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

$$\begin{aligned}
&\boldsymbol{g}(\boldsymbol{z}_0) = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{z}_0 \\
&\implies \left| \det \left(\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0} \right) \right|^{-1} = \left| \det \left(\frac{\partial}{\partial \boldsymbol{z}_0} (\boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{z}_0) \right) \right|^{-1} = |\text{diag}(\boldsymbol{\sigma})|^{-1} \\
&= \frac{1}{|\text{diag}(\boldsymbol{\sigma})|} = \frac{1}{\sqrt{|\text{diag}(\boldsymbol{\sigma}^2)|}} \qquad \text{[where } |.| \text{ represents determinant]}
\end{aligned}$$

$$q(\boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K) \implies q(\boldsymbol{z}_0) = \tfrac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2} \boldsymbol{z}_0^T \boldsymbol{z}_0)$$

$$\begin{aligned}
&\boldsymbol{g}(\boldsymbol{z}_0) = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{z}_0 \implies \boldsymbol{z}_0 = \frac{\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \qquad \text{[where } / \boldsymbol{\sigma} \text{ is element-wise division]} \\
\implies\ & q(\boldsymbol{z}_0) = \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{1}{2} \Big(\frac{\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \Big)^T \Big(\frac{\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \Big) \right) \\
&= \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{1}{2} \Big(\frac{\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu}}{\boldsymbol{\sigma}^2} \Big)^T \Big(\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu} \Big) \right) \\
&= \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{1}{2} \Big(\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu} \Big)^T \big(\text{diag}(\boldsymbol{\sigma}^2) \big)^{-1} \Big(\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu} \Big) \right)
\end{aligned}$$

$$\begin{aligned}
\therefore q(\boldsymbol{g}(\boldsymbol{z}_0)) &= q(\boldsymbol{z}_0) \left| \det \left(\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0} \right) \right|^{-1} \qquad \text{[from Equation 1]} \\
&= \frac{1}{\sqrt{2\pi\, |\text{diag}(\boldsymbol{\sigma}^2)|}} \exp \left(-\frac{1}{2} \Big(\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu} \Big)^T \big(\text{diag}(\boldsymbol{\sigma}^2) \big)^{-1} \Big(\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu} \Big) \right) \\
&= \mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2))
\end{aligned}$$

2. Similar to 1. :

$$
\begin{aligned}
&\boldsymbol{g}(\boldsymbol{z}_0) = \boldsymbol{\mu} + \boldsymbol{S}\boldsymbol{z}_0 \\
&\implies \left|\det\left(\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0}\right)\right|^{-1} = \left|\det\left(\frac{\partial}{\partial \boldsymbol{z}_0}(\boldsymbol{\mu} + \boldsymbol{S}\boldsymbol{z}_0)\right)\right|^{-1} = |\boldsymbol{S}|^{-1} \\
&= \frac{1}{|\boldsymbol{S}|} \qquad \text{[where } |.| \text{ represents determinant]}
\end{aligned}
$$

$$
\begin{aligned}
&\boldsymbol{g}(\boldsymbol{z}_0) = \boldsymbol{\mu} + \boldsymbol{S}\boldsymbol{z}_0 \implies \boldsymbol{z}_0 = \boldsymbol{S}^{-1}\big(\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu}\big) \\
\implies &q(\boldsymbol{z}_0) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\Big(\boldsymbol{S}^{-1}\big(\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu}\big)\Big)^T \Big(\boldsymbol{S}^{-1}\big(\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu}\big)\Big)\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\big(\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu}\big)^T \boldsymbol{S}^{-T}\boldsymbol{S}^{-1}\big(\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu}\big)\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\big(\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu}\big)^T (\boldsymbol{S}\boldsymbol{S}^T)^{-1}\big(\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu}\big)\right)
\end{aligned}
$$

$$
\begin{aligned}
\therefore q(\boldsymbol{g}(\boldsymbol{z}_0)) &= q(\boldsymbol{z}_0)\left|\det\left(\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0}\right)\right|^{-1} \qquad \text{[from Equation 1]} \\
&= \frac{1}{\sqrt{2\pi}|\boldsymbol{S}|} \exp\left(-\frac{1}{2}\big(\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu}\big)^T (\boldsymbol{S}\boldsymbol{S}^T)^{-1}\big(\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu}\big)\right) \\
&= \frac{1}{\sqrt{2\pi|\boldsymbol{S}\boldsymbol{S}^T|}} \exp\left(-\frac{1}{2}\big(\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu}\big)^T (\boldsymbol{S}\boldsymbol{S}^T)^{-1}\big(\boldsymbol{g}(\boldsymbol{z}_0) - \boldsymbol{\mu}\big)\right) \\
&= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S}\boldsymbol{S}^T)
\end{aligned}
$$

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 ?
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.

$$\begin{aligned}
\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] &= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{z})}\right] = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{z})}\right] \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{q(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{z})}\right] \left[\because p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x}), \text{ and } p(\boldsymbol{z}|\boldsymbol{x}) = q_\phi(\boldsymbol{z}|\boldsymbol{x})\right] \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{x})\right] + \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{q(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z})}\right] \\
&= \log p(\boldsymbol{x}) + D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))
\end{aligned}$$

$\therefore \arg\max_\theta\{\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]\} = \arg\max_\theta\{\log p_\theta(\boldsymbol{x}) + B(\theta)\}$, where $B(\theta) = D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))$.

2. Since θ is fixed, maximizing ELBO $\implies$ minimizing $D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))$.
   For a particular instance $\boldsymbol{x}_i$, q_i^* maximizes the ELBO $\implies$ the KL-divergence between q_i^* and $p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)$ is minimum. Hence, any other q_ϕ, such as q_{ϕ^*}, would only have a minimum possible KL-divergence with $p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)$ equal to $D_{\mathrm{KL}}(q_i^*(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))$.
   $\therefore D_{\mathrm{KL}}(q_i^*(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) \leq D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))$

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. (a) Since the bias of estimating the marginal likelihood via the ELBO *is* the KL-divergence, the conclusion of the second subquestion is the answer for this question as well.
   $D_{\mathrm{KL}}(q_i^*(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) \leq D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))$

   (b) In both cases, per iteration, there are the same number of parameters to be calculated. There is no difference in the number of calculations to be made per iteration. Thus, per-iteration efficiency is the same. However, the number of iterations required might be different. Assuming it takes the same number of iterations to obtain each optimal q_i^*, as the q^* in the amortized case, the non-amortized case requires n times as many iterations. Hence, the non-amortized case is computationally less efficient.

   (c) Since in 2, it is assumed that a separate q_i^* is modeled for each example, there are n times as many parameters in this case than in the amortized case. Hence, the amortized case is better from a storage point of view.

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. "Importance Weighted Autoencoders"

1.

$$\begin{aligned}
\mathcal{L}_K &= \mathbb{E}_{\boldsymbol{h}_j \sim q(\boldsymbol{h})} \left[\log \frac{1}{K} \sum_{j=1}^{K} \frac{p(\boldsymbol{x}, \boldsymbol{h}_j)}{q(\boldsymbol{h}_j)} \right] \\
&\leq \log \mathbb{E}_{\boldsymbol{h}_j \sim q(\boldsymbol{h})} \left[\frac{1}{K} \sum_{j=1}^{K} \frac{p(\boldsymbol{x}, \boldsymbol{h}_j)}{q(\boldsymbol{h}_j)} \right] \qquad \text{[by Jensen's inequality]} \\
&= \log \frac{1}{K} \sum_{j=1}^{K} \mathbb{E}_{\boldsymbol{h}_j \sim q(\boldsymbol{h})} \left[\frac{p(\boldsymbol{x}, \boldsymbol{h}_j)}{q(\boldsymbol{h}_j)} \right] \qquad \left[\because \mathbb{E} \left[\sum x \right] = \sum \mathbb{E}[x] \right] \\
&= \log \frac{1}{K} \sum_{j=1}^{K} p(\boldsymbol{x}) \qquad \left[\because \mathbb{E}_{\boldsymbol{h}_j \sim q(\boldsymbol{h})} \left[\frac{p(\boldsymbol{x}, \boldsymbol{h}_j)}{q(\boldsymbol{h}_j)} \right] = p(\boldsymbol{x}) \right] \\
&= \log p(\boldsymbol{x}) \cancel{\left(\frac{1}{K} \sum_{j=1}^{K} 1 \right)}^{1} = \log p(\boldsymbol{x})
\end{aligned}$$

$\therefore \mathcal{L}_K \leq \log p(\boldsymbol{x}) \implies \mathcal{L}_K$ is a lower bound on $p(\boldsymbol{x})$.

2. Let $\{a_1, a_2, ..., a_K\}$ be a set of K numbers. Let i be an integer uniformly sampled from $\{1, 2, ..., K\}$. We know that :

$$\mathbb{E}_{i\sim U(1,K)}[a_i] = \frac{1}{K}\sum_{j=1}^{K} a_j$$

We use this property to simplify $\mathcal{L}_K$:

$$\begin{aligned}
\mathcal{L}_K &= \mathbb{E}_{\boldsymbol{h}_j\sim q(\boldsymbol{h})}\left[\log \frac{1}{K}\sum_{j=1}^{K}\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right] \\
&= \mathbb{E}_{\boldsymbol{h}_j\sim q(\boldsymbol{h})}\left[\log \mathbb{E}_{i\sim U(1,K)}\Big[\frac{p(\boldsymbol{x},\boldsymbol{h}_i)}{q(\boldsymbol{h}_i)}\Big]\right] \qquad \text{[as seen above]} \\
&\geq \mathbb{E}_{\boldsymbol{h}_j\sim q(\boldsymbol{h})}\left[\mathbb{E}_{i\sim U(1,K)}\Big[\log\frac{p(\boldsymbol{x},\boldsymbol{h}_i)}{q(\boldsymbol{h}_i)}\Big]\right] \qquad \text{[by Jensen's inequality]} \\
&= \mathbb{E}_{\boldsymbol{h}_j\sim q(\boldsymbol{h})}\left[\frac{1}{K}\sum_{j=1}^{K}\log\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right] \qquad \text{[as seen above]} \\
&= \frac{1}{K}\sum_{j=1}^{K}\mathbb{E}_{\boldsymbol{h}_j\sim q(\boldsymbol{h})}\left[\log\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right] \qquad \Big[\because \mathbb{E}\Big[\sum x\Big] = \sum \mathbb{E}[x]\Big] \\
&= \frac{1}{K}\sum_{j=1}^{K}\mathcal{L}_1 = \mathcal{L}_1 \qquad \Big[\because \mathbb{E}_{\boldsymbol{h}_j\sim q(\boldsymbol{h})}\left[\log\frac{p(\boldsymbol{x},\boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right] = \mathcal{L}_1\Big]
\end{aligned}$$

$\therefore \mathcal{L}_K \geq \mathcal{L}_1$; i.e. $\mathcal{L}_K$ 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.

Answer 4. We require the receptive field for index 33. Since a 3×3 kernel is being used, the pixels in the image after the first convolutional layer that contribute to the pixel at index 33 are, respectively for MaskA and MaskB used in the second convolutional 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

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 of pixel 33 for Mask A and Mask B

Thus, only pixels 22, 23, 24, and 32 in the hidden image need to be considered to compute the receptive field in case of Mask A in the 2nd convolutional layer, and only pixels 22, 23, 24, 32, 33 in case of Mask B.

The receptive fields of these pixels in the hidden image on the input image in the case of Mask A and Mask B in the first convolutional layer are :

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

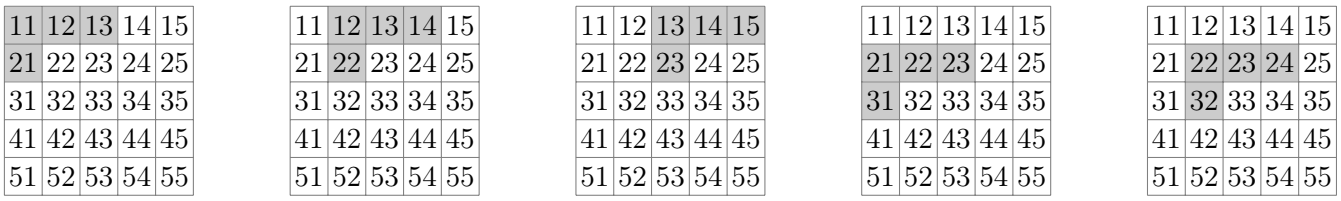

FIGURE 3 – Receptive fields of pixels 22, 23, 24, 32, 33 for Mask A

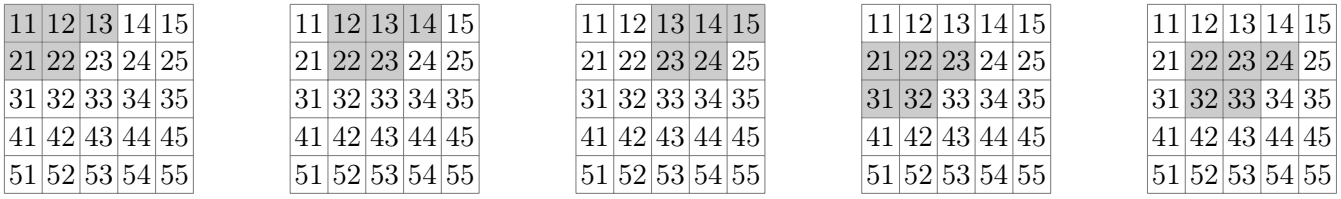

FIGURE 4 – Receptive fields of pixels 22, 23, 24, 32, 33 for Mask B

1. If we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^A$ for the second layer : we combine the receptive fields of pixels 22, 23, 24, and 32 from Fig. 3.
2. If we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^B$ for the second layer : we combine the receptive fields of pixels 22, 23, 24, 32, and 33 from Fig. 3.
3. If we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^A$ for the second layer : we combine the receptive fields of pixels 22, 23, 24, and 32 from Fig. 4.
4. If we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^B$ for the second layer : we combine the receptive fields of pixels 22, 23, 24, 32, and 33 from Fig. 4.

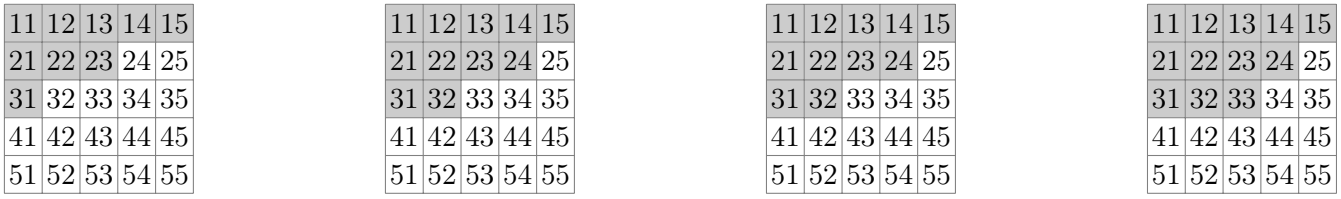

FIGURE 5 – Receptive field under different masking schemes.

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

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

Answer 5.

$$\begin{aligned}
D^* &= \arg\max_D \left(\mathbb{E}_{\boldsymbol{x}\sim P_1}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim P_0}[\log(1-D(\boldsymbol{x}))]\right) \\
&= \arg\max_D \left(\int_{\boldsymbol{x}} \log D(\boldsymbol{x})\ f_1(\boldsymbol{x})\ dx \ +\ \int_{\boldsymbol{x}} \log(1-D(\boldsymbol{x}))\ f_0(\boldsymbol{x})\ dx\right) \\
&= \arg\max_D \int_{\boldsymbol{x}} \left(\log D(\boldsymbol{x})\ f_1(\boldsymbol{x}) \ +\ \log(1-D(\boldsymbol{x}))\ f_0(\boldsymbol{x})\right) dx \\
&\quad [\because f_0(\boldsymbol{x}) \text{ and } f_1(\boldsymbol{x}) \text{ have the same support }]
\end{aligned}$$

We know that the $D(x)$ that maximizes the above integral, i.e. $D^*(\boldsymbol{x})$, makes the differentiation of the formula in the parantheses equal to 0 :

$$\begin{aligned}
&\frac{d}{dD^*(\boldsymbol{x})}\Big(\log D^*(\boldsymbol{x})\ f_1(\boldsymbol{x}) \ +\ \log(1-D^*(\boldsymbol{x}))\ f_0(\boldsymbol{x})\Big) = 0 \\
&\implies \frac{f_1(\boldsymbol{x})}{D^*(\boldsymbol{x})} - \frac{f_0(\boldsymbol{x})}{1-D^*(\boldsymbol{x})} = 0 \\
&\implies f_1(\boldsymbol{x}) = f_0(\boldsymbol{x})\frac{D^*(\boldsymbol{x})}{1-D^*(\boldsymbol{x})}
\end{aligned}$$

$\therefore$ Having established the identity $f_1(\boldsymbol{x}) = f_0(\boldsymbol{x})D^*(\boldsymbol{x})/(1-D^*(\boldsymbol{x}))$, we can see that $f_1(\boldsymbol{x})$ can be estimated by $f_0(\boldsymbol{x})D(\boldsymbol{x})/(1-D(\boldsymbol{x}))$

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

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

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

To find the supremum, let us differentiate $ut - u \log u$ w.r.t. u and equate it to 0 :

$\frac{d}{du}(ut - u \log u) = 0 \implies t - \log u^* - 1 = 0 \implies u^* = e^{t-1}$

$\implies f^*(t) = e^{t-1}t - e^{t-1}\log(e^{t-1}) = e^{t-1}t - e^{t-1}(t-1) = \cancel{e^{t-1}t} - \cancel{e^{t-1}t} + e^{t-1} = e^{t-1}$

$\therefore f^*(t) = e^{t-1}$

$f^{**}(u) = \sup_{t \in \mathrm{dom} f^*} tu - f^*(t) = \sup_{t \in \mathrm{dom} f^*} tu - e^{t-1}$

$\frac{d}{dt}(tu - e^{t-1}) = 0 \implies u - e^{t^*-1} = 0 \implies t^* - 1 = \log u \implies t^* = 1 + \log u$

$\implies f^{**}(u) = (1 + \log u)u - e^{(1+\log u)-1} = u + u \log u - e^{\log u} = \cancel{u} + u \log u - \cancel{u} = u \log u$

$\therefore f^{**}(u) = u \log u$

(b) To prove : $\sup_T R_1[T] = D_{KL}(p||q)$

$$\begin{aligned}
\sup_T R_1[T] = \sup_{T(x)} \mathbb{E}_p[T(x)] - \mathbb{E}_q[e^{T(x)-1}] &= \sup_{T(x)} \int p(x)T(x) - q(x)e^{T(x)-1}dx \\
&= \int \sup_{t \in \mathbb{R}} p(x)t - q(x)e^{t-1}dx = \int q(x) \sup_{t \in \mathbb{R}} (t\left(\frac{p(x)}{q(x)}\right) - e^{t-1})\, dx \\
&= \int \cancel{q(x)} \left(\frac{p(x)}{\cancel{q(x)}}\right) \log\left(\frac{p(x)}{q(x)}\right) dx \qquad \left[\because f(t) = e^{t-1} \implies f^*(u) = u \log u\right] \\
&= \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx = D_{KL}(p||q)
\end{aligned}$$

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-Moreau Theorem**).

2. (a) $\int r(x)\, q(x)\, dx = \int \frac{e^{T(x)}}{\mathbb{E}_q[e^{T(x)}]}\, q(x)\, dx = \frac{1}{\mathbb{E}_q[e^{T(x)}]} \int e^{T(x)}\, q(x)\, dx = \frac{1}{\mathbb{E}_q[e^{T(x)}]} \cancel{\mathbb{E}_q[e^{T(x)}]} = 1$
$\therefore rq$ is a proper density function, i.e. it integrates to 1.

(b)

$$\begin{aligned} R_2[T] &= \mathbb{E}_p[T(x)] - \log \mathbb{E}_q[e^{T(x)}] = \mathbb{E}_p[\log e^{T(x)}] - \log \mathbb{E}_q[e^{T(x)}] \\ &= \mathbb{E}_p\big[\log e^{T(x)} - \log \mathbb{E}_q[e^{T(x)}]\big] \qquad [\because \log \mathbb{E}_q[e^{T(x)}] \text{ is a constant w.r.t } p] \\ &= \mathbb{E}_p\left[\log \frac{e^{T(x)}}{\mathbb{E}_q[e^{T(x)}]}\right] = \mathbb{E}_p\left[\log \frac{e^{T(x)}}{\mathbb{E}_q[e^{T(x)}]} \frac{q(x)}{q(x)}\right] \\ &= \mathbb{E}_p\left[\log \frac{rq(x)}{q(x)}\right] \end{aligned}$$

Let Δ be the difference between $D_{\mathrm{KL}}(p \,||\, q)$ and $R_2[T]$.

$$\begin{aligned} \Delta &= D_{\mathrm{KL}}(p \,||\, q) - R_2[T] \\ &= \mathbb{E}_p\left[\log \frac{p(x)}{q(x)}\right] - \mathbb{E}_p\left[\log \frac{rq(x)}{q(x)}\right] \\ &= \mathbb{E}_p\left[\log \frac{p(x)}{rq(x)}\right] \\ &= D_{\mathrm{KL}}(p \,||\, rq) \end{aligned}$$

$\because$ KL divergence is always ≥ 0, $\Delta \geq 0 \implies D_{\mathrm{KL}}(p \,||\, q) \geq R_2[T]$
$\therefore D_{\mathrm{KL}}(p \,||\, q) \geq R_2[T]$

For the equality to hold,

$$\begin{aligned} D_{\mathrm{KL}}(p \,||\, rq) = 0 &\iff p(x) = rq(x) \\ &\iff p(x) = \frac{e^{T(x)}}{\mathbb{E}_q[e^{T(x)}]} q(x) \\ &\iff \frac{p(x)}{q(x)} = \frac{e^{T(x)}}{\mathbb{E}_q[e^{T(x)}]} \\ &\iff \log\left(\frac{p(x)}{q(x)}\right) = T(x) - \log \mathbb{E}_q[e^{T(x)}] \\ &\iff T(x) = \log\left(\frac{p(x)}{q(x)}\right) + c \end{aligned}$$

$\therefore D_{\mathrm{KL}}(p \,||\, q) \geq R_2[T] \iff T(x) = \log\left(\frac{p(x)}{q(x)}\right) + c$

3. $R_1[T] \;?\; R_2[T] \implies \mathbb{E}_p[T(x)] - \mathbb{E}_q[e^{T(x)-1}] \;?\; \mathbb{E}_p[T(x)] - \log \mathbb{E}_q[e^{T(x)}] \implies \log \mathbb{E}_q[e^{T(x)}] \;?\; \mathbb{E}_q[e^{T(x)-1}]$
Let $\mathbb{E}_q[e^{T(x)-1}] = a \implies \frac{\mathbb{E}_q[e^{T(x)}]}{e} = ea \implies \log \mathbb{E}_q[e^{T(x)}] = 1 + \log a$
$\implies R_1[T] \;?\; R_2[T] \iff 1 + \log a \;?\; a$
We know that $\log a \leq a - 1 \implies 1 + \log a \leq a$.
$\therefore R_1[T] \leq R_2[T]$

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. Recall that $D_{\mathrm{KL}}(p||q) = \int_x p(x) \log \frac{p(x)}{q(x)} dx$.

$$
\begin{aligned}
D_{JS}(p||q) &= \frac{1}{2}D_{KL}(p||r) + \frac{1}{2}D_{KL}(q||r) \\
&= \frac{1}{2}\int_{x|p(x)>0} p(x) \log \frac{p(x)}{r(x)}\, dx + \frac{1}{2}\int_{x|q(x)>0} q(x) \log \frac{q(x)}{r(x)}\, dx \\
&= \frac{1}{2}\int_{x|p(x)>0} p(x) \log \frac{2\cancel{p(x)}}{\cancel{p(x)} + \cancel{q(x)}^{0}} dx \;+\; \frac{1}{2}\int_{x|q(x)>0} q(x) \log \frac{2\cancel{q(x)}}{\cancel{p(x)}^{0} + \cancel{q(x)}} dx \\
&= \frac{1}{2}\log 2 \int_{x|p(x)>0} p(x)\, dx \;+\; \frac{1}{2}\log 2 \int_{x|q(x)>0} q(x)\, dx \\
&= \frac{1}{2}\log 2 + \frac{1}{2}\log 2 \\
&= \log 2
\end{aligned}
$$