Due Date: April 5th 23:59, 2019

Name: Aditya Joshi
Matricule:

Instructions

- *For all questions, show your work!*
- *Starred questions are **hard** questions, not **bonus** questions.*
- *Please use a document preparation system such as LaTeX, unless noted otherwise.*
- *Unless noted that questions are related, assume that notation and defintions for each question are self-contained and independent*
- *Submit your answers electronically via Gradescope.*
- ***TAs for this assignment are Shawn Tan, Samuel Lavoie, and Chin-Wei Huang.***

This assignment covers mathematical and algorithmic techniques underlying the three most popular families of deep generative models, variational autoencoders (VAEs, Questions 1-3), autoregressive models (Question 4), and generative adversarial networks (GANs, Questions 5-7).

Question 1 (8-8)**.** Reparameterization trick is a standard technique that makes the samples of a random variable differentiable. Consider a random vector $Z \in \mathbb{R}^K$ with a density function $q(\boldsymbol{z}; \phi)$. We want to find a deterministic function $\boldsymbol{g} : \mathbb{R}^K \to \mathbb{R}^K$ that depends on ϕ, to transform a random variable Z_0 having a ϕ-independent density function $q(\boldsymbol{z}_0)$, such that $\boldsymbol{g}(Z_0)$ has the same density as Z. Recall the change of density for a bijective, differentiable $\boldsymbol{g}$:

$$q(\boldsymbol{g}(\boldsymbol{z}_0)) = q(\boldsymbol{z}_0) \left| \det \left(\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0} \right) \right|^{-1} \tag{1}$$

1. Assume $q(\boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$ and $\boldsymbol{g}(\boldsymbol{z}_0) = \mu + \sigma \odot \boldsymbol{z}_0$, where $\mu \in \mathbb{R}^K$ and $\sigma \in \mathbb{R}^K_{>0}$. Show that $\boldsymbol{g}(\boldsymbol{z}_0)$ is distributed by $\mathcal{N}(\mu, \text{diag}(\sigma^2))$ using Equation (1).

We have that $q(z_0) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$. The reparameterization trick is used so that a simple distribution that is independent of parameters ϕ together with a function $g(.)$ that depends on parameters ϕ can be used to sample from a distribution that is parameterized by ϕ. In this case, the target distribution is $\mathcal{N}(\mu, \text{diag}(\sigma^2))$.

We use equation (1), we first need to find the quantity $\dfrac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0}$ (A Jacobian of dimensions $K \times K$.)

$$\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0} = \begin{pmatrix} \dfrac{\partial \boldsymbol{g}(\boldsymbol{z}_0)_1}{\partial (\boldsymbol{z}_0)_1} & \dfrac{\partial \boldsymbol{g}(\boldsymbol{z}_0)_1}{\partial (\boldsymbol{z}_0)_2} & \cdots & \dfrac{\partial \boldsymbol{g}(\boldsymbol{z}_0)_1}{\partial (\boldsymbol{z}_0)_K} \\ \vdots & \cdots & \cdots & \cdots \\ \vdots & \cdots & \cdots & \cdots \\ \dfrac{\partial \boldsymbol{g}(\boldsymbol{z}_0)_K}{\partial (\boldsymbol{z}_0)_1} & \cdots & \cdots & \dfrac{\partial \boldsymbol{g}(\boldsymbol{z}_0)_K}{\partial (\boldsymbol{z}_0)_K} \end{pmatrix}$$

Here, since we have an element wise product, all the non-diagonal terms will be 0 because $(\boldsymbol{z}_0)_i$ does not influence $g(\boldsymbol{z}_0)_j$ in any way when $i \neq j$. The diagonal terms are of the form $(\sigma)_i$ for $1 \leq i \leq K$. Thus the Jacobian is as follows:

$$= \begin{pmatrix} (\sigma)_1 & 0 & \cdots & 0 \\ 0 & (\sigma)_2 & \cdots & 0 \\ \vdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & (\sigma)_K \end{pmatrix}$$

The determinant of a diagonal matrix is equal to the product of the diagonal terms, and thus we have $\left|\det \dfrac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0}\right| = \prod_{i=1}^K (\sigma)_i = \sqrt{|\text{diag}(\sigma^2)|}$. The inverse of this expression is then:

$$\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0} = \frac{1}{\prod_{i=1}^K (\sigma)_i} = \frac{1}{\sqrt{|\text{diag}(\sigma^2)|}} \tag{2}$$

Next, we expand the term $q(\boldsymbol{z}_0) \sim \mathcal{N}(0, \boldsymbol{I}_K)$. This equals:

$$= \frac{1}{(2\pi)^{\frac{K}{2}}} \cdot \exp\left\{\frac{-1}{2} \cdot \boldsymbol{z}_0^\top \boldsymbol{z}_0\right\} \tag{3}$$

We need to express $\boldsymbol{z}_0$ in terms of $g(\boldsymbol{z}_0)$. We have that:

$$g(\boldsymbol{z}_0) = \mu + \sigma \odot \boldsymbol{z}_0 \tag{4}$$
$$\boldsymbol{z}_0 = \sigma^{-1} \odot (g(\boldsymbol{z}_0) - \mu) \tag{5}$$
$$= \text{diag}(\sigma^{-1})(g(\boldsymbol{z}_0) - \mu) \tag{6}$$

Let us now substitute equation 6 in equation 3.

$$= \frac{1}{(2\pi)^{\frac{K}{2}}} \cdot \exp\left\{\frac{-1}{2} \cdot \Big(\text{diag}(\sigma^{-1})(g(\boldsymbol{z}_0 - \mu))\Big)^\top \Big(\text{diag}(\sigma^{-1})(g(\boldsymbol{z}_0 - \mu))\Big)\right\} \tag{7}$$
$$= \frac{1}{(2\pi)^{\frac{K}{2}}} \cdot \exp\left\{\frac{-1}{2} \cdot \Big((g(\boldsymbol{z}_0 - \mu))^\top \text{diag}(\sigma^{-1})^\top\Big) \Big(\text{diag}(\sigma^{-1})(g(\boldsymbol{z}_0 - \mu))\Big)\right\} \tag{8}$$
$$= \frac{1}{(2\pi)^{\frac{K}{2}}} \cdot \exp\left\{\frac{-1}{2} \cdot (g(\boldsymbol{z}_0) - \mu)^\top \cdot \text{diag}(\sigma^2)^{-1} \cdot (g(\boldsymbol{z}_0) - \mu)\right\} \tag{9}$$

Finally, we combine 2 and 19 as per the formula given in the question. We then get:

$$
\begin{aligned}
g(\boldsymbol{z}_0) &= \frac{1}{(2\pi)^{\frac{K}{2}}} \cdot \exp\left\{\frac{-1}{2} \cdot (g(\boldsymbol{z}_0) - \mu)^\top \cdot \mathrm{diag}(\sigma^2)^{-1} \cdot (g(\boldsymbol{z}_0) - \mu)\right\} \cdot \frac{1}{\sqrt{|\mathrm{diag}(\sigma^2)|}} \\
&= \frac{1}{(2\pi)^{\frac{K}{2}} \cdot \sqrt{|\mathrm{diag}(\sigma^2)|}} \cdot \exp\left\{\frac{-1}{2} \cdot (g(\boldsymbol{z}_0) - \mu)^\top \cdot \mathrm{diag}(\sigma^2)^{-1} \cdot (g(\boldsymbol{z}_0) - \mu)\right\}
\end{aligned}
$$

$$
g(\boldsymbol{z}_0) \sim \mathcal{N}(\mu, \mathrm{diag}(\sigma^2))
$$

2. Assume instead $\boldsymbol{g}(\boldsymbol{z}_0) = \mu + \boldsymbol{S}\boldsymbol{z}_0$, where $\boldsymbol{S}$ is a non-singular $K \times K$ matrix. Derive the density of $\boldsymbol{g}(\boldsymbol{z}_0)$ using Equation (1).

   As with the first part of the question, we first find the quantity $\dfrac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0}$.

$$\begin{aligned}
\boldsymbol{g}(\boldsymbol{z}_0) &= \mu + \boldsymbol{S}\boldsymbol{z}_0 && (10)\\
\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0} &= \frac{\partial}{\partial \boldsymbol{z}_0}\Big(\mu + \boldsymbol{S}\boldsymbol{z}_0\Big) && (11)\\
&= \boldsymbol{S} && (12)\\
\left|\det\left(\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0}\right)\right|^{-1} &= |\det(\boldsymbol{S})|^{-1} = \frac{1}{\sqrt{|\det(\boldsymbol{S}\boldsymbol{S}^\top)|}} && (13)
\end{aligned}$$

   And following the same steps as question 1, we express $\boldsymbol{g}(\boldsymbol{z}_0)$ in terms of $\boldsymbol{z}_0$. We can find the inverse of $\boldsymbol{S}$ is it is non-singular.

$$\begin{aligned}
\boldsymbol{g}(\boldsymbol{z}_0) &= \mu + \boldsymbol{S}\boldsymbol{z}_0 && (14)\\
\boldsymbol{z}_0 &= \boldsymbol{S}^{-1}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu) && (15)
\end{aligned}$$

   Using equation 3 from question 1, we have:

$$\begin{aligned}
&= \frac{1}{(2\pi)^{\frac{K}{2}}} \cdot \exp\left\{\frac{-1}{2} \cdot \boldsymbol{z}_0^\top \boldsymbol{z}_0\right\} && (16)\\
&= \frac{1}{(2\pi)^{\frac{K}{2}}} \cdot \exp\left\{\frac{-1}{2} \cdot \Big(\boldsymbol{S}^{-1}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu)\Big)^\top \Big(\boldsymbol{S}^{-1}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu)\Big)\right\} && (17)\\
&= \frac{1}{(2\pi)^{\frac{K}{2}}} \cdot \exp\left\{\frac{-1}{2} \cdot \Big((\boldsymbol{g}(\boldsymbol{z}_0) - \mu)^\top (\boldsymbol{S}^{-1})^\top\Big) \Big(\boldsymbol{S}^{-1}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu)\Big)\right\} && (18)\\
&= \frac{1}{(2\pi)^{\frac{K}{2}}} \cdot \exp\left\{\frac{-1}{2} \cdot \Big(\boldsymbol{g}(\boldsymbol{z}_0) - \mu\Big)^\top \cdot (\boldsymbol{S}\boldsymbol{S}^\top)^{-1} \cdot \Big(\boldsymbol{g}(\boldsymbol{z}_0) - \mu\Big)\right\} && (19)
\end{aligned}$$

   Now combining 13 and 19 as per Equation 1, we have:

$$\begin{aligned}
&= \frac{1}{(2\pi)^{\frac{K}{2}}} \cdot \exp\left\{\frac{-1}{2} \cdot \Big(\boldsymbol{g}(\boldsymbol{z}_0) - \mu\Big)^\top \cdot (\boldsymbol{S}\boldsymbol{S}^\top)^{-1} \cdot \Big(\boldsymbol{g}(\boldsymbol{z}_0) - \mu\Big)\right\} \cdot \frac{1}{\sqrt{|\det(\boldsymbol{S}\boldsymbol{S}^\top)|}}\\
&= \frac{1}{(2\pi)^{\frac{K}{2}} \cdot \sqrt{|\det(\boldsymbol{S}\boldsymbol{S}^\top)|}} \cdot \exp\left\{\frac{-1}{2} \cdot \Big(\boldsymbol{g}(\boldsymbol{z}_0) - \mu\Big)^\top \cdot (\boldsymbol{S}\boldsymbol{S}^\top)^{-1} \cdot \Big(\boldsymbol{g}(\boldsymbol{z}_0) - \mu\Big)\right\}
\end{aligned}$$

$$\boldsymbol{g}(\boldsymbol{z}_0) \sim \mathcal{N}(\mu, \boldsymbol{S}\boldsymbol{S}^\top)$$

Question 2 (5-5-6)**.** Consider a latent variable model $\boldsymbol{z} \sim p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_K)$ where $\boldsymbol{z} \in \mathbb{R}^K$, and $\boldsymbol{x} \sim p_\theta(\boldsymbol{x}|\boldsymbol{z})$. The encoder network (aka "recognition model") of variational autoencoder, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, is used to produce an approximate (variational) posterior distribution over latent variables $\boldsymbol{z}$ for any input datapoint $\boldsymbol{x}$.[1] This distribution is trained to match the true posterior by maximizing the evidence lower bound (ELBO):

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = \mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) || p(\boldsymbol{z}))$$

Let $\mathcal{Q}$ be the family of variational distributions with a feasible set of parameters $\mathcal{P}$; i.e. $\mathcal{Q} = \{q(\boldsymbol{z}; \pi) : \pi \in \mathcal{P}\}$; for example π can be mean and standard deviation of a normal distribution. We assume q_ϕ is parameterized by a neural network (with parameters ϕ) that outputs the parameters, $\pi_\phi(\boldsymbol{x})$, of the distribution $q \in \mathcal{Q}$, i.e. $q_\phi(\boldsymbol{z}|\boldsymbol{x}) := q(\boldsymbol{z}; \pi_\phi(\boldsymbol{x}))$.

1. Show that maximizing the expected complete data log likelihood

$$\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]$$

for a fixed $q(\boldsymbol{z}|\boldsymbol{x})$, wrt the model parameter θ, gives the maximizer of the biased log marginal likelihood: $\arg\max_\theta\{\log p_\theta(\boldsymbol{x}) + B(\theta)\}$, where $B(\theta)$ is non-positive. Find $B(\theta)$.

Let us expand the term $\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]$:

$$\begin{aligned}
\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] &= \int_{z\sim q(\boldsymbol{z}|\boldsymbol{x})} q(\boldsymbol{z}|\boldsymbol{x}) \cdot \log p_\theta(\boldsymbol{x}|\boldsymbol{z}) \cdot dz \\
&= \int_{z\sim q(\boldsymbol{z}|\boldsymbol{x})} q(\boldsymbol{z}|\boldsymbol{x}) \left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{p_\theta(\boldsymbol{z})} \right] \cdot dz \\
&= \int_{z\sim q(\boldsymbol{z}|\boldsymbol{x})} q(\boldsymbol{z}|\boldsymbol{x}) \left[\log \frac{p_\theta(\boldsymbol{z}|\boldsymbol{x}) \cdot p_\theta(\boldsymbol{x})}{p_\theta(\boldsymbol{z})} \right] \cdot dz \\
&= \int_{z\sim q(\boldsymbol{z}|\boldsymbol{x})} q(\boldsymbol{z}|\boldsymbol{x}) \cdot \log p_\theta(\boldsymbol{x}) \cdot dz + \int_{z\sim q(\boldsymbol{z}|\boldsymbol{x})} q(\boldsymbol{z}|\boldsymbol{x}) \cdot \log \left[\frac{p_\theta(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z})} \right] \cdot dz \\
&= \log p_\theta(\boldsymbol{x}) + \int_{z\sim q(\boldsymbol{z}|\boldsymbol{x})} q(\boldsymbol{z}|\boldsymbol{x}) \cdot \log \left[\frac{p_\theta(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{x})} \cdot \frac{q(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z})} \right] \cdot dz \\
&= \log p_\theta(\boldsymbol{x}) - \int_{z\sim q(\boldsymbol{z}|\boldsymbol{x})} q(\boldsymbol{z}|\boldsymbol{x}) \cdot \log \left[\frac{q(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} \right] + \int_{z\sim q(\boldsymbol{z}|\boldsymbol{x})} q(\boldsymbol{z}|\boldsymbol{x}) \cdot \log \left[\frac{q(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z})} \right] \\
&= \log p_\theta(\boldsymbol{x}) - KL(q(\boldsymbol{z}|\boldsymbol{x}) || p_\theta(\boldsymbol{z}|\boldsymbol{x})) + KL(q(\boldsymbol{z}|\boldsymbol{x}) || \underbrace{p_\theta(\boldsymbol{z})}_{\text{independent of } \theta})
\end{aligned}$$

Thus, we have:

$$\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \underbrace{KL(q(\boldsymbol{z}|\boldsymbol{x}) || p(\boldsymbol{z}))}_{\text{independant of } \theta} = \log p_\theta(\boldsymbol{x}) - KL(q(\boldsymbol{z}|\boldsymbol{x}) || p_\theta(\boldsymbol{z}|\boldsymbol{x}))$$

$$\text{maximize } \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] = \text{argmax}_\theta \Big\{ \log p_\theta(\boldsymbol{x}) - KL(q(\boldsymbol{z}|\boldsymbol{x}) || p_\theta(\boldsymbol{z}|\boldsymbol{x})) \Big\}$$

Thus, we have $B(\theta) = -KL(q(\boldsymbol{z}|\boldsymbol{x}) || p_\theta(\boldsymbol{z}|\boldsymbol{x}))$.

[1] Using a recognition model in this way is known as "amortized inference" ; this can be contrasted with traditional variational inference approaches (see, e.g., Chapter 10 of Bishop's *Pattern Recognition an Machine Learning*), which fit a variational posterior independently for each new datapoint.

2. Consider a finite training set $\{\boldsymbol{x}_i : i \in \{1, ..., n\}\}$, n being the size the training data. Let ϕ^* be the maximizer of $\sum_{i=1}^n \mathcal{L}(\theta, \phi; \boldsymbol{x}_i)$ with θ fixed. In addition, for each $\boldsymbol{x}_i$ let $q_i \in \mathcal{Q}$ be an instance-dependent variational distribution, and denote by q_i^* the maximizer of the corresponding ELBO. Compare $D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$ and $D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$. Which one is bigger ?

   Let us first expand $D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$:

$$
\begin{aligned}
&= \int_z q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i) \cdot \log \frac{q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)}{p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)} \cdot dz \\
&= \int_z q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i) \cdot \log \left[\frac{q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)}{p_\theta(\boldsymbol{z})} \cdot \frac{1}{p_\theta(\boldsymbol{x}_i|\boldsymbol{z})} \cdot p_\theta(x_i) \right] \\
&= \int_z q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i) \cdot \log p_\theta(x_i) \cdot dz - \int_z q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i) \cdot \log p_\theta(\boldsymbol{x}_i|\boldsymbol{z}) \cdot dz + \int_z q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i) \cdot \log \frac{q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)}{p_\theta(\boldsymbol{z})} \cdot dz \\
&= \log p_\theta(\boldsymbol{x}_i) - \Bigg(\underbrace{\mathbb{E}_{z \sim q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)}[\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z})] - D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}))}_{\text{ELBO}} \Bigg) \\
&= \log p_\theta(\boldsymbol{x}_i) - \mathcal{L}(\theta, \phi^*, \boldsymbol{x}_i)
\end{aligned}
$$

   Now let us expand $D_{KL}(q_i^*(\boldsymbol{z})||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$ in the same manner:

$$
\begin{aligned}
&= \int_z q_i^*(\boldsymbol{z}) \cdot \log \frac{q_i^*(\boldsymbol{z})}{p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)} \cdot dz \\
&= \int_z q_i^*(\boldsymbol{z}) \cdot \log \left[\frac{q_i^*(\boldsymbol{z})}{p_\theta(\boldsymbol{z})} \cdot \frac{1}{p_\theta(\boldsymbol{x}_i|\boldsymbol{z})} \cdot p_\theta(\boldsymbol{x}_i) \right] \cdot dz \\
&= \int_z q_i^*(\boldsymbol{z}) \cdot \log p_\theta(\boldsymbol{x}_i) \cdot dz - \int_z q_i^*(\boldsymbol{z}) \cdot \log p_\theta(\boldsymbol{x}_i|\boldsymbol{z}) \cdot dz + \int_z q_i^*(\boldsymbol{z}) \cdot \log \frac{q_i^*(\boldsymbol{z})}{p_\theta(\boldsymbol{z})} \cdot dz \\
&= \log p_\theta(\boldsymbol{x}_i) - \Bigg(\underbrace{\mathbb{E}_{z \sim q_i^*}[\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z})] - D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})||p_\theta(\boldsymbol{z}))}_{\text{ELBO}} \Bigg) \\
&= \log p_\theta(\boldsymbol{x}_i) - \mathcal{L}(\theta, \phi_i^*, \boldsymbol{x}_i)
\end{aligned}
$$

   Now, we have the following inequality (because the ELBO for $\boldsymbol{x}_i$ is maximized by q^*):

$$
\mathcal{L}(\theta, \phi^*, \boldsymbol{x}_i) \leq \mathcal{L}(\theta, \phi_i^*, \boldsymbol{x}_i)
$$

$$
D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)) \geq D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))
$$

3. Following the previous question, compare the two approaches in the second subquestion

   (a) in terms of bias of estimating the marginal likelihood via the ELBO, in the best case scenario (i.e. when both approaches are optimal within the respective families)

   The non-positive bias is equal to the KL divergence between the q and p. Thus in the amortized approach it is $D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$ for the non-amortized approach it is $D_{KL}(q_i^*(\boldsymbol{z})||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$. We have the same relation as above for the bias, which is:

   $$D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)) \geq D_{KL}(q_i^*(\boldsymbol{z})||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$$

   (b) from the computational point of view (efficiency)

   The amortized approach is better as we would need to calculate the posterior only for one value of q_{ϕ^*} where as the other approach would be calculating one posterior for each data point $\boldsymbol{x}_i$, and that is more computationally expensive.

   (c) in terms of memory (storage of parameters)

   Similar the above reasoning, the amortized approach will need to store parameters only for calculating one value of q_{ϕ^*} but the non-amortized approach will need to store parameters proportional to the number of data points $\boldsymbol{x}_i$. Hence the amortized approach will be more efficient in terms of memory.

Question 3 (6-6)**.** Since variational inference provides a lower-bound on the log marginal likelihood of the data, it gives us a biased estimate of the marginal likelihood. Therefore, methods of "tightening" the bound (i.e. finding a higher valid lower bound) may be desirable.

Consider a latent variable model with the joint $p(\boldsymbol{x}, \boldsymbol{h})$ where $\boldsymbol{x}$ and $\boldsymbol{h}$ are the observed and unobserved random variables, respectively. Now let $q(\boldsymbol{h})$ be a variational approximation to $p(\boldsymbol{h}|\boldsymbol{x})$. Define

$$\mathcal{L}_K = \mathbb{E}_{\boldsymbol{h}_j \sim q(\boldsymbol{h})}\left[\log \frac{1}{K}\sum_{j=1}^{K} \frac{p(\boldsymbol{x}, \boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right]$$

Note that $\mathcal{L}_1$ is equivalent to the evidence lower bound (ELBO).

1. Show that $\mathcal{L}_K$ is a lower bound of the log marginal likelihood $\log p(\boldsymbol{x})$.
   In the following relation, we use the Jensen's inequality.

$$\begin{aligned}
\mathcal{L}_K &= \mathbb{E}_{\boldsymbol{h}_j \sim q(\boldsymbol{h})}\left[\log \frac{1}{K}\sum_{j=1}^{K} \frac{p(\boldsymbol{x}, \boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right] \\
&\leq \log \mathbb{E}_{\boldsymbol{h}_j \sim q(\boldsymbol{h})}\left[\frac{1}{K}\sum_{j=1}^{K} \frac{p(\boldsymbol{x}, \boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right] \\
&\leq \log \frac{1}{K}\sum_{j=1}^{K} \mathbb{E}_{\boldsymbol{h}_j \sim q(\boldsymbol{h})}\left[\frac{p(\boldsymbol{x}, \boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right] \\
&\leq \log \frac{1}{K}\sum_{j=1}^{K} \int_{\boldsymbol{h}_j} q(\boldsymbol{h}_j) \cdot \frac{p(\boldsymbol{x}, \boldsymbol{h}_j)}{q(\boldsymbol{h}_j)} \\
&\leq \log \frac{1}{K}\sum_{j=1}^{K} \int_{\boldsymbol{h}_j} p(\boldsymbol{x}, \boldsymbol{h}_j) \\
&\leq \log \frac{1}{K}\sum_{j=1}^{K} p(\boldsymbol{x}) \\
\mathcal{L}_K &\leq \log p(\boldsymbol{x})
\end{aligned}$$

   Thus, as $\log p(\boldsymbol{x}) \geq \mathcal{L}_K$, $\mathcal{L}_K$ is a lower bound of the log marginal likelihood $\log p(\boldsymbol{x})$.

2. Show that $\mathcal{L}_K \geq \mathcal{L}_1$; i.e. $\mathcal{L}_K$ is a family of lower bounds tighter than the ELBO.
   In the following derivation, we use the log sum inequality:

$$
\begin{aligned}
\mathcal{L}_K &= \mathbb{E}_{\boldsymbol{h}_j \sim q(\boldsymbol{h})}\left[\log \frac{1}{K}\sum_{j=1}^{K} \frac{p(\boldsymbol{x}, \boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right] \\
&\geq \underbrace{\mathbb{E}_{\boldsymbol{h}_j \sim q(\boldsymbol{h})}\left[\frac{1}{K}\sum_{j=1}^{K} \log \frac{p(\boldsymbol{x}, \boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right]}_{\text{log sum inequality}} \\
&\geq \frac{1}{K}\sum_{j=1}^{K} \mathbb{E}_{\boldsymbol{h}_j \sim q(\boldsymbol{h})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right] \\
&\geq \frac{1}{K}\sum_{j=1}^{K} \mathbb{E}_{\boldsymbol{h}_1 \sim q(\boldsymbol{h})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{h}_1)}{q(\boldsymbol{h}_1)}\right] \\
&\geq \mathbb{E}_{\boldsymbol{h}_1 \sim q(\boldsymbol{h})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{h}_1)}{q(\boldsymbol{h}_1)}\right] = \mathcal{L}_1 \\
\mathcal{L}_K &\geq \mathcal{L}_1
\end{aligned}
$$

   Thus, $\mathcal{L}_K$ is a family of lower bounds tighter than the ELBO.

Question 4 (5-5-5-5)**.** One way to enforce autoregressive conditioning is via masking the weight parameters.[2] Consider a two-layer convolutional neural network without kernel flipping, with kernel size 3×3 and padding size 1 on each border (so that an input feature map of size 5×5 is convolved into a 5×5 output). Define mask of type A and mask of type B as

$$(\boldsymbol{M}^A)_{::ij} := \begin{cases} 1 & \text{if } i < 2 \\ 1 & \text{if } i = 2 \text{ and } j < 2 \\ 0 & \text{elsewhere} \end{cases} \qquad (\boldsymbol{M}^B)_{::ij} := \begin{cases} 1 & \text{if } i < 2 \\ 1 & \text{if } i = 2 \text{ and } j \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

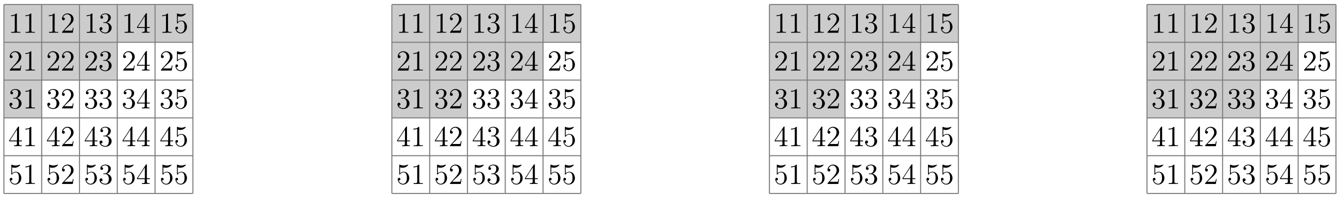
where the index starts from 1. Masking is achieved by multiplying the kernel with the binary mask (elementwise). Specify the receptive field of the output pixel that corresponds to the third row and the third column (index 33 of Figure 1) in each of the following 4 cases:

11	12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

FIGURE 1 – 5×5 convolutional feature map.

1. If we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^A$ for the second layer.
2. If we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^B$ for the second layer.
3. If we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^A$ for the second layer.
4. If we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^B$ for the second layer.

The concerned masks are $\boldsymbol{M}^A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, $\boldsymbol{M}^B = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ The receptive field is illustrated in figure 2:



11	**12**	**13**	**14**	**15**
21	**22**	**23**	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

11	**12**	**13**	**14**	**15**
21	**22**	**23**	**24**	25
31	**32**	33	34	35
41	42	43	44	45
51	52	53	54	55

11	**12**	**13**	**14**	**15**
21	**22**	**23**	**24**	25
31	**32**	33	34	35
41	42	43	44	45
51	52	53	54	55

11	**12**	**13**	**14**	**15**
21	**22**	**23**	**24**	25
31	**32**	**33**	34	35
41	42	43	44	45
51	52	53	54	55

FIGURE 2 – Receptive field under different masking schemes.

[2] An example of this is the use of masking in the Transformer architecture (Problem 3 of TP2 practical part).

Question 5 (10)**.** Let P_1 and P_0 be two probability distributions with densities f_0 and f_1 (respectively). This problem demonstrates that a optimal GAN Discriminator (i.e. one which is able to distinguish between examples from P_0 and P_1 with minimal NLL loss) can be used to express the probability density of a datapoint $\boldsymbol{x}$ under f_1, $f_1(\boldsymbol{x})$ in terms of $f_0(\boldsymbol{x})$.

Assume f_0 and f_1 have the same support. Show that $f_1(\boldsymbol{x})$ can be estimated by $f_0(\boldsymbol{x})D(\boldsymbol{x})/(1-D(\boldsymbol{x}))$ by establishing the identity $f_1(\boldsymbol{x}) = f_0(\boldsymbol{x})D^*(\boldsymbol{x})/(1-D^*(\boldsymbol{x}))$, where

$$D^* := \arg\max_D \mathbb{E}_{\boldsymbol{x}\sim P_1}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim P_0}[\log(1-D(\boldsymbol{x}))]$$

Let us write out the full integral. We combine the integrals as the densities have the same support.

$$\begin{aligned}
&= \int_x f_0(\boldsymbol{x}) \cdot \log D(\boldsymbol{x}) + \int_x f_1(\boldsymbol{x}) \cdot \log(1-D(\boldsymbol{x})) \\
&= \int_x f_0(\boldsymbol{x}) \cdot \log D(\boldsymbol{x}) + f_1(\boldsymbol{x}) \cdot \log(1-D(\boldsymbol{x}))
\end{aligned}$$

Differentiating with respect to $D(\boldsymbol{x})$ and setting to 0,

$$\begin{aligned}
0 &= \frac{f_0(\boldsymbol{x})}{D^*(\boldsymbol{x})} - \frac{f_1(\boldsymbol{x})}{1-D^*(\boldsymbol{x})} \\
0 &= \frac{f_0(\boldsymbol{x}) \cdot (1-D^*(\boldsymbol{x})) - f_1(\boldsymbol{x}) \cdot D^*(\boldsymbol{x})}{D^*(\boldsymbol{x}) \cdot 1 - D^*(\boldsymbol{x})} \\
f_1(\boldsymbol{x}) \cdot D^*(\boldsymbol{x}) &= f_0(\boldsymbol{x}) \cdot (1-D^*(\boldsymbol{x})) \\
f_1(\boldsymbol{x}) &= f_0(\boldsymbol{x}) \cdot \frac{D^*(\boldsymbol{x})}{1-D^*(\boldsymbol{x})}
\end{aligned}$$

Question 6 (5-5-6)**.** While generative adversarial networks were originally formulated as minimizing the Jensen-Shannon (JS)-divergence, the framework can be generalized to use other divergences, such as the Kullback–Leibler (KL)-divergence. In this exercise we see how KL can be approximated (bounded from below) via a function $T : \mathcal{X} \to \mathbb{R}$ (i.e. the discriminator). Let q and p be probability density functions and recall the definition of the KL divergence $D_{\text{KL}}(p||q) = \int p(x) \log \left(\frac{p(x)}{q(x)}\right) dx$.

*1. Let $R_1[T] := \mathbb{E}_p[T(x)] - E_q[e^{T(x)-1}]$.

(a) The convex conjugate of a function $f(u)$ is defined as $f^*(t) = \sup_{u \in \text{dom} f} ut - f(u)$. Show that the convex conjugate of $f(u) = u \log u$ is $f^*(t) = e^{t-1}$, and its biconjugate[3], i.e. the convex conjugate of its convex conjugate, is $f^{**}(u) := (f^*)^*(u) = u \log u$. The convex conjugate $f(u)$ is defined as:

$$\begin{aligned} f^*(t) &= \sup_{u \in \text{dom} f} ut - f(u) \\ &= \sup_{u \in \text{dom} f} ut - u \log u \end{aligned}$$

Now for a fixed value of t we maximize the function $ut - u \log u$ by setting it to 0.

$$\begin{aligned} &= \frac{d}{du}(ut - u \log u) \\ &= t - \log u - 1 \\ 0 &= t - [\log u^* + 1] \\ u^* &= e^{t-1} \end{aligned}$$

Therefore, $f^*(t) = e^{t-1}$.

To find the biconjugate, we have the following relation:

$$\begin{aligned} f^{**}(u) &= (f^*)^*(u) = u \log u \\ &= \sup_{t \in \text{dom}(u)} tu - \underbrace{f^*(t)}_{e^{t-1}} \end{aligned}$$

For a fixed value of u, we find the value of t that maximizes $tu - e^{t-1}$ by setting it to 0.

$$\begin{aligned} &= \frac{d}{dt}\left(tu - e^{t-1}\right) \\ 0 &= u - e^{t-1} \\ t^* &= \log u + 1 \end{aligned}$$

Now we substitute this back into the above equation to get:

$$\begin{aligned} f^{**}(u) &= t^* u - f^*(t^*) \\ &= \underbrace{u \cdot (\log u + 1)}_{u \log u + u} - \underbrace{e^{\log u + 1 - 1}}_{u} \\ f^{**}(u) &= u \log u \end{aligned}$$

[3] More generally, the biconjugate of f is equal to itself if f is a lower semi-continuous convex function (this is known as the **Fenchel-Monreau Theorem**).

(b) Use the fact found above to show that $D_{\text{KL}}(p||q) = \sup_T R_1[T]$, where the supremum is taken over the set of all (measurable) functions $\mathcal{X} \to \mathbb{R}$. Start from the following step

$$\sup_{T(x)} \int p(x)T(x) - q(x)e^{T(x)-1}dx = \int \sup_{t\in\mathbb{R}} p(x)t - q(x)e^{t-1}dx$$

which you don't need to prove.

$$\begin{aligned}\sup_T R_1[T] &= \int \sup_{t\in\mathbb{R}} p(x)t - q(x)e^{t-1}dx \\ &= \int \sup_{t\in\mathbb{R}} q(x) \cdot \left(t \cdot \frac{p(x)}{q(x)} - e^{t-1}\right)\end{aligned}$$

Now since $q(x) \geq 0$, $\sup q(x) \cdot F = q(x) \sup F$,

$$= \int q(x) \sup_{t\in\mathbb{R}} \left(\frac{t \cdot p(x)}{q(x)} \cdot -e^{t-1}\right)$$

Now, the convex conjugate of $f^*(t) = e^{t-1}$ from the above question and thus we have:

$$\begin{aligned}\sup_T R_1[T] &= \int q(x) \cdot \left(\frac{p(x)}{q(x)} \cdot \log \frac{p(x)}{q(x)}\right) \cdot dx \\ &= \int p(x) \cdot \log \frac{p(x)}{q(x)} \cdot dx \\ \sup_T R_1[T] &= KL(p(x)||q(x))\end{aligned}$$

*2. Let $r(x) = e^{T(x)}/\mathbb{E}_q[e^{T(x)}]$ and $R_2[T] := \mathbb{E}_p[T(x)] - \log \mathbb{E}_q[e^{T(x)}]$.

(a) Verify that rq is a proper density function, i.e. integrating to 1.

For $x \in \mathbb{R}$,

$$\begin{aligned} &= \int_x r(x) \cdot q(x) \cdot dx \\ &= \int_x \frac{e^{T(x)}}{\mathbb{E}_q[e^{T(x)}]} \cdot q(x) \cdot dx \\ &= \frac{1}{\mathbb{E}_q[e^{T(x)}]} \int_x e^{T(x)} \cdot q(x) \cdot dx \\ &= \frac{1}{\mathbb{E}_q[e^{T(x)}]} \cdot \mathbb{E}_q[e^{T(x)}] = 1 \end{aligned}$$

Therefore, rq is a proper density function.

(b) Show that $D_{\mathrm{KL}}(p||q) \geq R_2[T]$, with equality if and only if $T(x) = \log(p(x)/q(x)) + c$ where c is a constant independent of x.

3. Compare the two representations of the KL divergence. For fixed $T(x)$, $p(x)$ and $q(x)$, which one of $R_1[T]$ and $R_2[T]$ is greater than or equal to the other ?

For the domain of $x \in (0, +\infty)$, we have that $\log x \leq (x-1)$. And thus $(1 + \log x) \leq x$. Using this fact, we proceed as follows:

$$\begin{aligned}
\log \mathbb{E}_q[e^{T(x)}] &= \log E_q[e^{T(x)-1+1}] \\
&= \log \mathbb{E}_q[e^{T(x)-1} \cdot e] \\
&= \log e + \log \mathbb{E}_q[e^{T(x)-1}] \\
&= 1 + \log \mathbb{E}_q[e^{T(x)-1}] \\
1 + \log \mathbb{E}_q[e^{T(x)-1}] &\leq \mathbb{E}_q[e^{T(x)-1}]
\end{aligned}$$

For a fixed value of $T(x), p(x)$ and $q(x)$ and using the result in the above equation we have:

$$\begin{aligned}
R_2[T] &= \mathbb{E}_p[T(x)] - \log \mathbb{E}_q[e^{T(x)}] \\
\mathbb{E}_p[T(x)] - \log \mathbb{E}_q[e^{T(x)}] &\geq \mathbb{E}_p[T(x)] - \mathbb{E}_q[e^{T(x)-1}] \\
R_2[T] &\geq R_1[T]
\end{aligned}$$

Question 7 (10). Let $q, p : \mathcal{X} \to [0, \infty)$ be probability density functions with disjoint (i.e. non-overlapping) support ; more formally, $\{x \in \mathcal{X} : p(x) > 0 \text{ and } q(x) > 0\} = \varnothing$. What is the Jensen Shannon Divergence (JSD) between p and q ? Recall that JSD is defined as $D_{JS}(p||q) = \frac{1}{2}D_{KL}(p||r) + \frac{1}{2}D_{KL}(q||r)$ where $r(x) = \dfrac{p(x)+q(x)}{2}$.

Let us first compute $D_{KL}(p||r)$

$$\begin{aligned}
D_{KL}(p||r) &= \int_0^\infty p(x) \log \frac{2 \cdot p(x)}{p(x)+q(x)} \\
&= \int_0^\infty p(x) \Bigg[\log\Big(2 \cdot p(x)\Big) - \log\Big(p(x)+q(x)\Big)\Bigg] \\
&= \int_0^\infty p(x) \Bigg[\big(\log 2 + \log p(x)\big) - \log p(x) - \log\left(1 + \frac{q(x)}{p(x)}\right)\Bigg] \\
&= \int_0^\infty p(x) \log 2 + \int_0^\infty p(x) \log p(x) - \int_0^\infty p(x) \log p(x) - \int_0^\infty p(x) \log\left(1 + \frac{q(x)}{p(x)}\right) \\
&= \log 2 - \int_0^\infty p(x) \log\left(1 + \frac{q(x)}{p(x)}\right)
\end{aligned}$$

Let us examine the term $\int_0^\infty p(x) \log\left(1 + \dfrac{q(x)}{p(x)}\right)$. We are given that the two density functions are disjoint, that is, if $p(x) > 0$ then $q(x) = 0$ and vice versa. Therefore, we can divide the support into three regions:

- $p(x) > 0, q(x) = 0$
- $q(x) > 0, p(x) = 0$
- $p(x) = 0, q(x) = 0$

In each of these cases, the ratio $\dfrac{q(x)}{p(x)} = 0$[4]. As a consequence, we have $\int_0^\infty \log\Big(1+0\Big) = 0$. Thus, the term simplifies to:

$$D_{KL}(p||r) = \log 2 \tag{20}$$

$$\frac{1}{2} D_{KL}(p||r) = \frac{\log 2}{2} \tag{21}$$

We can derive $D_{KL}(q||r)$ in an identical manner, the final value of which is:

$$D_{KL}(q||r) = \frac{\log 2}{2} \tag{22}$$

Finally, we plug in 21 and 22 into the definition of the Jensen-Shannon divergence, and we get:

$$\begin{aligned}
D_{JS} &= \frac{1}{2} D_{KL}(p||r) + \frac{1}{2} D_{KL}(q||r) \\
&= \boxed{\frac{\log 2}{2} + \frac{\log 2}{2} = \log 2}
\end{aligned}$$

[4] Note that in some of these cases we have undefined scenarios when we divide by 0, but these are assumed to be 0.