

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

**Matin M.babaei 400102114****1 D_{kl} Intuitions****1.1 Which One?**

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

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

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

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:

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

این جیسا، ممکن است این دو مسیر را داشتیم: **کلینیکی** = **Maxim.**

پس از $\nabla_{\phi} \text{KL}(q(x | \phi), p(x | \theta))$ و $\nabla_{\phi} \text{KL}(p(x | \theta), q(x | \phi))$

$$\nabla_{\phi} \text{KL}(p(x | \theta) || q(x | \phi)) = \mathbb{E}_{q(x | \phi)} \left[\log \frac{p(x | \theta)}{q(x | \phi)} \right] \stackrel{\text{MC}}{=} \frac{1}{N} \sum_{i=1}^N \log \left(\frac{p(x_i | \theta)}{q(x_i | \phi)} \right)$$

$$\nabla_{\phi} \text{KL}(q(x | \phi) || p(x | \theta)) = \mathbb{E}_{p(x | \theta)} \left[\log \frac{q(x | \phi)}{p(x | \theta)} \right] \stackrel{\text{MC}}{=} \frac{1}{N} \sum_{i=1}^N \log \left(\frac{q(x_i | \phi)}{p(x_i | \theta)} \right)$$

Both are ∇ able and tractable w.r.t $\phi \Rightarrow$ ϕ \rightarrow $\text{arg min}_{\phi} \text{KL}$

\rightarrow B \rightarrow Proof

$$KL(p_{\text{model}} \parallel q_{\text{model}}) = E_{\text{model}} \left[\log \frac{p_{\text{model}}}{q_{\text{model}}} \right] = E_{\text{model}} \left[\log p_{\text{model}} \right] - E_{\text{model}} \left[\log q_{\text{model}} \right] \quad \checkmark$$

سید مختار نایم

تابع حفاظتی
نایم

الآن الگوریتم

$$KL(q_{\text{model}} \parallel p_{\text{model}}) = E_{\text{model}} \left[\log \frac{q_{\text{model}}}{p_{\text{model}}} \right] = E_{\text{model}} \left[\log q_{\text{model}} \right] - E_{\text{model}} \left[\log p_{\text{model}} \right] \quad \times$$

تابع پیوسته نایم

الآن الگوریتم

$$p_{\text{model}} = g(\theta) \quad \tilde{p}_{\text{model}}$$

\rightarrow C \rightarrow Proof

$$KL(p_{\text{model}} \parallel q_{\text{model}}) = E_{\text{model}} \left[\log p_{\text{model}} \right] - E_{\text{model}} \left[\log q_{\text{model}} \right]$$

لطفاً \rightarrow Proof \rightarrow \checkmark

$$KL(q_{\text{model}} \parallel p_{\text{model}}) = E_{\text{model}} \left[\log q_{\text{model}} \right] - E_{\text{model}} \left[\log p_{\text{model}} \right]$$

Or we say:

$$D_{KL}(q_{\text{model}} \parallel p_{\text{model}}) = E_{\text{model}} \left[\log \frac{q_{\text{model}}}{p_{\text{model}}} \right]$$

$$= E_{\text{model}} \left[\log \left(\frac{q_{\text{model}}}{\tilde{p}_{\text{model}}} \right) + \log \left(\frac{\tilde{p}_{\text{model}}}{p_{\text{model}}} \right) \right] = E_{\text{model}} \left[\log \left(\frac{q_{\text{model}}}{\tilde{p}_{\text{model}}} \right) + \log \left(\frac{\tilde{p}_{\text{model}}}{p_{\text{model}}} \right) \right]$$

پس \rightarrow Proof \rightarrow Proof \rightarrow Reparametrization \rightarrow \checkmark

$$\text{Arg min } D_{KL}(q_{\text{model}} \parallel p_{\text{model}}) \quad \checkmark$$

• mean KL is convex in θ

$$\begin{aligned} E_{\sim p(\mathbf{x}|\theta)} \left[\log \frac{p(\mathbf{x}|\theta)}{q(\mathbf{x}|\theta)} \right] &= \int_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x}|\theta) \log \left(\frac{p(\mathbf{x}|\theta)}{q(\mathbf{x}|\theta)} \right) d\mathbf{x} \\ &= \int_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}) \tilde{p}(\mathbf{x}|\theta) \log \left(\frac{\tilde{p}(\mathbf{x}|\theta)}{q(\mathbf{x}|\theta)} \right) d\mathbf{x} \\ &= g(\theta) \left(\int_{\mathbf{x}} \tilde{p}(\mathbf{x}|\theta) \log \left(\frac{\tilde{p}(\mathbf{x}|\theta)}{q(\mathbf{x}|\theta)} \right) d\mathbf{x} + \int_{\mathbf{x}} \tilde{p}(\mathbf{x}|\theta) \log(g(\mathbf{x})) d\mathbf{x} \right) \\ &= g(\theta) \left(\int_{\mathbf{x}} \tilde{p}(\mathbf{x}|\theta) \log \left(\frac{\tilde{p}(\mathbf{x}|\theta)}{q(\mathbf{x}|\theta)} \right) d\mathbf{x} + \log(g(\theta)) \underbrace{\int_{\mathbf{x}} \tilde{p}(\mathbf{x}|\theta) d\mathbf{x}} \right) \\ &= g(\theta) \int_{\mathbf{x}} \tilde{p}(\mathbf{x}|\theta) \log \left(\frac{\tilde{p}(\mathbf{x}|\theta)}{q(\mathbf{x}|\theta)} \right) d\mathbf{x} + \log(g(\theta)) \\ &= g(\theta) H[\tilde{p}(\mathbf{x}|\theta) - \int_{\mathbf{x}} \tilde{p}(\mathbf{x}|\theta) \log(g(\mathbf{x})) d\mathbf{x} + \log(g(\theta))] \end{aligned}$$

Thus,

$$\arg \min_{\theta} E \left[\log \left(\frac{p(\mathbf{x}|\theta)}{q(\mathbf{x}|\theta)} \right) \right] = \arg \min_{\theta} E \left[\log(g(\mathbf{x}|\theta)) \right]$$

• Indicates that subjective vs. objective is consistent \Rightarrow

1.2 Mode-Seeking or Mode-Coverage

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

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

- (a) If we choose to minimize $KL(q(x | \phi), p(x | \theta))$, 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 $KL(p(x | \theta), q(x | \phi))$, 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"?

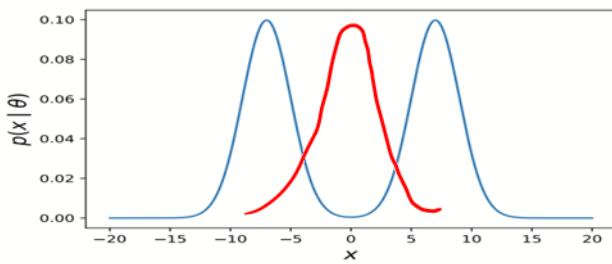
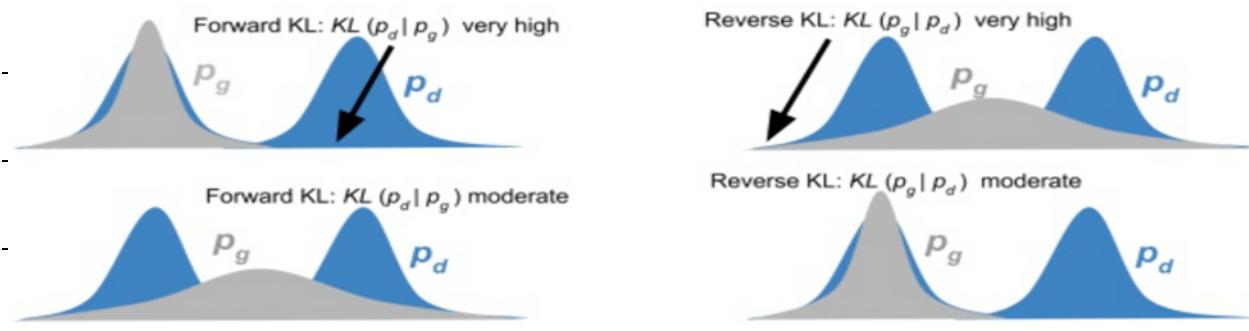


Figure 1: A two-mode Gaussian mixture

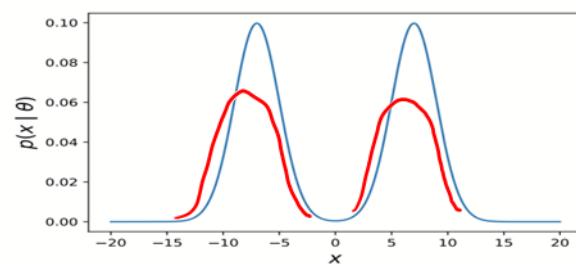


Figure 1: A two-mode Gaussian mixture

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.

$$z \sim \mathcal{N}(\mu, \Sigma) \rightarrow z = \mu + L \epsilon \Rightarrow z \sim \mathcal{N}(\mu, \Sigma - \Lambda \Lambda^T)$$

$$\mu = E[Z] = E[\mu + L\epsilon] = E[\mu] + E[L\epsilon] = \mu + L E[\epsilon] = \mu$$

$$\Sigma = E[22^T] - \mu\mu^T = E[(\mu + L\epsilon)(\mu + L\epsilon)^T] - \mu\mu^T$$

$$= E[MM^T + \mu \epsilon L^T + L \epsilon \epsilon^T L^T + L \epsilon M^T] = E[MM^T] + E[\mu \epsilon L^T] + E[L \epsilon \epsilon^T L^T] + E[L \epsilon M^T]$$

$$= L E[cc^T] L^T - L I L^T = LL^T$$

لے کر سرسری میں اسی طبقے کی مدد سے اپنے ایجاد کی تحریک کی جائے گی۔

$$\Sigma = L O L^T \quad (O = I) \quad : \quad \text{Inv. in } \Sigma \quad \Sigma \text{ da}$$

Note: $S_{\text{var}} = \text{Cov}(Z, Z) = \text{Var}(Z) \geq 0$

نہایت ل بی ملائک حارثہ نسبت نہیں۔

بن ترتیب این ایجاد شده است و بازگشتی برای Backprop است.

لأن $\frac{\partial L}{\partial \Sigma}$, $\frac{\partial L}{\partial \mu}$ متساوي

خواسته خود را در میان افراد خود می خواهد

لیکن ایک اسٹریچ ایکسپریس کے مقابلے میں ایک ایکسپریس کی سرعت کو کم کر دیتی ہے اس کا نام "softplus" ہے اس کا معنی "لٹھنے" ہے (لٹھنے کا دلیل یہ ہے کہ اس کی سرعت کو کم کر کے اس کو ایک اسٹریچ ایکسپریس کے مقابلے میں ایک ایکسپریس کی سرعت کے مقابلے میں بھی کم کر دیا جائے گا) اس کی سرعت کو کم کر کے اس کو ایک اسٹریچ ایکسپریس کے مقابلے میں ایک ایکسپریس کی سرعت کے مقابلے میں بھی کم کر دیا جائے گا۔

3 Gaussian Mixture Latent Model

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

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.

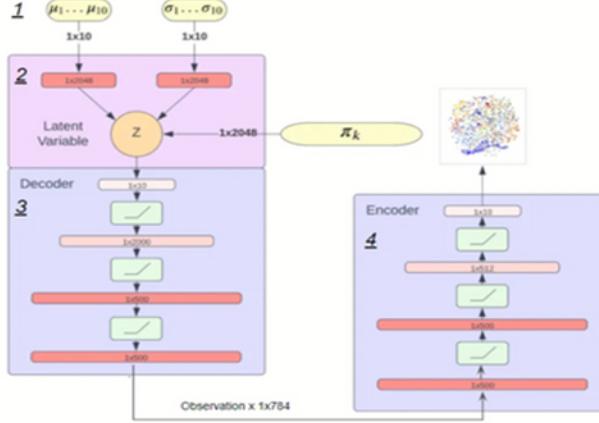


Figure 2: model architecture

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 precisely:

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 :
 - If x is binary:
 - Compute the expectation vector μ_x :
$$\mu_x = f(z; \theta)$$
 - Choose a sample $x \sim \text{Ber}(\mu_x)$
 4. If x is real-valued:
 - Compute μ_x and σ_x^2 :

$$[\mu_x; \log \sigma_x^2] = f(z; \theta)$$
 - 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 $\mathcal{N}(\mu_x, \sigma_x^2)$ 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

$$P(x, c, z) = P(c) P(z|c) P(x|z, c) = P(c) P(z|c) P(x|z)$$

$$P(c) = \text{Cat}(\pi) \quad P(z|c) = \mathcal{N}(\mu_c, \sigma_c^2 I) \quad \begin{matrix} \text{thus} \\ \text{Zero} \\ \downarrow \end{matrix} \quad \begin{matrix} \text{Ber}(\mu_c) \\ \text{Bin} \end{matrix}$$

Thus:

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

3.2 ELBO

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

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

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

$$\log p(x) = \log \int_z \sum_c p(x, z, c) dz$$

Given $q(z, c|x)$ with z and c independent

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

Note:

$$q(z, c|x) = q(z|x) q(c|x)$$

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

3.3 On derivation of $q(c_i | x)$

Prove that ELBO can be re expressed as below

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

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

$$\mathcal{L}_{\text{ELBO}} = E_{c \sim q(c|z)} \left[\log \frac{p(c)P(c|z)p(z)}{q(z|c)} \right] \xrightarrow{P(c)p(z) = P(z)p(c|z)} E_{c \sim q(c|z)} \left[\log \frac{P(c|z)P(z)p(c|z)}{q(z|c)} \right]$$

$$= E_{c \sim q(c|z)} \left[\log \frac{P(c|z)P(z)}{q(z|c)} + \log \frac{p(c|z)}{q(z|c)} \right]$$

$$\rightarrow q(z|c) = q(z|c) + q(z|c)$$

$$= \int_{z \in Z} q(z|c) \log \left(\frac{P(c|z)P(z)}{q(z|c)} \right) \sum_{c \in C} q(z|c) dz + \int_{z \in Z} q(z|c) \sum_{c \in C} q(z|c) \log \left(\frac{P(c|z)}{q(z|c)} \right) dz$$

$$= \int_{z \in Z} q(z|c) \log \left(\frac{P(c|z)P(z)}{q(z|c)} \right) + \int_{z \in Z} q(z|c) \left(- \sum_{c \in C} q(z|c) \log \left(\frac{q(z|c)}{P(c|z)} \right) \right) dz$$

$$= \int_{z \in Z} q(z|c) \log \left(\frac{P(c|z)P(z)}{q(z|c)} \right) dz - \int_{z \in Z} q(z|c) D_{KL}(q(z|c) \| P(c|z)) dz$$

$$\text{Thus: } \mathcal{L}_{\text{ELBO}} = \int_{z \in Z} q(z|c) \log \left(\frac{P(c|z)P(z)}{q(z|c)} \right) dz - \int_{z \in Z} q(z|c) D_{KL}(q(z|c) \| P(c|z)) dz$$

$$\text{Since } D_{KL} \geq 0 \rightarrow \text{To maximize ELBO} \rightarrow D_{KL}(q(z|c) \| P(c|z)) = 0$$

$$\rightarrow P(c|z) = q(z|c) = \frac{P(c)P(z|c)}{P(z)} \rightarrow \text{So, } \frac{P(c)P(z|c)}{P(z)} = \pi_c$$

$$\frac{q(z|c)}{P(z)} = \frac{P(c)P(z|c)}{P(z)} = \frac{N(\mu_c, \sigma_c^2 I) \text{Cat}(c|\pi)}{\sum_{c' \in C} N(\mu_{c'}, \sigma_{c'}^2 I) \text{Cat}(c'|\pi)} = \pi_c$$

3.4 A useful lemma for the next part

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

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

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} .

$$q(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\sigma}}^2 \mathbf{I}), \quad p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2 \mathbf{I})$$

$$\text{we know } \underset{\mathbf{z} \sim q(\mathbf{z})}{E} \log(p(\mathbf{z})) = \int q(\mathbf{z}) \log p(\mathbf{z}) d\mathbf{z}$$

$$\underset{\mathbf{z} \sim q(\mathbf{z})}{E} [\log(p(\mathbf{z}))] = \underset{\mathbf{z} \sim q(\mathbf{z})}{E} \left[\frac{J}{r} \log\left(\frac{1}{r\sigma^2}\right) - \frac{1}{r\sigma^2} (\mathbf{z} - \boldsymbol{\mu})^T (\mathbf{z} - \boldsymbol{\mu}) \right]$$

$$= \underset{\mathbf{z} \sim q(\mathbf{z})}{E} \left[-\frac{J}{r} \log(r\sigma^2) - \frac{1}{r\sigma^2} (\mathbf{z} - \boldsymbol{\mu})^T (\mathbf{z} - \boldsymbol{\mu}) \right]$$

: push (use u's in previous step)

$$\underset{\mathbf{z} \sim q(\mathbf{z})}{E} \left[-\frac{1}{r\sigma^2} (\mathbf{z} - \boldsymbol{\mu})^T (\mathbf{z} - \boldsymbol{\mu}) \right] = ?$$

Recall that $\underset{\mathbf{z} \sim q(\mathbf{z})}{E} [\mathbf{z}] = \tilde{\boldsymbol{\mu}}$, so we should try to appear $\tilde{\boldsymbol{\mu}}$.

$$\underset{\mathbf{z} \sim q(\mathbf{z})}{E} [(\mathbf{z} - \boldsymbol{\mu})^T (\mathbf{z} - \boldsymbol{\mu})] = \underset{\mathbf{z} \sim q(\mathbf{z})}{E} [(\mathbf{z} - \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})^T (\mathbf{z} - \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})]$$

$$= \underset{\mathbf{z} \sim q(\mathbf{z})}{E} [(\mathbf{z} - \tilde{\boldsymbol{\mu}})^T (\mathbf{z} - \tilde{\boldsymbol{\mu}}) + r(\mathbf{z} - \tilde{\boldsymbol{\mu}})^T (\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}) + (\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})^T]$$

$$= \underset{\mathbf{z} \sim q(\mathbf{z})}{E} [(\mathbf{z} - \tilde{\boldsymbol{\mu}})^T (\mathbf{z} - \tilde{\boldsymbol{\mu}})] + \underset{\mathbf{z} \sim q(\mathbf{z})}{E} [r(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})^T \underset{\mathbf{z} \sim q(\mathbf{z})}{E} [\mathbf{z} - \tilde{\boldsymbol{\mu}}]] + \underset{\mathbf{z} \sim q(\mathbf{z})}{E} [(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})^T]$$

$\tilde{\sigma}_j^2$ $\underset{\mathbf{z} \sim q(\mathbf{z})}{E} [\mathbf{z} - \tilde{\boldsymbol{\mu}}]$ $= 0$ $(\boldsymbol{\mu}_j - \tilde{\boldsymbol{\mu}}_j)^T$

$$= \sum_j \tilde{\sigma}_j^r + (\mu_j - \tilde{\mu}_j)^r$$

$$\underset{2 \text{nd term}}{E} \left[-\frac{1}{r\sigma_j^r} (z-\mu)^r (z-\mu) \right] = -\sum_j \left(\frac{\tilde{\sigma}_j^r}{r\sigma_j^r} + \frac{(\mu_j - \tilde{\mu}_j)^r}{r\sigma_j^r} \right)$$

$$\rightarrow \underset{2 \text{nd term}}{\sum} \left[\log(p(z)) \right] = E \left[-\frac{1}{r} \log(r\sigma_j^r) - \sum_j \frac{\tilde{\sigma}_j^r + (\mu_j - \tilde{\mu}_j)^r}{r\sigma_j^r} \right]$$

$$\boxed{\int q(z) \log(p(z)) dz = \sum_j -\frac{1}{r} \log(r\sigma_j^r) - \frac{\tilde{\sigma}_j^r}{r\sigma_j^r} - \frac{(\mu_j - \tilde{\mu}_j)^r}{r\sigma_j^r} \quad (\star)}$$

3.5 Putting everything together

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

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

- 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 $q(c|\mathbf{x})$

Just what we found in 3.2:

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

$$(I) \quad E_{q(z|c|x)} \left[\log \frac{p(c)}{q(c|x)} \right] \stackrel{q_c = q(c|x)}{=} \sum_{c \in C} q(c|x) \log \frac{p(c)}{q(c|x)} = \sum_{c \in C} r_c \log \left(\frac{\pi_c}{r_c} \right)$$

$$(III) \quad E_{q(z|c|x)} \left[\log p(x|z) \right] = \int q(z|x) \sum_{c \in C} \log p(x|z) dz$$

$$\begin{aligned}&= E_{q(z|c|x)} \left[\sum_{j=1}^J -\frac{1}{r} \log(\Gamma \delta_j) - \frac{N_j \cdot r}{\Gamma \delta_j} - \frac{(\mu_j - \tilde{\mu}_j)^T}{\Gamma \delta_j} \right] \\ &= \sum_{c \in C} r_c \left[\sum_{j=1}^J -\frac{1}{r} \log(\Gamma \delta_j) - \frac{N_j \cdot r}{\Gamma \delta_j} - \frac{(\mu_j - \tilde{\mu}_j)^T}{\Gamma \delta_j} \right]\end{aligned}$$

$$\text{II), } E_{q(z, c|m)} [\log p(m|z)] = E_{q(z|m)} [\log p(m|z)]$$

$$= E_{q(z|m)} \left[\prod_{i=1}^L \delta_{y_i} \text{Ber}(m_i | \mu_{m_i}) \right] = \frac{1}{L} \sum_{l=1}^L \sum_{j=1}^D \left[m_j \log \left(\frac{\mu_j^l}{m_j} \right) + (1-m_j) \log \left(1 - \frac{\mu_j^l}{m_j} \right) \right]$$

$\log \rightarrow \text{out?}$

III)

$H(q(z|m))$

$$E_{q(z|m)} [\log q(z|m)] = E_{q(z|m)} \int_{\mathbb{R}^D} q(z|m) \log q(z|m) dz$$

$$= H(q(z|m)) = -\frac{1}{r} \sum_{j=1}^J (1 + \log \theta_j)$$

$\theta_j \sim \mathcal{N}$

$$\therefore q(z|m) = \mathcal{N}(z; \tilde{\mu}, \tilde{\sigma}^2 I)$$

$$\text{So: } L_{\text{ELBO}} = \sum_{c \in C} r_c \log \left(\frac{\pi_c}{\theta_c} \right) + \sum_{c \in C} r_c \left[\sum_{j=1}^J -\frac{1}{r} \log (\theta_j) - \frac{\frac{N_j r}{r \theta_j}}{r \theta_j} - \frac{(\mu_j - \tilde{\mu}_j)^T}{r \theta_j} \right]$$

$$+ \frac{1}{L} \sum_{l=1}^L \sum_{j=1}^D \left[m_j \log \left(\frac{\mu_j^l}{m_j} \right) + (1-m_j) \log \left(1 - \frac{\mu_j^l}{m_j} \right) \right] + \frac{1}{r} \sum_{j=1}^J (1 + \log \theta_j)$$

4 Vector Quantized Variational Autoencoders (VQ-VAE)

***(Bonus)**

Answer the below questions.

1. Why VQ-VAE uses a discrete latent model instead of a continuous one? Give an example of a scenario where discrete latent is better than continuous 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?

(1)

يس بدلًا عن تفاصيل متصلة لاتئذ في VQ-VAE
يمكن تفاصيل لاتئذ في الاتجاهات المتصلة
لأنها تزيد من التباين بين المدخلات
• تفاصيل متصلة يمكنها أن تزيد
لذلك يمكن تفاصيل متصلة
أولًا تفاصيل متصلة التي يمكنها أن
تزيد من التباين بين المدخلات
• تفاصيل متصلة يمكنها أن تزيد
لذلك يمكن تفاصيل متصلة

(2)

codebook يمثله في الواقع في VQ-VAE
• تفاصيل متصلة يمكنها أن تزيد
codebook على التباين بين المدخلات
• تفاصيل متصلة يمكنها أن تزيد

—(μ)

• Interventions (Designs) in other two ways :

پر ایڈٹ نہ کر سکتے ہیں لیکن لینڈ ایڈٹ کر سکتے ہیں Forward اور *
کوئی ایڈٹ نہ کر سکتے ہیں Backward اور *
کوئی ایڈٹ نہ کر سکتے ہیں لیکن لینڈ ایڈٹ کر سکتے ہیں
اپنے Face یا Identity Face کو اپنے
update کر سکتے ہیں لیکن لینڈ ایڈٹ
پر ایڈٹ نہ کر سکتے ہیں

14

فرآیند: ۱) درین میان استفاده از هر بطر در Codebook

۲) بزرگترین بطر ها در Codebook برابر سیلیس خود را میزند

که ب آنها نهادست شوند