CSC 6137 Generative Models: Homework 2

Homework Due Date: **Nov 30**

Instructions (Please read carefully)

- Submit your answers as an electronic copy only in **pdf** format on Blackboard.
- No late submissions will be accepted. Zero credit will be assigned for late submissions. Email request for late submission will not be replied.
- Please type in Latex.
- Explicitly mention your collaborators if any. Collaborations should be limited to discussing and learning from each other but please do your own work and write your own codes. We will actively monitor any attempt to copy solutions from each other or from the internet.
- The full score of this homework is 100 pts.

1 VAE: ELBO [20 pts]

For a VAE, given the encoder, i.e., an amortized inference network $q_\phi(z \mid x)$, and a generative model $q_\theta(x \mid z)$,

1. Please derive the ELBO objective and show that it is a lower bound for the log-likelihood function $\log p_\theta(x)$
2. Please derive that the ELBO can be written in 3 equivalent ways as

$$\begin{aligned} \text{ELBO} &= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x, z) - \log q_\phi(z \mid x)\right] && (1)\\ &= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z) + \log p_\theta(z)\right] + \mathbb{H}\left(q_\phi(z \mid x)\right) && (2)\\ &= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] - D_{\mathrm{KL}}\left(q_\phi(z \mid x) \| p_\theta(z)\right) && (3)\end{aligned}$$

3. If we assume a diagonal Gaussian prior, $p_\theta(z) = \mathcal{N}(z \mid 0, \mathrm{I})$ where $z \in R^d$, and diagonal gaussian posterior, $q_\phi(z \mid x) = \mathcal{N}(z \mid \mu_\phi(x), \mathrm{diag}(\sigma_\phi(x)))$. Please describe how to optimize ELBO, such as

$$\mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] - D_{\mathrm{KL}}\left(q_\phi(z \mid x) \| p_\theta(z)\right) \tag{4}$$

w.r.t. θ and ϕ by SGD.

4. The above ELBO is derived for individual samples. Now, let's take an expectation with all the data, i.e., $\mathbb{E}_{p_{\mathcal{D}}(x)}[\cdots]$. Please derive that VAEs optimize in an augmented space. i.e., it optimize

$$-D_{\mathrm{KL}}\left(q_{\mathcal{D},\phi}(x, z) \| p_\theta(x, z)\right) + \mathbb{E}_{p_{\mathcal{D}}(x)}\left[\log p_{\mathcal{D}}(x)\right] \tag{5}$$

5. Please continue to verify that the above expression equals to

$$-D_{\mathrm{KL}}\left(p_{\mathcal{D}}(x) \| p_\theta(x)\right) - \mathbb{E}_{p_{\mathcal{D}}(x)}\left[D_{\mathrm{KL}}\left(q_\phi(z \mid x) \| p_\theta(z \mid x)\right)\right] \tag{6}$$

and

$$-D_{\mathrm{KL}}\left(q_{\mathcal{D},\phi}(z) \| p_\theta(z)\right) - \mathbb{E}_{q_{\mathcal{D},\phi}(z)}\left[D_{\mathrm{KL}}\left(q_\phi(x \mid z) \| p_\theta(x \mid z)\right)\right] \tag{7}$$

2 VAE: flow-based posterior model [20 pts]

We model the approximate posterior distribution as a flow-based model.

$$\begin{aligned}
&\mathbf{x} \leftarrow \{\text{ Get mini-batch }\} && (8)\\
&\mathbf{z}_0 \sim q_0(\bullet \mid \mathbf{x}) && (9)\\
&\mathbf{z}_K \leftarrow f_K \circ f_{K-1} \circ \ldots \circ f_1\left(\mathbf{z}_0\right) && (10)
\end{aligned}$$

1. We consider a family of transformations of the form:

   $$f(\mathbf{z}) = \mathbf{z} + \mathbf{u}h\left(\mathbf{w}^{\top}\mathbf{z} + b\right),$$

   where $\lambda = \left\{\mathbf{w} \in \mathbb{R}^D, \mathbf{u} \in \mathbb{R}^D, b \in \mathbb{R}\right\}$ are free parameters and $h(\cdot)$ is a smooth element-wise non-linearity, with derivative $h'(\cdot)$. Please prove that

   $$\ln q_K\left(\mathbf{z}_K\right) = \ln q_0(\mathbf{z}_0) - \sum_{k=1}^{K} \ln\left|1 + \mathbf{u}_k^{\top}\psi_k\left(\mathbf{z}_{k-1}\right)\right|.$$

   and

   $$\psi(\mathbf{z}) = h'\left(\mathbf{w}^{\top}\mathbf{z} + b\right)\mathbf{w}. \tag{11}$$

2. Given the above approximate posterior, if we parameterize the approximate posterior distribution with a flow of length $K, q_\phi(\mathbf{z} \mid \mathbf{x}) := q_K\left(\mathbf{z}_K\right)$, the ELBO can be written as an expectation over the initial distribution $q_0(\mathbf{z})$. Please derive the ELBO objective and show that

   $$\text{ELBO} = \mathbb{E}_{q_0(z_0)}\left[\ln q_0\left(\mathbf{z}_0\right)\right] - \mathbb{E}_{q_0(z_0)}\left[\log p_\theta\left(\mathbf{x}, \mathbf{z}_K\right)\right] - \mathbb{E}_{q_0(z_0)}\left[\sum_{k=1}^{K} \ln\left|1 + \mathbf{u}_k^{\top}\psi_k\left(\mathbf{z}_{k-1}\right)\right|\right] \tag{12}$$

3. Given the above derivation, please describe how to optimize the inference network parameter ϕ and the generative model parameter θ.

4. As we can see, in order to optimize this objective, we need to be able to efficiently sample from $q_\phi(z \mid x)$ and evaluate the probability density of these samples during optimization. We can use inverse autoregressive flows (IAFs)! Please describe the corresponding algorithm and the potential benefits brought by IAFs.

3 VAE: prior learning [20 pts]

In this question, we will improve VAE by making the prior more flexible. We will consider a new prior that is a mixture of variational posteriors conditioned on learnable pseudo-data. This allows the variational posterior to learn more a potent latent representation. Consider the ELBO objective,

$$\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[\ln p(\mathbf{x})] \geq \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}\mid\mathbf{x})}\left[\ln p_\theta(\mathbf{x} \mid \mathbf{z}) + \ln p_\lambda(\mathbf{z}) - \ln q_\phi(\mathbf{z} \mid \mathbf{x})\right]\right] := \mathcal{L}(\phi, \theta, \lambda), \tag{13}$$

where $q_\phi(\mathbf{z} \mid \mathbf{x})$ is the approximate inference, prior $p_\lambda(\mathbf{z})$, and generative model $p_\theta(\mathbf{x} \mid \mathbf{z})$. During learning we consider a Monte Carlo estimate of the second expectation using L sample points:

$$\tilde{\mathcal{L}}(\phi, \theta, \lambda) = \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\frac{1}{L}\sum_{l=1}^{L}\left(\ln p_\theta\left(\mathbf{x} \mid \mathbf{z}_\phi^{(l)}\right) + \ln p_\lambda\left(\mathbf{z}_\phi^{(l)}\right) - \ln q_\phi\left(\mathbf{z}_\phi^{(l)} \mid \mathbf{x}\right)\right)\right],$$

where $\mathbf{z}_\phi^{(l)}$ are sampled from $q_\phi(\mathbf{z} \mid \mathbf{x})$ through the reparameterization trick. Note that the second and third components constitute a kind of regularization that drives the encoder to match the prior. The prior plays a role of an "anchor" that keeps the posterior close to it. Typically, the encoder is assumed to have a diagonal covariance matrix, i.e., $q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{z} \mid \mu_\phi(\mathbf{x}), \operatorname{diag}\left(\sigma_\phi^2(\mathbf{x})\right)\right)$, where $\mu_\phi(\mathbf{x})$ and $\sigma_\phi^2(\mathbf{x})$ are parameterized by a NN with weights ϕ, and the prior is expressed using the standard normal distribution, $p_\lambda(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I})$. The decoder utilizes a suitable distribution for the data under consideration, e.g., the Bernoulli distribution for binary data or the normal distribution for continuous data, and it is parameterized by a NN with weights θ.

1. Please rewrite the ELBO as

$$\mathcal{L}(\phi, \theta, \lambda) = \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln p_\theta(\mathbf{x} \mid \mathbf{z})\right]\right] + \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})}\left[\mathbb{H}\left[q_\phi(\mathbf{z} \mid \mathbf{x})\right]\right] - \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})}\left[-\ln p_\lambda(\mathbf{z})\right] \tag{14}$$

where we name $q(\mathbf{z})$ as the aggregated posterior, and $q(\mathbf{z}) = \frac{1}{N}\sum_{n=1}^{N} q_\phi\left(\mathbf{z} \mid \mathbf{x}_n\right)$. Please explain the meaning of each three terms.

2. Instead of choosing the prior in advance, e.g., a standard normal prior, one could find a prior that optimizes the ELBO by maximizing the following Lagrange function with the Lagrange multiplier β :

$$\max_{p_\lambda(\mathbf{z})} -\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})}\left[-\ln p_\lambda(\mathbf{z})\right] + \beta\left(\int p_\lambda(\mathbf{z}) \mathrm{d}\mathbf{z} - 1\right).$$

Please verify that the solution of the above problem is simply the aggregated posterior:

$$p_\lambda^*(\mathbf{z}) = \frac{1}{N}\sum_{n=1}^{N} q_\phi\left(\mathbf{z} \mid \mathbf{x}_n\right).$$

3. However, the above choice for the prior may potentially lead to overfitting and is computationally expensive. On the other hand, having a simple prior, such as the standard normal may result in over-regularized models with only few active latent dimensions. We therefore choose to model the prior as a mixture of variational posteriors with pseudo-inputs:

$$p_\lambda(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K} q_\phi\left(\mathbf{z} \mid \mathbf{u}_k\right),$$

where K is the number of pseudo-inputs, and $\mathbf{u}_k$ is a D-dimensional vector we refer to as a pseudo-input. We can learn the prior by learning the pseudo-inputs. Now, $\lambda = \{\mathbf{u}_1, \ldots, \mathbf{u}_K, \phi\}$. Please describe the backpropagation algorithm that is used to learn the prior. Note that we are coupling the parameters of prior and approximate posteriors.

4. K is a hyperparameter. Please discuss the impact by tuning K (from the perspectives of computational complexity and model regularization.)

5. A simpler alternative to the previous prior is to model it as a mixture of Gaussians (MoG),

$$p_\lambda(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K} \mathcal{N}\left(\mu_k, \operatorname{diag}\left(\sigma_k^2\right)\right)$$

In this case, the hyperparameters of the prior $\lambda = \left\{\mu_k, \operatorname{diag}\left(\sigma_k^2\right)\right\}_{k=1}^{K}$. Describe how to train by backpropagation similarly to the pseudo-inputs.

6. Describe how the MoG prior influences the variational posterior in the same manner as the standard prior (e.g., standard Gaussian) and the gradient of the ELBO with respect to the encoder's parameters.

4 VAE and GAN: posterior collapse [10 pts]

1. Please describe what is the posterior collapse issue faced by VAE. Please justify that the posterior collapse is rooted in the ELBO objectives, which inherently favors fitting the data over performing the correct amortized inference.

2. Given the original ELBO objective

$$\mathcal{L}_{\mathrm{ELBO}}(x) = -D_{\mathrm{KL}}\left(q_\phi(z \mid x) \| p(z)\right) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] \tag{15}$$

   We propose the InfoVAE to balance the learning and inference in ELBO, with a modified objective as

$$\mathcal{L}_{\mathrm{InfoVAE}} = -\lambda D_{\mathrm{KL}}\left(q_\phi(z) \| p(z)\right) - \mathbb{E}_{q(z)}\left[D_{\mathrm{KL}}\left(q_\phi(x \mid z) \| p_\theta(x \mid z)\right)\right] + \alpha I_q(x; z) \tag{16}$$

   where $I_q(x; z)$ is the mutual information between x and z under the distribution $q_\phi(x, z)$. Please justify why this objective may address the posterior collapse issue. Please show what are the computational burden in terms of evaluating the objective function?

3. Please prove that the above objective can be rewritten as

$$\begin{aligned}\mathcal{L}_{\mathrm{InfoVAE}} =& \mathbb{E}_{p_{\mathcal{D}}(x)} \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] - (1-\alpha) \mathbb{E}_{p_{\mathcal{D}}(x)} D_{\mathrm{KL}}\left(q_\phi(z \mid x) \| p(z)\right) - \\ & (\alpha + \lambda - 1) D_{\mathrm{KL}}\left(q_\phi(z) \| p(z)\right)\end{aligned}$$

4. In fact we may replace the term $D_{\mathrm{KL}}\left(q_\phi(z) \| p(z)\right)$ with anther strict divergence $D\left(q_\phi(z) \| p(z)\right)$, which may be more efficient to optimize. Many divergence used in GAN that can be easily approximated by samples can be used here. Please propose a GAN model to aid the InfoVAE learning. Describe your ideas briefly.

5 VAE: Programming [30 pts]

1. Please implement the VAE codes (vae_example.ipynb) and summarize the output.

2. Introduce a β coefficient to the KL regularization term (i.e., β-VAE). By tuning various values of $\beta = 0, 0.5, 1, 10$, what are your discoveries and summarize your outputs.

3. For the same dataset, please use the labeling information, i.e., $\{0, 1, \ldots, 16\}$, in both training and data generating stage. Your goal is to build a conditional generator. Please provide your ideas. Revise the codes and report your outcomes.

4. Please consider learning the priors. First implement the VAE with learnable prior codes (vae_priors_example.ipynb). Consider the standard Gaussian prior, mixture of Gaussian priors, the mixture of variational posteriors with pseudo-inputs (i.e., VampPrior). Summarize your output. For the VampPrior, please tune the number of pseudo-inputs, and report your discoveries. For the mixture of Gaussains prior, please tune the number of Gaussian components, and report your discoveries.

5. Please implement the flow-based priors. Report your output and discoveries.