IFT6135-H2019 - Representation learning

Assignment 3, Theoretical Part

Sanae Lotfi

Num de matricule (Poly) : 1968682
Num de matricule (Udem) : 20147309

4th April, 2019

Due Date: April 5th 23:59, 2019

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).
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).

Answer 1.

1. Let's consider $\boldsymbol{g}(\boldsymbol{z}_0) = \mu + \sigma \odot \boldsymbol{z}_0$, where $\mu \in \mathbb{R}^K$ and $\sigma \in \mathbb{R}^K_{>0}$. The function $\boldsymbol{z}_0 \mapsto \boldsymbol{g}(\boldsymbol{z}_0)$ is differentiable because it is linear in the components of $\boldsymbol{z}_0$. Moreover, let $\boldsymbol{z}$ be a vector in $\mathbb{R}^K$, is there a unique $\boldsymbol{z}_0$ such that $\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{z}_0)$? We have:

$$\begin{aligned} \boldsymbol{z} = \boldsymbol{g}(\boldsymbol{z}_0) &\iff \boldsymbol{z} = \mu + \sigma \odot \boldsymbol{z}_0 \\ &\iff \sigma \odot \boldsymbol{z}_0 = \boldsymbol{z} - \mu \\ &\iff diag(\sigma)\boldsymbol{z}_0 = \boldsymbol{z} - \mu \end{aligned}$$

   Since $\sigma \in \mathbb{R}^K_{>0}$, $diag(\sigma)$ is invertible and $(diag(\sigma)^{-1})_{ij} = \delta_{ij}\frac{1}{\sigma_i}$. We write:

$$\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{z}_0) \iff \boldsymbol{z}_0 = diag(\sigma)^{-1}(\boldsymbol{z} - \mu)$$

   $\boldsymbol{z}_0$ is unique by expression. Thus, we conclude that $\boldsymbol{g}$ is also bijective. We can then use the expression of the change of density:

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

Let's put $\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{z}_0)$, then we need to find the density function of $\boldsymbol{z}$. We have:

$$\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0} = \frac{\partial\Big(\mu + diag(\sigma)\boldsymbol{z}_0\Big)}{\partial \boldsymbol{z}_0} = diag(\sigma)$$

Then:

$$\begin{aligned}
q(\boldsymbol{z}) &= q(\boldsymbol{z}_0)\left|\det\left(diag(\sigma)\right)\right|^{-1} \\
&= q(diag(\sigma)^{-1}(\boldsymbol{z}-\mu))\left|\det\left(diag(\sigma)\right)\right|^{-1} \\
&= \frac{1}{\sqrt{(2\pi)^K}}\exp\Big[-\frac{1}{2}\left[diag(\sigma)^{-1}(\boldsymbol{z}-\mu)\right]^\top\left[diag(\sigma)^{-1}(\boldsymbol{z}-\mu)\right]\Big]\left|\det\left(diag(\sigma)\right)\right|^{-1} \\
&= \frac{1}{\sqrt{(2\pi)^K}}\exp\Big[-\frac{1}{2}\left(\boldsymbol{z}-\mu\right)^\top\left(diag(\sigma)^{-1}\right)^\top diag(\sigma)^{-1}(\boldsymbol{z}-\mu)\Big]\frac{1}{\det\left(diag(\sigma)\right)} \\
&= \frac{1}{\sqrt{(2\pi)^K}}\exp\Big[-\frac{1}{2}\left(\boldsymbol{z}-\mu\right)^\top(diag(\sigma)^{-1})^2(\boldsymbol{z}-\mu)\Big]\frac{1}{\sqrt{\det\left(diag(\sigma^2)\right)}} \\
&= \frac{1}{\sqrt{(2\pi)^K\ \det\left(diag(\sigma^2)\right)}}\exp\Big[-\frac{1}{2}\left(\boldsymbol{z}-\mu\right)^\top diag(\sigma^2)^{-1}(\boldsymbol{z}-\mu)\Big]
\end{aligned}$$

We conclude that $\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{z}_0)$ is distributed by $\mathcal{N}(\mu, \text{diag}(\sigma^2))$.

2. In the same fashion as the previous question, let's consider $\boldsymbol{g}(\boldsymbol{z}_0) = \mu + \boldsymbol{S}\boldsymbol{z}_0$, where $\boldsymbol{S}$ is a non-singular $K \times K$ matrix and $\mu \in \mathbb{R}^K$. The function $\boldsymbol{z}_0 \mapsto \boldsymbol{g}(\boldsymbol{z}_0)$ is differentiable because it is linear in the components of $\boldsymbol{z}_0$. Moreover, let $\boldsymbol{z}$ be a vector in $\mathbb{R}^K$, is there a unique $\boldsymbol{z}_0$ such that $\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{z}_0)$? We have:

$$\begin{aligned}
\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{z}_0) &\iff \boldsymbol{z} = \mu + \boldsymbol{S}\boldsymbol{z}_0 \\
&\iff \boldsymbol{S}\boldsymbol{z}_0 = \boldsymbol{z} - \mu
\end{aligned}$$

$\boldsymbol{S}$ is an invertible matrix, so we can write:

$$\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{z}_0) \iff \boldsymbol{z}_0 = \boldsymbol{S}^{-1}(\boldsymbol{z} - \mu)$$

$\boldsymbol{z}_0$ is unique by expression. Thus, we conclude that $\boldsymbol{g}$ is also bijective. We can then use the expression of the change of density:

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

Let's put $\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{z}_0)$, then we need to find the density function of $\boldsymbol{z}$. We have:

$$\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0} = \frac{\partial\Big(\mu + \boldsymbol{S}\boldsymbol{z}_0\Big)}{\partial \boldsymbol{z}_0} = \boldsymbol{S}$$

Then:

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

We know that: $\det(\boldsymbol{S}) = \det\left(\boldsymbol{S}^\top\right)$ and

$$|\det(\boldsymbol{S})| = \sqrt{\det(\boldsymbol{S})^2} = \sqrt{\det(\boldsymbol{S})\,\det(\boldsymbol{S}^\top)} = \sqrt{\det(\boldsymbol{S}\boldsymbol{S}^\top)}$$

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

We conclude that $\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{z}_0)$ is distributed by $\mathcal{N}(\mu, \boldsymbol{S}\boldsymbol{S}^\top)$.

Question 2 (5-5-6)**.** Consider a latent variable model $\boldsymbol{z} \sim p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$ 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 $B(\theta)$.

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 ?

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.

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)
   (b) from the computational point of view (efficiency)
   (c) in terms of memory (storage of parameters)

Answer 2.

1. We have:

$$\begin{aligned}
\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - \log p_\theta(\boldsymbol{x}) &= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - \int_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{x}) \log p_\theta(\boldsymbol{x}) d\boldsymbol{z} \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x})\right] \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}) - \log p_\theta(\boldsymbol{x})\right] \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x}|\boldsymbol{z})}{p_\theta(\boldsymbol{x})}\right] \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{z}|\boldsymbol{x})\, p_\theta(\boldsymbol{x})}{p(\boldsymbol{z}) p_\theta(\boldsymbol{x})}\right] \quad \text{(using Bayes' theorem)} \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z})}\right] \quad (p(\boldsymbol{z}) \text{ doesn't depend on } \theta) \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{z}|\boldsymbol{x})\, q(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{x})\, p(\boldsymbol{z})}\right] \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{x})} + \log \frac{q(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z})}\right] \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[-\log \frac{q(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{q(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z})}\right] \\
&= -D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})) + D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))
\end{aligned}$$

Thus, we can write:

$$\begin{aligned}
\arg\max_\theta \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] &= \arg\max_\theta \{\log p_\theta(\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})) + D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))\} \\
&= \arg\max_\theta \{\log p_\theta(\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))\} \\
&= \arg\max_\theta \{\log p_\theta(\boldsymbol{x}) + B(\theta)\}
\end{aligned}$$

where $B(\theta) = -D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$. $B(\theta)$ is non-positive since the Kullback-Leibler divergence is always positive.

2. Let's consider a finite training set $\{\boldsymbol{x}_i : i \in \{1, ..., n\}\}$, n being the size the training data. For given θ, ϕ and $\boldsymbol{x}_i$, we have:

$$\begin{aligned}
\mathcal{L}(\theta, \phi; \boldsymbol{x}_i) &= \mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x}_i \mid \boldsymbol{z})] - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}_i)||p(\boldsymbol{z})) \\
&= \log p_\theta(\boldsymbol{x}_i) - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}_i)||p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i)) \quad \text{(from previous question)}
\end{aligned}$$

For fixed θ, we define:

$$\begin{aligned}\phi^* &= \arg\max_{\phi} \sum_{i=1}^{n} \mathcal{L}(\theta, \phi; \boldsymbol{x}_i) \\ &= \arg\max_{\phi} \sum_{i=1}^{n} \log p_\theta(\boldsymbol{x}_i) - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}_i)||p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i)) \\ &= \arg\max_{\phi} \sum_{i=1}^{n} -D_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}_i)||p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i)) \\ &= \arg\min_{\phi} \sum_{i=1}^{n} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}_i)||p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i))\end{aligned}$$

and for a given $\boldsymbol{x}_i$, we define:

$$\begin{aligned}q_i^* &= \arg\max_{q_i \in \mathcal{Q}} \mathcal{L}(\theta, \phi; \boldsymbol{x}_i) \\ &= \arg\max_{q_i \in \mathcal{Q}} \log p_\theta(\boldsymbol{x}_i) - D_{\mathrm{KL}}(q_i(\boldsymbol{z})||p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i)) \\ &= \arg\max_{q_i \in \mathcal{Q}} -D_{\mathrm{KL}}(q_i(\boldsymbol{z})||p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i)) \\ &= \arg\min_{q_i \in \mathcal{Q}} D_{\mathrm{KL}}(q_i(\boldsymbol{z})||p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i))\end{aligned}$$

Since $q_{\phi^*}(\boldsymbol{z} \mid \boldsymbol{x}_i) \in \mathcal{Q}$, by definition of $q_i^* = q^*(\boldsymbol{z} \mid \boldsymbol{x}_i)$ as the global minimizer of

$$D_{\mathrm{KL}}(q(\boldsymbol{z} \mid \boldsymbol{x}_i)||p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i))$$

over $\mathcal{Q}$ for a given $\boldsymbol{x}_i$, we have:

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

3. Let's compare the two approaches:

   (a) in terms of bias of estimating the marginal likelihood via the ELBO, in the best case scenario: the bias is defined as seen in the first question as

   $$B(\theta) = -D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$$

   thus:

   $$D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})||p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i)) \leq D_{\mathrm{KL}}(q_\phi^*(\boldsymbol{z} \mid \boldsymbol{x}_i)||p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i)) \implies B(\theta)_{\boldsymbol{x}_i,\phi^*} \leq B(\theta)_{\boldsymbol{x}_i,q_i^*} \leq 0$$

   As the bias is negative, we conclude that it is closer to zero when we estimate q_i^* for each training example, compared to when we calculate ϕ^* for the entire training data. Thus, it can allow us to have a more accurate estimation of the marginal likelihood via the ELBO.

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

The comparison will depend on the algorithm we use to obtain q_i^* for each data point $\boldsymbol{x}_i$. Since generally obtaining q_i^* might involve the calculation of an integral that might be computationally heavy, we can assume that calculating ϕ^* for the entire training data is generally more efficient than calculating q_i^* for each data point $\boldsymbol{x}_i$, especially if n has a large value. Training a single neural network to get the value of ϕ^* seems to be more efficient than calculating q_i^* for each data point.

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

This will depend on whether the number of parameters of the network to get ϕ^* is higher than the size of the training set n (since we store n calculations of q_i^*). In deep learning, generally the number of parameters of the the network is higher than the size of the training set. Thus, storing the parameters for the entire training might take more memory than storing q_i^* for each element individually. The second approach seems to be better on this aspect. There is a tradeoff between the memory and the computation time to manage.

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 $\log p(\boldsymbol{x})$.
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.

Answer 3.

1. Let's 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 $p(\boldsymbol{h}|\boldsymbol{x})$ and $\boldsymbol{h}_j$, $j \in \{1, \ldots, K\}$ be i.i.d random unobserved variables. We have:

$$
\begin{aligned}
p(\boldsymbol{x}) &= \frac{1}{K}\sum_{j=1}^{K} p(\boldsymbol{x}) \\
&= \frac{1}{K}\sum_{j=1}^{K} \int_{\boldsymbol{h}_j} p(\boldsymbol{x},\boldsymbol{h}_j)d\boldsymbol{h}_j \ \text{(using the law of total probability)} \\
&= \frac{1}{K}\sum_{j=1}^{K} \int_{\boldsymbol{h}_j} \frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)} q(\boldsymbol{h}_j)d\boldsymbol{h}_j \ \text{(since: } q(\boldsymbol{h}_j)=0 \implies p(\boldsymbol{x},\boldsymbol{h}_j)=0\text{)} \\
&= \frac{1}{K}\sum_{j=1}^{K} \int_{\boldsymbol{h}_j} \frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)} q(\boldsymbol{h}_j)d\boldsymbol{h}_j \prod_{i=1,i\neq j}^{K} \int_{\boldsymbol{h}_i} q(\boldsymbol{h}_i)d\boldsymbol{h}_i \ \text{(because } \int_{\boldsymbol{h}_i} q(\boldsymbol{h}_i)d\boldsymbol{h}_i = 1\text{)} \\
&= \frac{1}{K}\sum_{j=1}^{K} \int_{\boldsymbol{h}_1}\dots\int_{\boldsymbol{h}_K} \frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)} q(\boldsymbol{h}_1)\dots q(\boldsymbol{h}_K)\, d\boldsymbol{h}_1\dots d\boldsymbol{h}_K \ \left(\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)} \text{ doesn't depend on } \boldsymbol{h}_i \text{ for } i\neq j\right) \\
&= \mathbb{E}_{\boldsymbol{h}_j\sim q(\boldsymbol{h})}\left[\frac{1}{K}\sum_{j=1}^{K}\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right]
\end{aligned}
$$

$\mathbb{E}_{\boldsymbol{h}_j\sim q(\boldsymbol{h})}\left[\frac{1}{K}\sum_{j=1}^{K}\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right]$ denotes $\mathbb{E}_{\boldsymbol{h}_1,\dots,\boldsymbol{h}_K}\left[\frac{1}{K}\sum_{j=1}^{K}\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right]$. Using Jensen's inequality, since log is a concave function:
$\log\left(p(\boldsymbol{x})\right) = \log\left(\mathbb{E}_{\boldsymbol{h}_1,\dots,\boldsymbol{h}_K}\left[\frac{1}{K}\sum_{j=1}^{K}\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right]\right) \geq \mathbb{E}_{\boldsymbol{h}_1,\dots,\boldsymbol{h}_K}\left[\log\frac{1}{K}\sum_{j=1}^{K}\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right]$

We conclude that: $\log\left(p(\boldsymbol{x})\right) \geq \mathcal{L}_K$.

2. Using Jensen's inequality, since log is a concave function, we have:

$$
\begin{aligned}
\log\left(\frac{1}{K}\sum_{j=1}^{K}\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right) &\geq \frac{1}{K}\sum_{j=1}^{K}\log\left(\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right) \\
\mathbb{E}_{\boldsymbol{h}_1,\dots,\boldsymbol{h}_K}\left[\log\left(\frac{1}{K}\sum_{j=1}^{K}\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right)\right] &\geq \frac{1}{K}\sum_{j=1}^{K}\mathbb{E}_{\boldsymbol{h}_1,\dots,\boldsymbol{h}_K}\left[\log\left(\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right)\right]
\end{aligned} \tag{1}
$$

Since $\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}$ doesn't depend on $\boldsymbol{h}_i$ for $i\neq j$, we can marginalize over the random variables $\boldsymbol{h}_i$ for $i\neq j$ and we have:

$$
\mathbb{E}_{\boldsymbol{h}_1,\dots,\boldsymbol{h}_K}\left[\log\left(\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right)\right] = \mathbb{E}_{\boldsymbol{h}_j}\left[\log\left(\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right)\right] = \mathcal{L}_1 \quad \text{(by definition of } \mathcal{L}_1\text{)}
$$

Thus, inequality (1) becomes:

$$
\mathcal{L}_K \geq \frac{1}{K}\sum_{j=1}^{K}\mathcal{L}_1 \implies \mathcal{L}_K \geq \mathcal{L}_1
$$

We obtain a tighter lower bound of the log marginal likelihood $\log p(\boldsymbol{x})$.

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.

Answer 4. The kernel obtained when we apply the mask of type A is the following:

k_{11}	k_{12}	k_{13}
k_{21}	0	0
0	0	0

The kernel obtained when we apply the mask of type B is the following:

k_{11}	k_{12}	k_{13}
k_{21}	k_{22}	0
0	0	0

Since we apply no kernel flipping and the padding size is 1 on each border, the receptive field (of the output pixel of index 33) in feature map obtained after the first layer(input $\rightarrow$ [layer 1] $\rightarrow$ feature map 1 $\rightarrow$ [layer 2] $\rightarrow$ output) is:

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

- If we use $\boldsymbol{M}^A$ for the second layer:

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 in feature map 1 if we use $\boldsymbol{M}^A$ for the second layer.

- If we use $\boldsymbol{M}^B$ for the second layer:

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 3 – Receptive field in feature map 1 if we use $\boldsymbol{M}^B$ for the second layer.

In fact, we only consider the pixels that actually participate in the calculation of the pixel value of index 33 in the output. As the kernel masked by type A has 5 inactive pixels, we are only left with 4 active pixels as can be seen in the figure 2 (same for mask of type B). The value of the pixel of index 33 from the output if we use the mask of type A in the second layer is calculated as follows: $value_{33} = k_{11} * fmap_{22} + k_{12} * fmap_{23} + k_{13} * fmap_{24} + k_{21} * fmap_{32}$ (since there is no kernel flipping).

To find the receptive field (of the output pixel of index 33) in the input obtained if we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^A$ for the second layer, we need to find the receptive field of each active pixel in (2) in the input. For example, the pixel of index 22 has the following receptive field f we use $\boldsymbol{M}^A$ for the first layer as well:

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 4 – Receptive field for pixel 22 of feature map 1 in the input if we use $\boldsymbol{M}^A$ for the first layer.

In the same fashion we find the receptive field for pixels of indexes 23, 24 and 32. Thus, after doing these **(repetitive)** calculations, we obtain the answers for the four questions of the exercise:

1. If we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^A$ for the second layer:

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 5 – Receptive field 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:

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 6 – Receptive field 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:

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 7 – Receptive field 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:

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 8 – Receptive field if we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^B$ for the second layer.

Question 5 (10)**.** Let P_0 and P_1 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

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

Answer 5. Since f_0 and f_1 have the same support, we can write:

$$\arg\max_D \mathbb{E}_{\boldsymbol{x}\sim P_1}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim P_0}[\log(1 - D(\boldsymbol{x}))] = \int_{\boldsymbol{x}} [f_1(\boldsymbol{x}) \log(D(\boldsymbol{x})) + f_0(\boldsymbol{x}) \log(1 - D(\boldsymbol{x}))]\, d\boldsymbol{x} \tag{2}$$

Since we integrate over all possible values of $\boldsymbol{x}$, in order to optimize

$$\int_{\boldsymbol{x}} [f_1(\boldsymbol{x}) \log D(\boldsymbol{x}) + f_0(\boldsymbol{x}) \log(1 - D(\boldsymbol{x}))]\, d\boldsymbol{x}$$

with respect to D, we can simply optimize

$$f_1(\boldsymbol{x}) \log D(\boldsymbol{x}) + f_0(\boldsymbol{x}) \log(1 - D(\boldsymbol{x}))$$

with respect to D for a given value of $\boldsymbol{x}$. Thus:

$$\arg\max_D \mathbb{E}_{\boldsymbol{x}\sim P_1}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim P_0}[\log(1 - D(\boldsymbol{x}))] = \arg\max_D f_1(\boldsymbol{x}) \log(D(\boldsymbol{x})) + f_0(\boldsymbol{x}) \log(1 - D(\boldsymbol{x}))$$

For a fixed $\boldsymbol{x}$, let's note $D(\boldsymbol{x}) = y$, $f_1(\boldsymbol{x}) = \alpha$ and $f_0(\boldsymbol{x}) = \beta$ since $\boldsymbol{x}$ is fixed and doesn't depend on $D(\boldsymbol{x})$. It is D that varies here, thus y varies as well. We put then $g : y \mapsto \alpha \log(y) + \beta \log(1 - y)$. g is differentiable on the interval $]0, 1[$ and we have:

$$g'(y) = \alpha \frac{1}{y} - \beta \frac{1}{1 - y}$$

A first order condition for y^* to be the global maximizer of g is that:

$$\begin{aligned} g'(y^*) = 0 &\implies \alpha \frac{1}{y^*} - \beta \frac{1}{1 - y^*} = 0 \quad (\text{we admit that } \alpha = 0 \iff \beta = 0) \\ &\implies f_1(\boldsymbol{x}) \frac{1}{D^*(\boldsymbol{x})} - f_0(\boldsymbol{x}) \frac{1}{1 - D^*(\boldsymbol{x})} = 0 \end{aligned} \tag{3}$$

Before developing this equality, let's prove that g is concave to make sure that the stationary point we obtain is a maximizer of g. We have:

$$g''(y) = -\alpha \frac{1}{y^2} - \beta \frac{1}{(1 - y)^2}$$

Recall $f_1(\boldsymbol{x}) = \alpha \geq 0$ and $f_0(\boldsymbol{x}) = \beta \geq 0$, thus $g''(y) \leq 0$ and g is concave. Thus, the stationary point that we found is a maximizer. Back to equation (3), we have:

$$f_1(\boldsymbol{x})\frac{1}{D^*(\boldsymbol{x})} - f_0(\boldsymbol{x})\frac{1}{1-D^*(\boldsymbol{x})} = 0 \implies f_1(\boldsymbol{x})(1-D^*(\boldsymbol{x})) - f_0(\boldsymbol{x})D^*(\boldsymbol{x}) = 0$$
$$\implies f_1(\boldsymbol{x}) = f_0(\boldsymbol{x})D^*(\boldsymbol{x})/(1-D^*(\boldsymbol{x}))$$

We showed

$$f_1(\boldsymbol{x}) = f_0(\boldsymbol{x})D^*(\boldsymbol{x})/(1-D^*(\boldsymbol{x}))$$

then $f_1(\boldsymbol{x})$ can be estimated by $f_0(\boldsymbol{x})D(\boldsymbol{x})/(1-D(\boldsymbol{x}))$, where:

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

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 $D_{\mathrm{KL}}(p||q) = \int p(x)\log\left(\frac{p(x)}{q(x)}\right)dx$.

*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 $f^*(t) = \sup_{u\in\mathrm{dom}f} ut - f(u)$. 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 $f^{**}(u) := (f^*)^*(u) = u\log u$.

(b) Use the fact found above to show that $D_{\mathrm{KL}}(p||q) = \sup_T R_1[T]$, 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.

*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.

(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 ?

Answer 6. 1. Let $R_1[T] := \mathbb{E}_p[T(x)] - E_q[e^{T(x)-1}]$.

(a) Let's consider the following function $f(u) = u\log u$ defined on $\mathbb{R}_+^*$. The convex conjugate of $f(u)$ is defined as $f^*(t) = \sup_{u\in\mathrm{dom}f} ut - f(u)$. For a fixed $t \in \mathbb{R}$, the function $g : u \mapsto ut - f(u)$ defined on $\mathbb{R}_+^*$ is twice differentiable and we have:

$$g'(u) = (ut - ulog(u))' = t - log(u) - 1$$

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**).

A necessary first order condition for u^* to be a global maximizer of g is $g'(u^*) = 0$, then:

$$log(u^*) = t - 1 \implies u^* = \exp(t-1)$$

Morever,

$$\forall u \in \mathbb{R}_+^*,\ g''(u) = -\frac{1}{u} < 0$$

thus g is concave and u^* is a global maximizer of g. So, we have:

$$\sup_{u \in \text{dom} f} ut - f(u) = \max_{u \in \text{dom} f} ut - f(u) = u^*t - f(u^*) = t\exp(t-1) - \exp(t-1)(t-1) = \exp(t-1)$$

The convex conjugate of $f(u)$ is:

$$f^*(t) = \sup_{u \in \text{dom} f} ut - f(u) = \exp(t-1),\ \forall t \in \mathbb{R}$$

The biconjugate of $f(u)$ is defined as:

$$f^{**}(u) := (f^*)^*(u) = \sup_{t \in \text{dom} f^*} tu - f^*(t)$$

For a fixed $u \in \mathbb{R}$ for now, the function $h : t \mapsto tu - f^*(t)$ defined on $\mathbb{R}$ is twice differentiable and we have:

$$h'(t) = (tu - \exp(t-1))' = u - \exp(t-1)$$

A necessary for order condition for t^* to be a maximizer of h is $h'(t^*) = 0$, then:

$$u = \exp(t^* - 1)$$

If $u < 0$, then:

$$\lim_{t \to -\infty} tu - \exp(t-1) = +\infty \implies sup_{t \in \text{dom} f^*} tu - f^*(t) = +\infty \implies f^{**} \text{ is not defined on } R^-$$

If $u = 0$:

$$sup_{t \in \text{dom} f^*} - f^*(t) = 0 \implies f^{**}(0) = 0$$

Let's then consider that $u \in \mathbb{R}_*^+$, we have:

$$h'(t^*) = 0 \implies u = \exp(t^* - 1) \implies t^* = log(u) + 1$$

Morever,

$$\forall t \in \mathbb{R},\ h''(t) = -\exp(t-1) < 0$$

thus h is concave and t^* is a global maximizer of h. So, we have:

$$\sup_{t \in \text{dom} f^*} tu - f^*(t) = \max_{t \in \text{dom} f^*} tu - f^*(t) = u^*t - f(u^*) = u(log(u)+1) - \exp(log(u)+1-1) = ulog(u)$$

The biconjugate of $f(u)$ is:

$$f^{**}(u) = sup_{t\in \text{dom} f^*} tu - f^*(t) = ulog(u),\ \forall u \in \mathbb{R}^+ \quad \text{with convention} \quad 0 \times log(0) = 0$$

(b) $\int \sup_{t\in\mathbb{R}} p(x)t - q(x)e^{t-1}dx$ is only defined if $\forall x \in \mathcal{X},\ q(x) = 0 \implies p(x) = 0$. In fact:

$$\exists x_0 \in \mathcal{X}, q(x_0) = 0 \text{ and } p(x_0) > 0 \implies \sup_{t\in\mathbb{R}} p(x_0)t - q(x_0)e^{t-1} = +\infty$$

Then we necessarily have $\forall x \in \mathcal{X},\ q(x) = 0 \implies p(x) = 0$, thus when $q(x) = 0$ for a given x, the contribution of the entire function $\sup_{t\in\mathbb{R}} p(x)t - q(x)e^{t-1}$ to the integral is null. So we have:

$$\begin{aligned}
\int \sup_{t\in\mathbb{R}} p(x)t - q(x)e^{t-1}dx &= \int_{x,q(x)\neq 0} \sup_{t\in\mathbb{R}} p(x)t - q(x)e^{t-1}dx \\
&= \int_{x,q(x)\neq 0} \sup_{t\in\mathbb{R}} q(x)(\frac{p(x)}{q(x)}t - e^{t-1})dx \\
&= \int_{x,q(x)\neq 0} q(x)(f^*)^*\left(\frac{p(x)}{q(x)}\right)dx
\end{aligned} \tag{4}$$

For values of x for which $p(x) = 0$, we have $\sup_{t\in\mathbb{R}} q(x)(\frac{p(x)}{q(x)}t - e^{t-1}) = \sup_{t\in\mathbb{R}} q(x)e^{t-1} = 0$, thus the contribution to the integral is null. We can write using the previous question:

$$\begin{aligned}
\int \sup_{t\in\mathbb{R}} p(x)t - q(x)e^{t-1}dx &= \int_{x,q(x)\neq 0} q(x)\frac{p(x)}{q(x)}\, log\left(\frac{p(x)}{q(x)}\right)dx \\
&= \int_{x,q(x)\neq 0} p(x)\, log\left(\frac{p(x)}{q(x)}\right)dx \\
&= D_{\text{KL}}(p||q)
\end{aligned} \tag{5}$$

We use:

$$\sup_T R_1[T] = \sup_T \mathbb{E}_p[T(x)] - E_q[e^{T(x)-1}] = \sup_{T(x)} \int p(x)T(x) - q(x)e^{T(x)-1}dx = \int \sup_{t\in\mathbb{R}} p(x)t - q(x)e^{t-1}dx$$

Thus:

$$\boxed{D_{\text{KL}}(p||q) = \sup_T R_1[T]}$$

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) Let's verify that rq is a proper density function, i.e. integrating to 1:

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

(b) We have:

$$\begin{aligned}R_2[T] &:= \mathbb{E}_p[T(X)] - \log \mathbb{E}_q[e^{T(X)}] \\ &= \int_x p(x)T(x)dx - \log \mathbb{E}_q[e^{T(X)}] \\ &= \int_x p(x)T(x)dx - \log \mathbb{E}_q[e^{T(X)}]\int_x p(x)dx \\ &= \int_x p(x)T(x)dx - \int_x p(x)\log \mathbb{E}_q[e^{T(X)}]dx \\ &= \int_x p(x)\left[T(x) - \log \mathbb{E}_q[e^{T(X)}]\right]dx \\ &= \int_x p(x)\left[log\left[e^{T(X)}\right] - \log \mathbb{E}_q[e^{T(X)}]\right]dx \\ &= \int_x p(x)log\left[\frac{e^{T(X)}}{\mathbb{E}_q[e^{T(X)}]}\right]dx \\ &= \int_x p(x)log\left[r(x)\right]dx \\ &= \int_x p(x)log\left[r(x)\times\frac{p(x)q(x)}{p(x)q(x)}\right]dx \\ &= \int_x p(x)log\left[\frac{p(x)}{q(x)}\times\frac{r(x)q(x)}{p(x)}\right]dx \\ &= \int_x p(x)\left[log\left[\frac{p(x)}{q(x)}\right] - log\left[\frac{p(x)}{r(x)q(x)}\right]\right]dx \\ &= \int_x p(x)log\left[\frac{p(x)}{q(x)}\right]dx - \int_x p(x)log\left[\frac{p(x)}{r(x)q(x)}\right]dx \\ &= D_{\mathrm{KL}}(p||q) - D_{\mathrm{KL}}(p||qr)\end{aligned}$$

Since $D_{\mathrm{KL}}(p||qr) \geq 0$ (property of KL-divergence), we get:

$$\boxed{D_{\mathrm{KL}}(p||q) = R_2[T] + D_{\mathrm{KL}}(p||qr) \geq R_2[T]}$$

and we have :

$$\begin{aligned}
D_{\mathrm{KL}}(p||q) = R_2[T] &\iff D_{\mathrm{KL}}(p||qr) = 0 \\
&\iff p = qr, \quad \text{almost everywhere} \\
&\iff p(x) = q(x)r(x),\ \forall x \in \mathcal{X} \setminus \mathcal{N}, \text{ (s.t measure of } \mathcal{N} \text{ is null: } \mu(\mathcal{N}) = 0) \\
&\iff \frac{p(x)}{q(x)} = r(x),\ \forall x \in \mathcal{X} \setminus \mathcal{N}, \mu(\mathcal{N}) = 0 \quad (\text{since } q(x) = 0 \implies p(x), \text{ see 1.(b)}) \\
&\iff \frac{p(x)}{q(x)} = e^{T(x)} / \mathbb{E}_q[e^{T(x)}],\ \forall x \in \mathcal{X} \setminus \mathcal{N}, \mu(\mathcal{N}) = 0 \\
&\iff T(x) = log\left[\frac{p(x)}{q(x)}\right] + log\left[\mathbb{E}_q[e^{T(x)}]\right],\ \forall x \in \mathcal{X} \setminus \mathcal{N}, \mu(\mathcal{N}) = 0 \\
&\iff T(x) = log\left[\frac{p(x)}{q(x)}\right] + c,\ \forall x \in \mathcal{X} \setminus \mathcal{N}, \mu(\mathcal{N}) = 0 \quad (\text{with } c = log\left[\mathbb{E}_q[e^{T(x)}]\right])
\end{aligned}$$

c doesn't depend on x because the expectation does only depend on the random variable X.

Conclusion:

$$D_{\mathrm{KL}}(p||q) = R_2[T] \iff T(x) = log\left[\frac{p(x)}{q(x)}\right] + c,\ \forall x \in \mathcal{X} \setminus \mathcal{N}, \mu(\mathcal{N}) = 0$$

where $c = log\left[\mathbb{E}_q[e^{T(x)}]\right]$ is a constant that doesn't depend on x

3. We have:

$$\begin{aligned}
R_1[T] - R_2[T] &= log(E_q[e^{T(x)}]) - E_q[e^{T(x)-1}] \\
&= log(E_q[e^{T(x)-1} \times e]) - E_q[e^{T(x)-1}] \\
&= log(e \times E_q[e^{T(x)-1}]) - E_q[e^{T(x)-1}] \\
&= log(E_q[e^{T(x)-1}]) - E_q[e^{T(x)-1}] + 1
\end{aligned}$$

If we put $y = E_q[e^{T(x)-1}] \in \mathbb{R}_*^+$, we have:

$$R_1[T] - R_2[T] = log(y) - y + 1$$

The function defined on $\mathbb{R}_*^+$ by $\phi : y \mapsto log(y) - y + 1$ is twice differentiable and we have:

$$\phi'(y) = \frac{1}{y} - 1 \quad \text{and} \quad \phi''(y) = -\frac{1}{y^2} \quad \forall y > 0$$

We have $\phi''(y) < 0$, $\forall y > 0$, thus ϕ is concave and we have:

$$\phi'(y) = \frac{1}{y} - 1 = 0 \implies y = 1$$

Thus 1 is a global maximizer of ϕ and $\phi(1) = 0$. We conclude that:

$$\phi(y) \leq 0,\ \forall y > 0$$

Thus: $R_1[T] - R_2[T] \leq 0$ for any given $T(x)$, $p(x)$ and $q(x)$.

$\boxed{R_1[T] \leq R_2[T] \text{ for fixed T(x),p(x) and q(x)}}$

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 $r(x) = \dfrac{p(x) + q(x)}{2}$.

Answer 7. We have:

$$\begin{aligned} D_{JS}(p||q) &= \frac{1}{2}D_{KL}(p||r) + \frac{1}{2}D_{KL}(q||r) \\ &= \frac{1}{2}\int_x p(x)\log(\frac{2p(x)}{p(x)+q(x)})dx + \frac{1}{2}\int_x q(x)\log(\frac{2q(x)}{q(x)+p(x)})dx \end{aligned} \tag{6}$$

We know that $D_{\mathrm{KL}}(p||q)$ is defined for continuous distributions as an integral over the support of p (where $p > 0$). In fact, whenever $p(x)$ is zero, the contribution of the corresponding term is interpreted as zero because: $\lim_{x\to 0^+} x\log(x) = 0$.

Also, for a given $x \in \mathcal{X}$ such that $p(x) > 0$, we have by hypothesis that: $q(x) = 0$, thus: $\log(\frac{2p(x)}{p(x)+q(x)}) = \log(\frac{2p(x)}{p(x)}) = \log(2)$. So we have:

$$\begin{aligned} \int_x p(x)\log(\frac{2p(x)}{p(x)+q(x)})dx &= \int_{x,p(x)=0} p(x)\log(\frac{2p(x)}{p(x)+q(x)})dx + \int_{x,p(x)>0} p(x)\log(\frac{2p(x)}{p(x)+q(x)})dx \\ &= 0 + \int_{x,p(x)>0} p(x)\log(2)dx \\ &= \log(2)\int_{x,p(x)>0} p(x)dx \\ &= \log(2)\int_x p(x)dx \\ &= \log(2)\times 1 \\ &= \log(2) \end{aligned}$$

In the same fashion, we know that $D_{\mathrm{KL}}(q||p)$ is defined as an integral over the support of q (where $q > 0$): whenever $q(x)$ is zero the contribution of the corresponding term is interpreted as zero because: $\lim_{x\to 0^+} x\log(x) = 0$.
From the other side, for a given $x \in \mathcal{X}$ such that $q(x) > 0$, we have by hypothesis that: $p(x) = 0$, thus: $\log(\frac{2q(x)}{p(x)+q(x)}) = \log(\frac{2q(x)}{q(x)}) = \log(2)$. So we have:

$$\begin{aligned}
\int_x q(x)\log(\frac{2q(x)}{p(x)+q(x)})dx &= \int_{x,q(x)=0} q(x)\log(\frac{2q(x)}{p(x)+q(x)})dx + \int_{x,q(x)>0} q(x)\log(\frac{2q(x)}{p(x)+q(x)})dx \\
&= 0 + \int_{x,q(x)>0} q(x)\log(2)dx \\
&= \log(2)\int_{x,q(x)>0} q(x)dx \\
&= \log(2)\int_x q(x)dx \\
&= \log(2)\times 1 \\
&= \log(2)
\end{aligned}$$

Back to equation (6), we have:

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