}{q(x|\Phi)}\right)\right] \\
&= \arg\min \mathbb{H}(q(x|\Phi)) + \mathbb{E}_{x\sim p(x|\theta)}\left[\log\left(\frac{1}{q(x|\Phi)}\right)\right] \\
&= \arg\min \mathbb{E}_{x\sim p(x|\theta)}\left[\log\left(\frac{1}{q(x|\Phi)}\right)\right] \approx \arg\min \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{q(x_i|\Phi)}\right)
\end{aligned}$$

Here, we have access to samples which means we can estimate the expectation with a sum over the observed samples. Furthermore, we can evaluate the $\log\left(\frac{1}{q(x_i|\Phi)}\right)$ at any x_i. Finally, since this function is differentiable, we can minimize it using gradient descent algorithms, ... for instance.
So our choice will be: B. We can minimize $D_{KL}(p(x|\theta) \parallel q(x|\Phi))$.

(c) This context is a bit critical. we assume that distribution $p(x|\theta)$ is as $p(x|\theta) = \frac{f(x;\theta)}{g(\theta)}$ where $g(\theta)$ is the normalization factor of distribution as $g(\theta) = \int_\chi f(x;\theta)dx$. We will show that both KL-Divergences can be minimized since the knowledge of unnormalized part of $p(.)$ which is $f(x;\theta)$ can still help us.
We wish to minimize $\mathbb{E}_{x\sim p}\left[\log\left(\frac{p(x|\theta)}{q(x|\Phi)}\right)\right] = \int_\chi p(x|\theta)\log\left(\frac{p(x|\theta)}{q(x|\Phi)}\right)dx = \frac{1}{g(\theta)}\int_\chi f(x;\theta)\log\left(\frac{f(x;\theta)}{q(x|\Phi)g(\theta)}\right)dx$ which results in

$$\begin{aligned}\mathbb{E}_{x\sim p}\left[\log\left(\frac{p(x|\theta)}{q(x|\Phi)}\right)\right] &= \frac{1}{g(\theta)}\int_\chi f(x;\theta)\log\left(\frac{f(x;\theta)}{q(x|\Phi)}\right)dx - \frac{\log\left(g(\theta)\right)}{g(\theta)}\int_\chi f(x;\theta)dx \\ &= \frac{1}{g(\theta)}\int_\chi f(x;\theta)\log\left(\frac{f(x;\theta)}{q(x|\Phi)}\right)dx - \log\left(g(\theta)\right) \\ &= \frac{\mathbb{H}[f(x;\theta)]}{g(\theta)} - \int_\chi p(x|\theta)\log\left(q(x|\Phi)\right)dx - \log\left(g(\theta)\right)\end{aligned}$$

Thus, $\arg\min_\Phi \mathbb{E}_{x\sim p}\left[\log\left(\frac{p(x|\theta)}{q(x|\Phi)}\right)\right] = \arg\min_\Phi -\int_\chi f(x;\theta)\log\left(q(x|\Phi)\right)dx = \arg\max_\Phi \mathbb{E}_{x\sim p}\left[\log\left(q(x|\Phi)\right)\right]$.
Now this optimization as mentioned in the class can be done using Reinforce algorithm or reparametrization tricks.

For the other KL-Divergence, we have access to the samples of $q(x|\Phi)$ which means we can estimate $D_{KL}(q(x|\phi)\parallel p(x|\theta))$ with Monte-Carlo simulation. However, having no knowledge of $g(\theta)$ will not be an issue since we want to do the optimization with respect to Φ.

$$\begin{aligned}D_{KL}(q(x|\phi)\parallel p(x|\theta)) &= \mathbb{E}_{x\sim q}\left[\log\left(\frac{q(x|\Phi)}{p(x|\theta)}\right)\right] \\ &= \mathbb{E}_{x\sim q}\left[\log\left(\frac{q(x|\Phi)}{f(x;\theta)}\right) + \log\left(g(\theta)\right)\right] \\ &= \mathbb{E}_{x\sim q}\left[\log\left(\frac{q(x|\Phi)}{f(x;\theta)}\right)\right] + \log\left(g(\theta)\right)\end{aligned}$$

Therefore, $\arg\min_\Phi D_{KL}(q(x|\phi)\parallel p(x|\theta)) = \arg\min_\Phi \mathbb{E}_{x\sim q}\left[\log\left(\frac{q(x|\Phi)}{f(x;\theta)}\right)\right]$ and in order to minimize $\mathbb{E}_{x\sim q}\left[\log\left(\frac{q(x|\Phi)}{f(x;\theta)}\right)\right]$ with respect to Φ again we can use Reinforce algorithm or reparametrization tricks.
So our choice will be: C. We can minimize both $D_{KL}(q(x|\Phi)\parallel p(x|\theta))$ and $D_{KL}(p(x|\theta)\parallel q(x|\Phi))$.

B. (a) We wish to minimize the KL-Divergence between the mixture of a Gaussian $p(x|\theta)$ and a Gaussian $q(x|\Phi)$. For$D_{KL}(q(x|\phi)\parallel p(x|\theta))$, since the expectation is with respect to $q(x|\theta)$, then $q(x|\Phi$ will tend to cover the modes(regions with the highest probability density) of $p(x|\Phi)$ and since it can minimize it when sticking to one of the modes(having the same mean), it will have two local minima but no global minima, since $q(x|\Phi)$ can't capture (aka Mode-Seeking) both modes of a bi-mode Gaussian.
Thus, our answer is: Two local minima, but no global minima.

(b) Minimizing $D_{KL}(p(x|\theta)\parallel q(x|\phi))$ will yield $q(x|\phi)$ to be probable wherever $p(x|\theta)$ is highly probable. it will tend to stick to one of the modes. Furthermore, if this unknown mean lies between these two modes(Mode-Coverage), then it minimizes the divergence between both modes to some extent since no mode is favorable as our bi-mode distribution is symmetric.
Thus, our answer is: One local minima, One global minima.

(c) **Mode-Seeking (minimizing KL$(q\|p)$)**: In this scenario, $q(x|\phi)$ aims to concentrate its probability mass on one of the modes of $p(x|\theta)$, rather than representing the entire distribution. This occurs because the penalty for q not covering the low-probability regions of p is minimal. Therefore, q only "seeks" one of the modes of p that results in the lowest divergence.
Mode-Coverage (minimizing KL$(p\|q)$): Here, $q(x|\phi)$ is driven to encompass all high-probability regions of $p(x|\theta)$. Any region where $p(x|\theta)$ has a significant probability mass that $q(x|\phi)$ does not

cover incurs a large penalty. Thus, q tries to "cover" both modes by positioning itself between them.

Problem 2

We know that $\epsilon \sim \mathcal{N}(0, I)$), and $z = \mu + L\epsilon$ is a linear combination of ϵ. Thus, z will also follow a Normal distribution with parameters $(\mu, \Sigma = LL^T)$. The mean of z is also μ since

$$\mathbb{E}[z] = \mathbb{E}[\mu + L\epsilon] = \mu + L\mathbb{E}[\epsilon] = \mu.$$

Furthermore, we can find $\Sigma = LL^T$ as well:

$$\Sigma = \mathbb{E}[zz^T] - \mu\mu^T = \mathbb{E}[(\mu + L\epsilon)(\mu + L\epsilon)^T] = \mathbb{E}[L\epsilon\epsilon^T L^T + \mu\epsilon^T L^T + L\epsilon\mu^T] = L\mathbb{E}[\epsilon\epsilon^T]L^T = LIL^T = LL^T.$$

Since the LL^T matrix is the covariance matrix of random variable z, in order to assure that the main diagonal of Σ is always non-negative (as it must be since $\Sigma(i,i) = \text{Cov}(z_i, z_i) = \text{Var}(z_i) \geq 0$), Since L is upper-triangular, it requires $LL^T = \Sigma$ to have non-negative diagonal elements.
This is the Reparametrization trick, when we have a random node z in a neural network with some parameters (μ, Σ). As we want to minimize some objective $\mathcal{L}$, we need back propagation with respect to node z which is impossible since $\frac{\partial \mathcal{L}}{\partial z}$ is nonsense since we can only back propagate for deterministic nodes. As a result, we do this trick by adding a random variable ϵ and then back propagate using the gradient $\frac{\partial \mathcal{L}}{\partial \Sigma}$ and $\frac{\partial \mathcal{L}}{\partial \mu}$.
In order to choose a differentiable and positive activation functions many use the exponential activation function or ReLU for Σ.

Problem 3

A. in order to get the joint probability $p(x, c, z)$ we can consider them as Markov chains and write them conditionally as :

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

B. We start by defining ELBO first:

$$\log(p(x)) = \log \sum_{c \in \mathcal{C}} \int_{\mathcal{Z}} p(x, c, z)dz = \log \left(\mathbb{E}_{c,z \sim q(c,z)} \left[\frac{p(x, c, z)}{q(c, z)} \right] \right)$$

By Jensen inequality we can obtain:

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

So now we will rewrite the ELBO(x):

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

(a) $\mathbb{E} \log p(x|z)$ tells us how the well-defined the latent space is.

(b) $\mathbb{E} \log p(z|c)$ is the prior of latent space over the cluster c.

(c) $\mathbb{E} \log p(c)$ is the prior on the classes.

(d) $\mathbb{E} \log q(z|x)$ is the variational posterior over z.

(e) $\mathbb{E}\log q(c|x)$ is the variational posterior over c.

C. Firstly, we notice that $p(z|c)p(c) = p(z,c) = p(z)p(c|z)$. plugging this into the result form above will imply that:

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

Now we notice that $q(c,z|x) = q(c|x)q(z|x)$ and also $\sum_{c\in\mathcal{C}} q(c|x) = 1$, Thus, we get:

$$\begin{aligned}
\mathcal{L}_{\text{ELBO}}(x) &= \mathbb{E}_{c,z\sim q(c,z|x)}\left[\log\left(\frac{p(x|z)p(z)}{q(z|x)}\right) + \log\left(\frac{p(c|z)}{q(c|x)}\right)\right] \\
&= \int_{z\in\mathcal{Z}} q(c|z)q(z|x)\sum_{c\in\mathcal{C}} q(c|z)\log\left(\frac{p(c|z)}{q(c|x)}\right)dz + \int_{z\in\mathcal{Z}} q(z|x)\log\left(\frac{p(x|z)p(z)}{q(z|x)}\right)\sum_{c\in\mathcal{C}} q(c|x)dz \\
&= \int_{z\in\mathcal{Z}} q(z|x)D_{KL}(p(c|z)\parallel q(c|x))dz + \int_{z\in\mathcal{Z}} q(c|x)q(z|x)\log\left(\frac{p(x|z)p(z)}{q(z|x)}\right)dz
\end{aligned}$$

So Now in order to maximize ELBO, we need to set $D_{KL}(p(c|z)\parallel q(c|x)) = 0$ since the kl-divergence is always positive. Thus we get(However this is our ideal):

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

Therefore, we estimate $q(c|x)$ by applying Bayes Rule on $p(c|z)$ which is:

$$q(c|x) = \frac{p(z|c)p(c)}{p(z)} = \frac{cat(c|\pi)\mathcal{N}(\mu_c,\sigma_c^2 I)}{\sum_{c'\in\mathcal{C}} cat(c'|\pi)\mathcal{N}(\mu_{c'},\sigma_{c'}^2 I)} = \frac{\pi_c\mathcal{N}(\mu_c,\sigma_c^2 I)}{\sum_{c'\in\mathcal{C}}\pi_{c'}\mathcal{N}(\mu_{c'},\sigma_{c'}^2 I)}$$

D. Firstly, we compute $\mathbb{E}_{z\sim q}[\log p(z)]$:

$$\mathbb{E}_{z\sim q}[\log p(z)] = \mathbb{E}_{z\sim q}\left[-\frac{J}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(z-\mu)^T(z-\mu)\right]$$

Now we compute $\mathbb{E}_{z\sim q}\left[(z-\mu)^T(z-\mu)\right]$ and we note that $\mathbb{E}_{z\sim q}[z-\tilde{\mu}] = 0$:

$$\begin{aligned}
\mathbb{E}_{z\sim q}\left[(z-\mu)^T(z-\mu)\right] &= \mathbb{E}_{z\sim q}\left[(z-\tilde{\mu}+\tilde{\mu}-\mu)^T(z-\tilde{\mu}+\tilde{\mu}-\mu)\right] \\
&= \mathbb{E}_{z\sim q}\left[(z-\tilde{\mu})^T(z-\tilde{\mu}) + 2(z-\tilde{\mu})^T(\tilde{\mu}-\mu) + (\tilde{\mu}-\mu)^T(\tilde{\mu}-\mu)\right] \\
&= \sum_{j=1}^{J}\tilde{\sigma_j^2} + (\mu_j - \tilde{\mu_j})^2
\end{aligned}$$

Eventually it follows that:

$$\mathbb{E}_{z\sim q(z)}[\log p(z)] = \sum_{j=1}^{J}[\frac{-1}{2}\log(2\pi\sigma_j^2) - \frac{\tilde{\sigma_j}^2 + (\mu_j - \tilde{\mu_j})^2)}{2\sigma_j^2}]$$

E. Firstly we note that $x|z \sim \text{Ber}(\mu_x)$. Additionally, $q(c|x) = \gamma_c$ and $p(c) = \pi_c$.

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

Now we will extend the first term $\mathbb{E}_{q(c,z|x)}[\log(p(x|z))]$:

$$\begin{aligned}
\mathbb{E}_{q(c,z|x)}[\log(p(x|z))] &= \mathbb{E}_{q(z|x)}[\log(p(x|z))] \\
&= \mathbb{E}_{q(z|x)}\left[\prod_{j=1}^{D}\log\text{Ber}(x_j|\mu_x)\right] \\
&= \frac{1}{L}\sum_{l=1}^{L}\sum_{j=1}^{D}[x\log(\mu_{x_i}^l) + (1-x_i)\log(1-\mu_{x_i}^l)]
\end{aligned}$$

Now the second term $\mathbb{E}_{q(c,z|x)}[\log \frac{p(c)}{q(c|x)}]$:

$$\mathbb{E}_{q(c,z|x)}[\log \frac{p(c)}{q(c|x)}] = \sum_{c\in\mathcal{C}} \gamma_c \log(\frac{\pi_c}{\gamma_c})$$

For simplifying the third term $\mathbb{E}_{q(c,z|x)}[\log p(z|c)]$, we use the result of the last section, since both distributions have Normal distributions (we must take expectation over the clusters as well):

$$\begin{aligned}\mathbb{E}_{q(c,z|x)}[\log p(z|c)] &= \mathbb{E}_{q(c|x)}\left[\sum_{j=1}^{J}\left(\frac{-1}{2}\log(2\pi{\sigma_{jc}}^2) - \frac{\tilde{\sigma_{jc}}^2 + (\mu_j - \tilde{\mu_{jc}})^2}{2{\sigma_{jc}}^2}\right)\right] \\ &= \sum_{c\in\mathcal{C}}\gamma_c\left[\sum_{j=1}^{J}\left(\frac{-1}{2}\log(2\pi{\sigma_{jc}}^2) - \frac{\tilde{\sigma_{jc}}^2 + (\mu_j - \tilde{\mu_{jc}})^2}{2{\sigma_{jc}}^2}\right)\right]\end{aligned}$$

At last step we know that for a Normal distribution $\mathcal{P}_§$ with mean μ and variance σ^2, its entropy is $\mathbb{H}(\mathcal{P}_§) = \frac{1}{2}\log(1+\sigma^2)$, thus we get:

$$\mathbb{E}_{q(c,z|x)}[\log p(z|x)] = \frac{1}{2}\log(1+\tilde{\sigma}^2)$$

Eventually summing up these 5 results together well lead to the proof:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}}(x) = &\frac{1}{L}\sum_{l=1}^{L}\sum_{j=1}^{D}[x\log(\mu^l_{x_i}) + (1-x_i)\log(1-\mu^l_{x_i})] \\ &+ \mathbb{E}_{q(c,z|x)}\left[\log\frac{p(c)}{q(c|x)}\right] \\ &+ \sum_{c\in\mathcal{C}}\gamma_c\log\left(\frac{\pi_c}{\gamma_c}\right) \\ &+ \sum_{c\in\mathcal{C}}\gamma_c\left[\sum_{j=1}^{J}\left(\frac{-1}{2}\log(2\pi{\sigma_{jc}}^2) - \frac{\tilde{\sigma_{jc}}^2 + (\mu_j - \tilde{\mu_{jc}})^2}{2{\sigma_{jc}}^2}\right)\right] \\ &- \frac{1}{2}\log(1+\tilde{\sigma}^2)\end{aligned}$$

1 Problem 4

A. The discrete latent space in "VQ-VAE" consists of embedding vectors, often called the codebook or dictionary, denoted as $e = \{e_1, e_2, ..., e_K\}$, where K represents the number of discrete embeddings. By mapping each input to a discrete embedding, VQ-VAE captures distinct patterns in the data. This quantization operation replaces the continuous latent vector $z_e(x)$ with the nearest discrete embedding, $z_q(x)$, which becomes the latent representation of x.
For many cases when a compressed form of data suffices like speech or for cases where we keep the data as discrete like text, the VQ-VAE is much better. Furthermore, with a Quantizae VAE, We wouldn't have the Decoding which was stochastic, since we assign the nearest quantized vector in our codebook to the input vector. Like the traditinal VAE, the encoder is still stochastic.

B. The encoder's continuous output, $z_e(x)$, is mapped to the closest embedding in the codebook, effectively creating a discrete latent representation:

$$z_q(x) = e_{j^*}$$

where $j^* = \arg\min_i \|z_e(x) - e_i\|^2$. Here, j^* is the index of the closest codebook embedding. This process discretizes the latent space, resulting in a finite number of possible states—each corresponding to one of the embedding vectors in the codebook. By quantizing the continuous output of the encoder, VQ-VAE ensures that the latent space is represented by a set of discrete, predefined embeddings.

C. The quantization step $z_e(x) \rightarrow z_q(x)$ is inherently non-differentiable. VQ-VAE employs a straight-through estimator to approximate the gradients. Specifically, it bypasses the quantizer entirely during backpropagation:

$$\frac{\partial L}{\partial z_e(x)} \approx \frac{\partial L}{\partial z_q(x)}$$

This identity operation allows the gradients to flow smoothly from the decoder's output back to the encoder, effectively sidestepping the non-differentiable nature of the quantization step. This approach ensures that the entire model remains trainable using standard gradient-based optimization techniques.

D. The codebook embeddings e_i are updated using an exponential moving average (EMA) approach. For each embedding e_i selected in the quantization step, its value is updated as follows:

$$e_i \leftarrow \gamma e_i + (1 - \gamma) z_e(x)$$

where γ is a decay parameter (e.g., 0.99), and $z_e(x)$ is the encoder output assigned to e_i. This moving average update ensures that the codebook gradually adapts to represent typical encoder outputs, improving both the stability and the alignment of the codebook with the latent space structure.