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.
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.

Your answer should look like Figure 1 (Right).

Answer 1.

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

1.

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**

2.

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

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

3.

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**

4.

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**

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

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

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

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. Show that $\boldsymbol{g}(\boldsymbol{z}_0)$ is distributed by $\mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$ using Equation (1).
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.
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).
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 $|\det \boldsymbol{J}_{\boldsymbol{z}_0}\boldsymbol{g}(\boldsymbol{z}_0)|$?

Answer 2.

Question 1

$$g(z_0) = \mu + \sigma \odot z_0$$

$$\sigma \odot z_0 = g(z_0) - \mu$$

$$z_0 = \mathrm{diag}(\sigma)^{-1}(g(z_0) - \mu)$$

From Equation 1

$$q(g(z_0)) = q(z_0)|det(\frac{\partial g(z_0)}{\partial z_0})|^{-1}$$

$$= \frac{exp(\frac{1}{2}(z_0 - 0)^T I_k^{-1}(z_0 - 0))}{\sqrt{(2\pi)^K |I_k|}}|det(\frac{\partial g(z_0)}{\partial z_0})|^{-1} = \frac{-\frac{1}{2}z_0^T z_0}{\sqrt{(2\pi)^K}}|det(\frac{\partial g(z_0)}{\partial z_0})|^{-1}$$

$$= \frac{exp(-\frac{1}{2}z_0^T z_0)}{\sqrt{(2\pi)^K}}|det(\frac{\partial(\mu + \sigma \odot z_0)}{\partial z_0})|^{-1}$$

$$= \frac{exp(-\frac{1}{2}z_0^T z_0)}{\sqrt{(2\pi)^K}}|det(\mathrm{diag}(\sigma))|^{-1}$$

$$= \frac{exp(-\frac{1}{2}z_0^T z_0)}{\sqrt{(2\pi)^K |det(\mathrm{diag}(\sigma))|^{-1}}}$$

$$= \frac{exp(-\frac{1}{2}(\mathrm{diag}(\sigma)^{-1}(g(z_0) - \mu)^T(\mathrm{diag}(\sigma)^{-1}(g(z_0 - \mu)))}{\sqrt{(2\pi)^K |det(\mathrm{diag}(\sigma^2)|}}$$

$$= \frac{exp(-\frac{1}{2}((g(z_0) - \mu)^T(\mathrm{diag}(\sigma)^2)^{-1}(g(z_0 - \mu)))}{\sqrt{(2\pi)^K |det(\mathrm{diag}(\sigma^2)|}}$$

$$= \mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$$

Question 2

$$|detJ_{z_0}g(z_0)| = |det(\frac{\partial g(z_0)}{\partial z_0})| = |det(\mathrm{diag}(\sigma))|$$

Since Jacobian is diagonal, time complexity should be $O(K)$

[3] https://en.wikipedia.org/wiki/Computational_complexity_of_mathematical_operations

Question 3

$$g(z_0) = \mu + Sz_0 Sz_0 = g(z_0) - \mu z_0 = S^{-1}(g(z_0) - \mu)$$

From Equation 1,

$$q(g(z_0)) = q(z_0)|det(\frac{\partial g(z_0)}{\partial z_0})|^{-1} = \frac{exp(-\frac{1}{2}z_0^T z_0)}{\sqrt{(2\pi)^K}}|det(\frac{\partial g(z_0)}{\partial z_0})|^{-1}$$

$$= \frac{exp(-\frac{1}{2}z_0^T z_0)}{\sqrt{(2\pi)^K}}|det(\frac{\partial(\mu + Sz_0)}{\partial z + 0})|^{-1} = \frac{exp(-\frac{1}{2}z_0^T z_0)}{\sqrt{(2\pi)^K}}|det(S)|^{-1} = \frac{exp(-\frac{1}{2}z_0^T z_0)}{\sqrt{(2\pi)^K|det(SS^T)|}}$$

$$= \frac{exp(-\frac{1}{2}(S^{-1}(g(z_0) - \mu))^T(S^{-1}(g(z_0) - \mu)))}{\sqrt{(2\pi)^K|det(SS^T)|}}$$

$$= \frac{exp(-\frac{1}{2}(g(z_0) - \mu)^T(SS^T)^{-1}(g(z_0) - \mu))}{\sqrt{(2\pi)^K|det(SS^T)|}} = \mathcal{N}(\mu, SS^T)$$

Question 4

$$|det(J_{z_0}g(z_0))| = |det(\frac{\partial g(z_0)}{\partial z_0})| = |det(S)|$$

Since S is a lower triangular matrix, time complexity is $O(K)$

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

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

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

[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.

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 ?
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 3.

Question 1

$$ECLL = E_{q(z|x)}[logp_\theta(x|z)p(z)] = E_{q(z|x)}[logp_\theta(z|x)p_\theta(x)]$$
$$= E_{q(z|x)}[logp_\theta(z|x) + logp_\theta(x)] = E_{q(z|x)}[logp_\theta(z|x)] + E_{q(z|x)}[logp_\theta(x)]$$
$$= E_{q(z|x)}[log(\frac{p_\theta(z|x)}{q(z|x)} \times q(z|x))] + logp_\theta(x)$$
$$= E_{q(z|x)}[log(\frac{p_\theta(z|x)}{q(z|x)}) + log(q(z|x))] + logp_\theta(x)$$
$$= logp_\theta(x) + E_{q(z|x)}[log(\frac{p_\theta(z|x)}{q(z|x)})] + E_{q(z|x)}[log(q(z|x))]$$
$$= logp_\theta(x) - E_{q(z|x)}[log(\frac{q(z|x)}{p_\theta(z|x)})] + E_{q(z|x)}[log(q(z|x))]$$

The second term is KL divergence, then,

$$ECLL = logp_\theta(x) - D_{KL}(q(z|x)||p_\theta(z|x)) + E_{q(z|x)}[log(q(z|x))]$$

The third term is independent of maximization with respect to the model parameter θ, hence it is still equivalent as stated in the question.

Question 2

$$D_{KL}(q_{\phi*}(z|x_i)||p_\theta(z|x_i)) = E_{q_{\phi*}(z|x_i)}[log\frac{q_{\phi*}(z|x_i)}{p_\theta(z|x_i)}] = E_{q_{\phi*}}[log(q_{\phi*}(z|x_i)) - log(p_\theta(z|x_i))]$$
$$= E_{q_{\phi*}}[log(q_{\phi*}(z|x_i)) - log(p_\theta(z|x_i))p_\theta(z) + log(p_\theta(x_i)] = E_{q_{\phi*}}[log(q_{\phi*}(z|x_i)) - log(p_\theta(x_i|z)) - log(p_\theta(z)) + log(p_\theta(z$$
$$= E_{q_{\phi*}}[log(q_{\phi*}(z|x_i)) - log(p_\theta(z))] - E_{q_{\phi*}}[log(p_\theta(x_i|z))] + E_{q_{\phi*}}[log(p_\theta(z))]$$
$$= E_{q_{\phi*}}[\frac{log(q_{\phi*}(z|x_i))}{log(p_\theta(z))}] - E_{q_{\phi*}}[log(p_\theta(x_i|z))] + E_{q_{\phi*}}[log(p_\theta(z))] = D_{KL}(q_{\phi*}(z|x_i)||p_\theta(z) - E_{q_{\phi*}}[log(p_\theta(x_i|z))] + E_{q_{\phi*}}[l$$
$$= -ELBO_{q_{\phi*}} + log(p_\theta(x_i))$$

Similarly,

$$D_{KL}(q_{i*}(z|x_i)||p_\theta(z|x_i)) = -ELBO_{q_{i*}} + log(p_\theta(x_i))$$

Since $ELBO_{q_{i*}} \geq ELBO_{q_{\phi*}}$,

$$D_{KL}(q_{\phi*}(z|x_i)||p_\theta(z|x_i)) \geq D_{KL}(q_{i*}(z|x_i)||p_\theta(z|x_i)) = -ELBO_{q_{i*}} + log(p_\theta(x_i))$$

Question 3

a. Estimation bias is given by KL between two distributions. Given last part, it has less bias than amortized approach (itself being non-amortized)

b. Amortized approach is more efficient. Non-amortized calculates posterior of n points but amortized only does it once

c. From last part, non-amortized approach is n times more spatially expensive.

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.
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.)

Answer 4.

Question 1

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

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

$$= E_{z_{1:k}}[log(\frac{1}{k}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)})]$$

$$= E_{z_{1:k}}[\frac{\sum_{i=1}^{K}\frac{p(x,z_i}{q(z_i|x})}{\sum_{i=1}^{K}\frac{p(x,z_i}{q(z_i|x})}log(\frac{1}{K}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)})]$$

$$= E_{z_{2:k}}[K \times \frac{\frac{p(x,z_i}{q(z_i|x})}{\sum_{i=1}^{K}\frac{p(x,z_j)}{q(z_j|x)})}log(\frac{1}{K}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)})]$$

$$= E_{z_{2:k}}[K\int\frac{\frac{p(x,z_j)}{q(z_j|x)})}{q}(z|x)log(\frac{1}{K}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)})dz]$$

$$= E_{z_{2:k}}[\int\hat{q}(z|x,z_2,...,z_K)log(\frac{1}{K}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)})dz]$$

$$= E_{z_{2:k}}[\int\hat{q}(z|x,z_2,...,z_K)log(\frac{p(x,z_j)}{\frac{p(x,z_j)}{\frac{1}{K}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}}})dz]$$

$$= E_{z_{2:k}}[\int\hat{q}(z|x,z_2,...,z_K)log(\frac{p(x,z_j)}{\frac{p(x,z_j)}{\hat{q}(z|x,z_2,...,z_K)}})dz]$$

$$= E_{z_{2:k}}[\mathcal{L}_1[\hat{q}(z|x,z_2,...,z_K)]]$$

Question 2

$$\int q_K(z|x)dz = \int E_{z_{2:k}}[\mathcal{L}_1[\hat{q}(z|x,z_2,...,z_K)]]dz$$

$$= \int E_{z_{2:k}}\frac{p(x,z)}{q(z,x)}\frac{1}{K}(\frac{p(x,z_j)}{q(z_j|x)} + \sum_{j=2}^{K}\frac{p(x,z_j)}{q(z_j|x)})]dz$$

$$= E_{z_{1:K}}[\frac{p(x,z)}{q(z,x)}\frac{1}{K}/(\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)})]$$

$$= K \times E_{z_{1:K}}[\frac{p(x,z)}{q(z,x)}/(\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)})]$$

$$= \sum_{j=1}^{K}E_{z_{1:K}}[\frac{p(x,z)}{q(z,x)}/(\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)})]$$

$$E_{z_{1:K}}[\frac{\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}}{\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}}]$$

$$= 1$$

Which proves $q_K(z|x)$ is a pdf. Let

$$\hat{p}(z|z1:k) = \frac{1}{K}(\frac{p(x,z)}{q(z|x)} + \sum_{j=2}^{K}\frac{p(x,z)}{q(z|x)})$$

$$\begin{aligned} \mathcal{L}_z[q_K(z|x)] &= E_z[log\frac{p(x,z)}{q(z|x)}) \\ &= E_z[-log(E_{q(z2:k|x}[\hat{p}(z|z1:k)^{-1}]] \\ &= -\int p(x,z)E_{q(z2:k|x}[\hat{p}(z|z1:k)^{-1}]log(E_{q(z2:k|x}[\hat{p}(z|z1:k)^{-1}]dz \\ &\geq -\int p(x,z)E_{q(z2:k|x}[\hat{p}(z|z1:k)^{-1}]log([\hat{p}(z|z1:k)^{-1}]dz \\ &... \\ &= E_q(z1:k)[log(\frac{1}{K}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)})] \end{aligned}$$

Which means $\mathcal{L}_1[q_K(z|x)] \geq \mathcal{L}_K[q(z|x)]$

As in the question ;

$$D_{KL}(q_K(z|x)||p(z|x)) = log(p(x)) - \mathcal{L}_1[q_K(z|x)] \leq log(p(x)) - \mathcal{L}_K[q(z|x)] \leq log(p(x)) - \mathcal{L}_1[q(z|x)] = D_{KL}(q(z|x)$$

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.
2. Let $g(z) = \sigma^{-1}(\sum_{i=1}^{N} w_i\sigma(a_i z + b_i))$, $0 < w_i < 1$, 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.
3. Consider a residual function of the form $g(z) = z + f(z)$. Show that $df/dz > -1$ implies g is invertible.
4. Consider the following transformation:

$$g(\boldsymbol{z}) = \boldsymbol{z} + \beta h(\alpha, r)(\boldsymbol{z} - \boldsymbol{z}_0) \tag{1}$$

where $\boldsymbol{z}_0 \in \mathbb{R}^D$, $\alpha \in \mathbb{R}^+$, $\beta \in \mathbb{R}$, 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 (1). (ii) Given r and $\boldsymbol{y}$, show that equation (1) has a unique solution $\tilde{\boldsymbol{z}}$.

Answer 5.

Question 1

Expand the equation,

$$g(z) = a \times max(0, bz + c) = max(0, a(bz + c))$$

To prove invertibility, one has to show a function is either strictly increasing or decreasing, i.e. derivative must not equal to 0 at any time,

$$\frac{dg(z)}{dz} = max(0, ab)$$

Assuming a and b are non-zero, the derivative could still be 0 from the above expression. This counters with the definition of invertibility. Hence proved.

Question 2 Let $f(z)$ denote $\sum_{i=1}^{N} w_i \sigma(a_i z + b_i)$

$$g(z) = \sigma^{-1}(f(z))$$

Then,

$$\frac{dg(z)}{dz} = \frac{dg(z)}{df(z)} \frac{df(z)}{z}$$

$$= \frac{1}{f(z)[1 - f(z)]} \sum_{i=1}^{N} \omega_i a_i \sigma(a_i z + b_i))[1 - \sigma(a_i z + b_i))]$$

Sigmoid function has limits at infinities:

$$\lim_{x \to +\infty} \sigma(x) = 1, \lim_{x \to +\infty} 1 - \sigma(x) = 0$$

$$\lim_{x \to -\infty} \sigma(x) = 0, \lim_{x \to -\infty} 1 - \sigma(x) = 1$$

Which means that in the domain of $(-\infty, +\infty$, the range is $(0, 1)$. Back to the eqaution. The first part, $\frac{1}{f(z)[1-f(z)]}$, can be expanded to,

$$\frac{1}{f(z)[1 - f(z)]} = [\sum_{i=1}^{N} w_i \sigma(a_i z + b_i)(1 - \sum_{i=1}^{N} w_i \sigma(a_i z + b_i))]^{-1}$$

Given the weights sum to 1 and that the range for a sigmoid function is $(0, 1)$, for non-zero weights, each element's value is also between 0 and 1, non-inclusive. Also, since sum of all weights equals to 1, even if all sigmoid activation are large (close to 1), the upper-bond is still smaller than 1. The lower-bound is larger than 0. The two multiplier would have the same upper/lower-bounds, which is $(0, 1)$. Therefore, the product is strictly positive, and as a denominator, the expression evaluates to positive as well.

The second part can be grouped as,

$$\sum_{i=1}^{N} (\omega_i a_i) \sigma(a_i z + b_i))[1 - \sigma(a_i z + b_i))]$$

Where from given, $\omega_i a_i > 0$. The two multipliers have same upper and lower bounds as analyzed from above, which means the three multipliers give a product that is strictly positive.

Multiplying the two parts which are both strictly positive, the product is also strictly positive. Since this is the derivative of $g(z)$, the original function would be monotonically strictly increasing as its derivative is strictly positive within the domain $(-\infty, +\infty)$. Therefore it is invertible

Question 3

We can get an expression for derivative of $g(z)$ by:

$$\frac{dg(z)}{dz} = \frac{d(z+f(z))}{dz} = 1 + \frac{df}{dz}$$

Invertibility is satisfied IFF the function is strictly increasing or decreasing. So the derivative cannot be if we wanted it to be invertible.

$$\frac{dg(z)}{dz} \neq 0$$
$$1 + \frac{df}{dz} \neq 0$$
$$\frac{df}{dz} \neq -1$$

We can then either have $\frac{df}{dz} > -1$ or $\frac{df}{dz} < -1$. Hence statement proved

Question 4

a)

$$y = g(z) = z + \beta h(\alpha, r)(z - z_0)$$
$$y - z_0 = z - z_0 + \frac{\beta}{\alpha + r}(z - z_0)$$
$$||y - z_0||_2 = ||z - z_0||_2 + \frac{\beta}{\alpha + r}||(z - z_0)||_2 = ||(z - z_0)||_2(1 + \frac{\beta}{\alpha + r}) = r(1 + \frac{\beta}{\alpha + r})$$

RHS needs to have a positive derivative,

$$\frac{d}{dr}(r(1 + \frac{\beta}{\alpha + r})) \geq 0$$
$$\frac{d}{dr}(r + \beta r(\alpha + r)^{-1}) \geq 0$$
$$1 + \beta((\alpha + r)^{-1} - r(\alpha + r)^2) \geq 0$$
$$\beta \geq -\frac{(\alpha + r)^2}{\alpha}$$

$r \geq 0$

$$\beta \geq -\alpha$$

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

$$V(d, g) = dg \tag{2}$$

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.
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 ?
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 $k \to \infty$.

Answer 6.

Question 1 Take partials,

$$\frac{\partial V}{\partial d} = g_k$$
$$\frac{\partial V}{\partial g} = d_k$$
$$d_{k+1} = d_k + \alpha g_k$$
$$g_{k+1} = g_k - \alpha d_k$$

Written in matrix,

$$\begin{bmatrix} d_{k+1} \\ g_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & \alpha \\ -\alpha & 1 \end{bmatrix} \begin{bmatrix} d_k \\ g_k \end{bmatrix}$$

Where A $= \begin{bmatrix} 1 & \alpha \\ -\alpha & 1 \end{bmatrix}$

Question 2

At stationary point, derivative is 0.

$$(d*, g*) = (0, 0)$$

Question 3

$$|A - \lambda I| = 0$$
$$\begin{vmatrix} 1 & \alpha \\ -\alpha & 1 \end{vmatrix} - \lambda \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 0$$
$$(1-)^2 + \alpha^2 = 0$$
$$\lambda = 1 \pm i\alpha$$

A has complex eigenvalues. Real and imaginary parts are both positive, d_k and g_k will converge to stationary point as $k \longrightarrow \infty$ and will spiral

[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