Due Date: April 29th 23:59, 2020

Instructions

- *For all questions, show your work!*
- *Please use a document preparation system such as LaTeX, unless noted otherwise.*
- *Unless noted that questions are related, assume that notation and definitions for each question are self-contained and independent.*
- *Submit your answers electronically via Gradescope.*
- ***TAs for this assignment are Samuel Lavoie, Jae Hyun Lim, Sanae Lotfi.***

This assignment covers mathematical and algorithmic techniques underlying the four most popular families of deep generative models. Thus, we explore autoregressive models (Question 1), reparameterization trick (Question 2), variational autoencoders (VAEs, Questions 3-4), normalizing flows (Question 5), and generative adversarial networks (GANs, Question 6).

Question 1 (4-4-4-4)**.** One way to enforce autoregressive conditioning is via masking the weight parameters.[1] Consider a two-hidden-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

$$(\boldsymbol{M}^A)_{::ij} := \begin{cases} 1 & \text{if } i = 2 \text{ and } j < 2 \\ 1 & \text{if } i = 3 \\ 0 & \text{elsewhere} \end{cases} \qquad (\boldsymbol{M}^B)_{::ij} := \begin{cases} 1 & \text{if } i = 2 \text{ and } j \leq 2 \\ 1 & \text{if } i = 3 \\ 0 & \text{elsewhere} \end{cases}$$

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 fourth column (index 34 of Figure 1 (Left)) 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

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 – (Left) 5×5 convolutional feature map. (Right) Template answer.

1. If we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^A$ for the second layer.
   Using $\boldsymbol{M}^A$ for the first and second layer, below is the receptive field of pixel 34,

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 – (Left) Receptive field of output pixel 34 when Mask-A is applied to both the layers.

1. An example of this is the use of masking in the Transformer architecture (Problem 3 of HW2 practical part).

2. If we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^B$ for the second layer. Using $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^B$ for second layer, below is the receptive field of pixel 34,

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 of output pixel-34 when Mask-A is applied to layer-1 and Mask-B is applied to Layer-2.

3. If we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^A$ for the second layer. Using $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^A$ for second layer, below is the receptive field of pixel 34,

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 of output pixel-34 when Mask-B is applied to layer-1 and Mask-A is applied to Layer-2.

4. If we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^B$ for the second layer.
Using $\boldsymbol{M}^B$ for the first layer and second layer, below is the receptive field of pixel 34,

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 of output pixel-34 when Mask-B is applied to layer-1 and Layer-2.

Question 2 (6-3-6-3)**.** Reparameterization trick is a standard technique that makes the samples of a random variable differentiable. The trick represents the random variable as a simple mapping from another random variable drawn from some simple distribution [2]. If the reparameterization is a bijective function, the induced density of the resulting random variable can be computed using the change-of-variable density formula, whose computation requires evaluating the determinant of the Jacobian of the mapping.

Consider a random vector $Z \in \mathbb{R}^K$ with a density function $q(\boldsymbol{z}; \phi)$ and a random variable $Z_0 \in \mathbb{R}^K$ having a ϕ-independent density function $q(\boldsymbol{z}_0)$. We want to find a deterministic function $\boldsymbol{g} : \mathbb{R}^K \to \mathbb{R}^K$ that depends on ϕ, to transform Z_0, such that the induced distribution of the transformation

2. More specifically, these mapping should be differentiable wrt the density function's parameters.

has the same density as Z. Recall the change of density for a bijective, differentiable $\boldsymbol{g}$:

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

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 $\sigma \in \mathbb{R}^K_{>0}$. Note that $\odot$ is element-wise product.

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

$$\begin{aligned}\boldsymbol{g}(\boldsymbol{z}_0) = \mu + \sigma \odot \boldsymbol{z}_0 &\implies \boldsymbol{z}_0 = \frac{\boldsymbol{g}(\boldsymbol{z}_0) - \mu}{\sigma} \quad \text{(division with } \sigma \text{ is element wise)} \\ \boldsymbol{q}(\boldsymbol{z}_0) &= \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}\boldsymbol{z}_0^T \boldsymbol{z}_0) \\ &= \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}(\frac{\boldsymbol{g}(\boldsymbol{z}_0) - \mu}{\sigma})^T(\frac{\boldsymbol{g}(\boldsymbol{z}_0) - \mu}{\sigma})) \\ &= \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu)^T(diag(\sigma^2)^{-1})(\boldsymbol{g}(\boldsymbol{z}_0) - \mu)) \\ \implies q(\boldsymbol{g}(\boldsymbol{z}_0)) &= q(\boldsymbol{z}_0)\left|\det\left(\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0}\right)\right|^{-1} \\ &= \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu)^T(diag(\sigma^2)^{-1}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu))\frac{1}{\sqrt{diag(\sigma^2)}} \\ \implies &\boxed{q(\boldsymbol{g}(\boldsymbol{z}_0)) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu)^T(diag(\sigma^2)^{-1}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu)) = \mathcal{N}(\mu, diag(\sigma^2))}\end{aligned}$$

2. Compute the time complexity of evaluating $|\det \boldsymbol{J}_{\boldsymbol{z}_0}\boldsymbol{g}(\boldsymbol{z}_0)|$ when $\boldsymbol{g}(\boldsymbol{z}_0) = \mu + \sigma \odot \boldsymbol{z}_0$. Use the big $\mathcal{O}$ notation and expressive the time complexity as a function of K.

   In this case, $|\det \boldsymbol{J}_{\boldsymbol{z}_0}\boldsymbol{g}(\boldsymbol{z}_0)| = \sigma^K$, computing power of a number is of log complexity. Hence, $|\det \boldsymbol{J}_{\boldsymbol{z}_0}\boldsymbol{g}(\boldsymbol{z}_0)|$ can be calculated in $\mathcal{O}(logK)$.

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

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

$$
\begin{aligned}
\boldsymbol{g}(\boldsymbol{z}_0) &= \mu + \boldsymbol{S}\boldsymbol{z}_0 \implies \boldsymbol{z}_0 = \boldsymbol{S}^{-1}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu) \\
\boldsymbol{q}(\boldsymbol{z}_0) &= \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}\boldsymbol{z}_0^T \boldsymbol{z}0) \\
&= \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}((\boldsymbol{S}^{-1}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu)))^T(\boldsymbol{S}^{-1}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu))) \\
&= \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}((\boldsymbol{g}(\boldsymbol{z}_0) - \mu)^T(\boldsymbol{S}^{-1})^T\boldsymbol{S}^{-1}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu))) \\
&= \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}((\boldsymbol{g}(\boldsymbol{z}_0) - \mu)^T(\boldsymbol{S}\boldsymbol{S}^T)^{-1}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu))) \\
\implies q(\boldsymbol{g}(\boldsymbol{z}_0)) &= q(\boldsymbol{z}_0)\left|\det\left(\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0}\right)\right|^{-1} \\
&= \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu)^T(\boldsymbol{S}\boldsymbol{S}^T)^{-1}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu))\frac{1}{|\boldsymbol{S}|} \\
\implies & \boxed{q(\boldsymbol{g}(\boldsymbol{z}_0)) = \frac{1}{\sqrt{2\pi|\boldsymbol{S}\boldsymbol{S}^T|}} exp(-\frac{1}{2}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu)^T(\boldsymbol{S}\boldsymbol{S}^T)^{-1}(\boldsymbol{g}(\boldsymbol{z}_0) - \mu)) = \mathcal{N}(\mu, \boldsymbol{S}\boldsymbol{S}^T)}
\end{aligned}
$$

4. The time complexity of the general Jacobian determinant is at least $\mathcal{O}(K^{2.373})$ [3]. Assume instead $\boldsymbol{g}(\boldsymbol{z}_0) = \mu + \boldsymbol{S}\boldsymbol{z}_0$ with $\boldsymbol{S}$ being a $K \times K$ lower triangular matrix; i.e. $\boldsymbol{S}_{ij} = 0$ for $j > i$, and $\boldsymbol{S}_{ii} > 0$. What is the time complexity of evaluating $|\det \boldsymbol{J}_{\boldsymbol{z}_0}\boldsymbol{g}(\boldsymbol{z}_0)|$?
   Here, $\boldsymbol{S}$ is a lower-triangular matrix, hence $\boldsymbol{S}^{-1}$ is a lower triangular matrix too. So $\boldsymbol{J}_{\boldsymbol{z}_0}\boldsymbol{g}(\boldsymbol{z}_0)$ is a lower-triangular matrix too. Computing determinant of a lower triangular matrix is multiplication of all the diagonal elements. There are total K such elements, hence the time complexity of evaluating $|\det \boldsymbol{J}_{\boldsymbol{z}_0}\boldsymbol{g}(\boldsymbol{z}_0)|$ is $\mathcal{O}(K)$.

Question 3 (5-5-6)**.** Consider a latent variable model $p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})dz$, where $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$ and $\boldsymbol{z} \in \mathbb{R}^K$.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 $\boldsymbol{x}$. [4] 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 (ECLL)

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

   for a fixed $q(\boldsymbol{z}|\boldsymbol{x})$, wrt the model parameter θ, is equivalent to maximizing

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

3. `https://en.wikipedia.org/wiki/Computational_complexity_of_mathematical_operations`

4. 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.

This means the maximizer of the ECLL coincides with that of the marginal likelihood only if $q(\boldsymbol{z}|\boldsymbol{x})$ perfectly matches $p(\boldsymbol{z}|\boldsymbol{x})$.

$$\begin{aligned}
\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})] &= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x},\boldsymbol{z})] \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})] \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{z}|\boldsymbol{x})] + \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x})] \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{z}|\boldsymbol{x})] + \log p_\theta(\boldsymbol{x}) \text{ (Because } \log p_\theta(\boldsymbol{x}) \text{ is not dependent on z)} \\
&= \log p_\theta(\boldsymbol{x}) + \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log \frac{p_\theta(\boldsymbol{z}|\boldsymbol{x})q(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{x})}] \\
&= \log p_\theta(\boldsymbol{x}) + \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log \frac{p_\theta(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{x})}] + \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log q(\boldsymbol{z}|\boldsymbol{x})] \\
&= \log p_\theta(\boldsymbol{x}) - \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log \frac{q(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})}] + \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log q(\boldsymbol{z}|\boldsymbol{x})] \\
&\text{(Since } \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log q(\boldsymbol{z}|\boldsymbol{x})] \text{ is constant with respect to } \theta,\ \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log q(\boldsymbol{z}|\boldsymbol{x})] = C) \\
&= \log p_\theta(\boldsymbol{x}) - \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log \frac{q(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})}] + C \\
&= \log p_\theta(\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})) + C
\end{aligned}$$

Hence,

$$\boxed{\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})] = \log p_\theta(\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})) + C}$$

From the above, we can say that maximizing ECLL with respect to θ is equivalent to maximizing $\log p_\theta(\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$.

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 $\arg\max_\phi \sum_{i=1}^n \mathcal{L}(\theta, \phi; \boldsymbol{x}_i)$ 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 ?

Here, θ is fixed. Hence, maximizing ELBO $\implies$ minimizing $D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))$ For any instance $\boldsymbol{x}_i$, $q_i^*(\boldsymbol{z})$ would maximize the ELBO or minimize the $D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$. Hence, $D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$ is smaller or equal to $D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$. In best case, q_{ϕ^*} will be same as q_i^*, that's when $D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$ and $D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$ will be equal. So the relation is as below,

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

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) When we have optimal encoding mechanism for both the approaches, the bias for $D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$ is expected to be lower than the bias for $D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$ based on the relation in subquestion 2, the conclusion would also be same as subquestion 2.

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

   (b) from the computational point of view (efficiency):
   In terms of computation, per iteration cost would be the same as we have same number of

parameters. However, required number of iterations will be more when we have instance-specific variational distribution. Ideally, we will have to train n times more parameters for instance specific variational distribution over instance-agnostic distribution(q_ϕ).

(c) in terms of memory (storage of parameters):
Instance specific method would require n times more memory than instance-agnostic method, as we are storing 1 distribution for each example and we have n training examples. So instance-specific variational optimization is expensive in terms of storage requirements as well as optimization.

Question 4 (8-8)**.** Let $p(x, z)$ be the joint probability of a latent variable model where x and z denote the observed and unobserved variables, respectively. Let $q(z|x)$ be an auxiliary distribution which we call the *proposal*, and define [5]

$$\mathcal{L}_K[q(z|x)] = \int \cdots \int \left(q(z_1|x)...q(z_K|x) \log \frac{1}{K} \sum_{j=1}^{K} \frac{p(x, z_j)}{q(z_j|x)} \right) dz_1 dz_2 ... dz_K$$

We've seen in class that this objective is a tighter lower bound on $\log p(x)$ than the evidence lower bound (ELBO), which is equal to $\mathcal{L}_1$; that is $\mathcal{L}_1[q(z|x)] \leq \mathcal{L}_K[q(z|x)] \leq \log p(x)$.

In fact, $\mathcal{L}_K[q(z|x)]$ can be interpreted as the ELBO with a refined proposal distribution. For z_j drawn i.i.d. from $q(z|x)$ with $2 \leq j \leq K$, define the *unnormalized* density

$$\tilde{q}(z|x, z_2, ..., z_K) := \frac{p(x, z)}{\frac{1}{K}\left(\frac{p(x,z)}{q(z|x)} + \sum_{j=2}^{K} \frac{p(x,z_j)}{q(z_j|x)}\right)}$$

(Hint: in what follows, you might need to use the fact that if $w_1, ..., w_K$ are random variables that have the same distribution, then $K\mathbb{E}[w_1] = \sum_i \mathbb{E}[w_i] = \mathbb{E}[\sum_i w_i]$. You need to identify such w_i's before applying this fact for each subquestion.)

1. Show that $\mathcal{L}_K[q(z|x)] = \mathbb{E}_{z_{2:K}}[\mathcal{L}_1[\tilde{q}(z|x, z_2, ..., z_K)]]$; that is, the importance-weighted lower bound with K samples is equal to the average ELBO with the unnormalized density as a refined proposal.

$\mathbb{E}_{z_{2:K}}[\mathcal{L}_1[\tilde{q}(z|x, z_2, ..., z_K)]]$ is,

$$= \int \cdots \int q(z_2|x)...q(z_K|x) \left(\int \tilde{q}(z|x, z_2, ..., z_K) \log \frac{p(x, z_1)}{\tilde{q}(z|x, z_2, ..., z_K)} dz_1 \right) dz_2 dz_3 ... dz_K$$

Substituting $\tilde{q}(z|x, z_2, ..., z_K) := \dfrac{p(x, z)}{\frac{1}{K}\left(\frac{p(x,z)}{q(z|x)} + \sum_{j=2}^{K} \frac{p(x,z_j)}{q(z_j|x)}\right)}$ and multiplying with $q(z_1|x)$,

$$= \int \cdots \int q(z_1|x)q(z_2|x)...q(z_K|x) \frac{\frac{p(x,z_1)}{q(z_1|x)}}{\frac{1}{K}\sum_{j=1}^{K} \frac{p(x,z_j)}{q(z_j|x)}} \log \frac{1}{K} \sum_{j=1}^{K} \frac{p(x, z_j)}{q(z_j|x)} dz_1 dz_2 ... dz_K$$

$$= \mathbb{E}_{z_{1:K}} \left[\frac{\frac{p(x,z_1)}{q(z_1|x)}}{\frac{1}{K}\sum_{j=1}^{K} \frac{p(x,z_j)}{q(z_j|x)}} \log \frac{1}{K} \sum_{j=1}^{K} \frac{p(x, z_j)}{q(z_j|x)} \right]$$

5. Note that $\mathcal{L}_K[\cdot]$ is a "functional" whose input argument is a "function" $q(\cdot|x)$.

Using $K\mathbb{E}[w_1] = \sum_i \mathbb{E}[w_i] = \mathbb{E}[\sum_i w_i]$ in the above equation,

$$= \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_{z_{1:K}}\left[\frac{\frac{p(x,z_i)}{q(z_i|x)}}{\frac{1}{K}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}} \log \frac{1}{K}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}\right] \text{ because } z_i\text{s are IID}$$

Let's take summation inside the expectation,

$$= \mathbb{E}_{z_{1:K}}\left[\frac{\frac{1}{K}\sum_{i=1}^{K}\frac{p(x,z_i)}{q(z_i|x)}}{\frac{1}{K}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}} \log \frac{1}{K}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}\right]$$

$$= \int\cdots\int \left(q(z_1|x)...q(z_K|x) \log \frac{1}{K}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}\right) dz_1 dz_2 ... dz_K$$

$$= \mathcal{L}_K[q(z|x)]$$

$$\implies \boxed{\mathcal{L}_K[q(z|x)] = \mathbb{E}_{z_{2:K}}[\mathcal{L}_1[\tilde{q}(z|x, z_2, ..., z_K)]]}$$

2. Show that $q_K(z|x) := \mathbb{E}_{z_{2:K}}[\tilde{q}(z|x, z_2, ..., z_K)]$ is in fact a probability density function. Also, show that $\mathcal{L}_1[q_K(z|x)]$ is an even tighter lower bound than $\mathcal{L}_K[q(z|x)]$. This implies $q_K(z|x)$ is closer to the true posterior $p(z|x)$ than $q(z|x)$ due to resampling, since $\mathcal{L}_K[q(z|x)] \geq \mathcal{L}_1[q(z|x)]$. (Hint: $f(x) := -x \log x$ is concave.)

$\mathbb{E}_{z_{2:K}}[\tilde{q}(z|x, z_2, ..., z_K)$ is,

$$= \int\cdots\int q(z_2|x)...q(z_K|x)\tilde{q}(z|x, z_2, ..., z_K)\, dz_2 dz_3 ... dz_K$$

$$= \int\cdots\int q(z_2|x)...q(z_K|x)\frac{p(x,z)}{\frac{1}{K}\left(\frac{p(x,z)}{q(z|x)} + \sum_{j=2}^{K}\frac{p(x,z_j)}{q(z_j|x)}\right)} dz_2 dz_3 ... dz_K$$

To prove that $\mathbb{E}_{z_{2:K}}[\tilde{q}(z|x, z_2, ..., z_K)$ is a probability distribution in z, let's integrate it in z space.

$$\int_{z_1} \mathbb{E}_{z_{2:K}}[\tilde{q}(z_1|x, z_2, ..., z_K) dz_1$$

$$= \int_{z_1}\left(\int\cdots\int q(z_2|x)...q(z_K|x)\frac{p(x,z_1)}{\frac{1}{K}\left(\frac{p(x,z_1)}{q(z_1|x)} + \sum_{j=2}^{K}\frac{p(x,z_j)}{q(z_j|x)}\right)} dz_2 dz_3 ... dz_K\right) dz_1$$

$$= \int\cdots\int q(z_1|x)...q(z_K|x)\frac{\frac{p(x,z_1)}{q(z_1|x)}}{\frac{1}{K}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}} dz_1 dz_2 ... dz_K$$

$$= \mathbb{E}_{z_{1:K}}\left[\frac{\frac{p(x,z_1)}{q(z_1|x)}}{\frac{1}{K}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}}\right]$$

Using $K\mathbb{E}[w_1] = \sum_i \mathbb{E}[w_i] = \mathbb{E}[\sum_i w_i]$ in the above equation,

$$= \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_{z_{1:K}}\left[\frac{\frac{p(x,z_i)}{q(z_i|x)}}{\frac{1}{K}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}}\right]$$

Taking summation inside expectation,

$$= \mathbb{E}_{z_{1:K}}\left[\frac{\frac{1}{K}\sum_{i=1}^{K}\frac{p(x,z_i)}{q(z_i|x)}}{\frac{1}{K}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}}\right] = \mathbb{E}_{z_{1:K}}[1] = 1$$

$$\implies \boxed{\int_{z_1}\mathbb{E}_{z_{2:K}}[\tilde{q}(z_1|x,z_2,...,z_K)dz_1 = \int_z q_K(z|x)dz = 1}$$

Hence, we can say that $q_K(z|x)$ is a probability distribution.

Here, $f(x) = -x\log x$ is a concave function, from Jensen's inequality, it implies that $f(\mathbb{E}(x)) \geq \mathbb{E}[f(x)]$. We will use this property. From answer of the previous question, we know that $\mathcal{L}_K[q(z|x)] = \mathbb{E}_{z_{2:K}}[\mathcal{L}_1[\tilde{q}(z|x,z_2,...,z_K)]]$. $\mathcal{L}_K[q(z|x)] =$

$$= \mathbb{E}_{z_{2:K}}[\mathcal{L}_1[\tilde{q}(z|x,z_2,...,z_K)]$$
$$= \mathbb{E}_{z_{2:K}}\left[\int \tilde{q}(z_1|x,z_2,...,z_K)\log\frac{p(x,z_1)}{\tilde{q}(z|x,z_2,...,z_K)}\,dz_1\right]$$
$$= \mathbb{E}_{z_{2:K}}\left[\int \tilde{q}(z_1|x,z_2,...,z_K)\log p(x,z_1)\,dz_1\right] - \mathbb{E}_{z_{2:K}}\left[\int \tilde{q}(z_1|x,z_2,...,z_K)\log\tilde{q}(z_1|x,z_2,...,z_K)\,dz_1\right]$$
$$\leq \int \mathbb{E}_{z_{2:K}}[\tilde{q}(z_1|x,z_2,...,z_K)]\log p(x,z_1)\,dz_1 - \int \mathbb{E}_{z_{2:K}}[\tilde{q}(z_1|x,z_2,...,z_K)]\log\mathbb{E}_{z_{2:K}}[\tilde{q}(z_1|x,z_2,...,z_K)]\,dz_1$$
$$\leq \int \mathbb{E}_{z_{2:K}}[\tilde{q}(z_1|x,z_2,...,z_K)]\frac{\log p(x,z_1)}{\mathbb{E}_{z_{2:K}}[\tilde{q}(z_1|x,z_2,...,z_K)]}\,dz_1$$
$$\leq \int q_K(z|x)\frac{\log p(x,z)}{q_K(z|x)}\,dz = \mathcal{L}_1[q_K(z|x)]$$
$$\implies \boxed{\mathcal{L}_K[q(z|x)] \leq \mathcal{L}_1[q_K(z|x)]}$$

Thus, we can say that $\mathcal{L}_1[q_K(z|x)]$ is a tigher bound than $\mathcal{L}_K[q(z|x)]$.

Question 5 (5-5-5-6)**.** Normalizing flows are expressive invertible transformations of probability distributions. In this exercise, we will see how to satisfy the invertibility constraint of some family of parameterizations. For the first 3 questions, we assume the function $g : \mathbb{R} \to \mathbb{R}$ maps from real space to real space.

1. Let $g(z) = af\,(bz + c)$ where f is the ReLU activation function $f(x) = \max(0, x)$. Show that g is non-invertible.

$$g(z) = \begin{cases} a(bz+c) & \text{if } z > \frac{-c}{b} \\ 0 & \text{otherwise} \end{cases} \qquad \frac{dg(z)}{dz} = \begin{cases} ab & \text{if } z > \frac{-c}{b} \\ 0 & \text{otherwise} \end{cases}$$

   Since $\frac{dg(z)}{dz}$ is 0 for $z \leq \frac{-c}{b}$, $g(z)$ is neither monotinically increasing or decreasing over its domain. Hence $g(z)$ is non-invertible function.

2. Let $g(z) = \sigma^{-1}(\sum_{i=1}^{N} w_i\sigma(a_iz + b_i))$, $0 < w_i < 1$, where $\sum_i w_i = 1$, $a_i > 0$, and $\sigma(x) = 1/(1 + \exp(-x))$ is the logistic sigmoid activation function and σ^{-1} is its inverse. Show that g is *strictly monotonically increasing* on its domain $(-\infty, \infty)$, which implies invertiblity.

   Let $y = h(z) = \sum_{i=1}^{N} w_i\sigma(a_iz + b_i)$, since $\sum_i w_i = 1$ and $\sigma(a_iz + b_i) \in [0, 1]$, for $z \in [-\infty, \infty]$,

$\sum_{i=1}^{N} w_i\sigma(a_iz + b_i) \in [0, 1]$. Calculating σ^{-1},

$$t = \sigma(x) = \frac{1}{1 + exp(-x)}$$
$$1 + exp(-x) = \frac{1}{t}$$
$$exp(-x) = \frac{1-t}{t}$$
$$-x = \log\frac{1-t}{t}$$
$$\boxed{\sigma^{-1}(t) = \log\frac{t}{1-t}}$$

$$\begin{aligned}\frac{d\sigma^{-1}(t)}{dt} &= \frac{d\log\frac{t}{1-t}}{dt} \\ &= \frac{d\log t}{dt} - \frac{d\log(1-t)}{dt} \\ &= \frac{1}{t} + \frac{1}{1-t} \\ \frac{d\sigma^{-1}(t)}{dt} &= \frac{1}{t(1-t)} \text{ (we will use this ahead.)}\end{aligned}$$

$$\begin{aligned}\frac{dg(z)}{dz} &= \frac{d\sigma^{-1}(h(z))}{dz} \\ &= \frac{1}{h(z)(1-h(z))}\frac{dh(z)}{dz} \\ &= \frac{1}{(\sum_{i=1}^{N} w_i\sigma(a_iz+b_i))(1-(\sum_{i=1}^{N} w_i\sigma(a_iz+b_i)))}\frac{d}{dz}(\sum_{i=1}^{N} w_i\sigma(a_iz+b_i)) \\ &= \frac{\sum_{i=1}^{N} w_ia_i\sigma(a_iz+b_i)(1-\sigma(a_iz+b_i))}{\sum_{i=1}^{N} w_i\sigma(a_iz+b_i))(1-(\sum_{i=1}^{N} w_i\sigma(a_iz+b_i)))} \\ \implies &\boxed{\frac{dg(z)}{dz} = \frac{\sum_{i=1}^{N} w_ia_i\sigma(a_iz+b_i)(1-\sigma(a_iz+b_i))}{\sum_{i=1}^{N} w_i\sigma(a_iz+b_i))(1-(\sum_{i=1}^{N} w_i\sigma(a_iz+b_i)))}}\end{aligned}$$

Since $a_i > 0$, we can say that $\frac{dg(z)}{dz} > 0$, which implies that $g(z)$ is monotonically increasing function. Implies that $g(z)$ is invertible !

3. Consider a residual function of the form $g(z) = z + f(z)$. Show that $df/dz > -1$ implies g is

invertible.

$$\frac{df(z)}{dz} > -1$$
$$\frac{df(z)}{dz} + 1 > 0$$
$$g(z) = z + f(z)$$
$$\frac{dg(z)}{dz} = \frac{df(z)}{dz} + 1$$
$$\implies \boxed{\frac{dg(z)}{dz} > 0}$$

Since $\frac{dg(z)}{dz} > 0$, we can say that $g(z)$ is monotonically increasing function, hence it is invertible.

4. Consider the following transformation:

$$g(\boldsymbol{z}) = \boldsymbol{z} + \beta h(\alpha, r)(\boldsymbol{z} - \boldsymbol{z}_0) \tag{2}$$

where $\boldsymbol{z}_0 \in \mathbb{R}^D$, $\alpha \in \mathbb{R}_{>0}$, $\beta \in \mathbb{R}$, and $r = ||\boldsymbol{z} - \boldsymbol{z}_0||_2$, $h(\alpha, r) = 1/(\alpha + r)$. Consider the following decomposition of $\boldsymbol{z} = \boldsymbol{z}_0 + r\tilde{\boldsymbol{z}}$. (i) Given $\boldsymbol{y} = g(\boldsymbol{z})$, show that $\beta \geq -\alpha$ is a sufficient condition to derive the unique r from equation (2). (ii) Given r and $\boldsymbol{y}$, show that equation (2) has a unique solution $\tilde{\boldsymbol{z}}$. Here, we have $(\boldsymbol{z}) = \boldsymbol{z} + \beta h(\alpha, r)(\boldsymbol{z} - \boldsymbol{z}_0)$

$$\begin{aligned} g(\boldsymbol{z}) &= vz + \beta h(\alpha, r)(\boldsymbol{z} - \boldsymbol{z}_0) \\ &= \boldsymbol{z}_0 + r\tilde{\boldsymbol{z}} + \frac{\beta r \tilde{\boldsymbol{z}}}{\alpha + r} \\ g(\boldsymbol{z}) - \boldsymbol{z}_0 &= r(\tilde{\boldsymbol{z}} + \frac{\beta \tilde{\boldsymbol{z}}}{\alpha + r}) \end{aligned}$$

Taking norm on both sides,

$$|g(\boldsymbol{z}) - \boldsymbol{z}_0| = r(1 + \frac{\beta}{\alpha + r}) = f(r)$$

For $f(r)$ to be invertible, it has to be non-decreasing in nature. Meaning $\frac{df(r)}{dr} \geq 0$.

$$\begin{aligned} \frac{df(r)}{dr} &\geq 0 \\ &= 1 + \frac{\beta}{\alpha + r} - \frac{\beta r}{(\alpha + r)^2} \geq 0 \\ &= \frac{(\alpha + r + \beta)(\alpha + r) - r\beta}{(\alpha + r)^2} \geq 0 \\ &= \frac{(\alpha + r)^2 + \alpha\beta + \cancel{r\beta} - \cancel{r\beta}}{(\alpha + r)^2} \geq 0 \implies \frac{\alpha\beta}{(\alpha + r)^2} \geq -1 \implies \boxed{\beta \geq \frac{-(\alpha + r)^2}{\alpha}} \end{aligned}$$

Since $r \geq 0$ it is sufficient to imply that $\boxed{\beta \geq -\alpha}$

Hence, it's a sufficient condition. To get unique solution of $\tilde{\boldsymbol{z}}$

$$\begin{aligned} g(\boldsymbol{z}) &= vz + \beta h(\alpha, r)(\boldsymbol{z} - \boldsymbol{z}_0) \\ &= \boldsymbol{z}_0 + r\tilde{\boldsymbol{z}} + \frac{\beta r \tilde{\boldsymbol{z}}}{\alpha + r} \\ \implies g(\boldsymbol{z}) - \boldsymbol{z}_0 &= r\tilde{\boldsymbol{z}}(1 + \frac{\beta}{\alpha + r}) \\ \implies \boxed{\tilde{\boldsymbol{z}} &= \frac{\boldsymbol{y} - \boldsymbol{z}_0}{r(1 + \frac{\beta}{\alpha + r})}} \end{aligned}$$

The above equation can be uniquely solved give r and $\boldsymbol{y}$.

Question 6 (4-3-6)**.** In this question, we are concerned with analyzing the training dynamics of GANs. Consider the following value function

$$V(d, g) = dg \tag{3}$$

with $g \in \mathbb{R}$ and $d \in \mathbb{R}$. We will use this simple example to study the training dynamics of GANs.

1. Consider gradient descent/ascent with learning rate α as the optimization procedure to iteratively minimize $V(d, g)$ w.r.t. g and maximize $V(d, g)$ w.r.t. d. We will apply the gradient descent/ascent to update g and d simultaneously. What is the update rule of g and d? Write your answer in the following form
$$[d_{k+1}, g_{k+1}]^\top = A[d_k, g_k]^\top$$
where A is a 2×2 matrix; i.e. specify the value of A. The update would look like below,
$$[d_{k+1}, g_{k+1}]^\top = \begin{pmatrix} 1 & \alpha \\ -\alpha & 1 \end{pmatrix} [d_k, g_k]^\top$$

2. The optimization procedure you found in 6.1 characterizes a map which has a stationary point [6], what are the coordinates of the stationary points? For stationary points, we have to set partial derivative of $V(d, g)$ with respect to d and g to 0. $\frac{\partial V(d,g)}{\partial d} = 0 \implies g = 0$ and $\frac{\partial V(d,g)}{\partial g} = 0 \implies d = 0$. Hence the stationary point is $(g, d) = (0, 0)$.

3. Analyze the eigenvalues of A and predict what will happen to d and g as you update them jointly. In other word, predict the behaviour of d_k and g_k as $k \to \infty$.
Eigenvalues of A are $1 + \alpha i$ and $1 - \alpha i$. The norm of eigenvalue is $\sqrt{1 + \alpha^2}$. The norm of the eigenvalue is greater than 1. Which denotes that the vector $[d_k, g_k]^\top$ is scaling up as training progresses. So for this kind of simultaneous training, the variables being optimized are gonna explode. Hence, such a training procedure will lead to divergence and not appropriate for training such minimax objectives.

6. A stationary point is a point on the surface of the graph (of the function) where all its partial derivatives are zero (equivalently, the gradient is zero). Source: https://en.wikipedia.org/wiki/Stationary_point