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.

The pixels in the image after the first convolutional layer that contribute to the pixel at index 34 are pixels 33, 43, 44, 45 for mask A and pixels 33, 34, 43, 44, 45 for mask B.

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

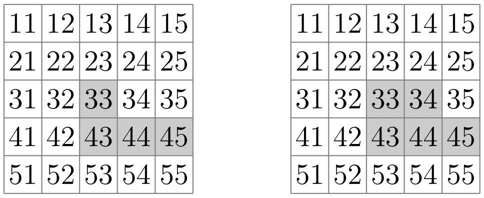

FIGURE 2 – Receptive field of pixel 34 for mask A (left) and mask B (right)

Thus, only the pixels respective to the mask need to be considered to compute the receptive field of each mask in the second convolutional layer.

The repective field of these pixels in the first convolutional layer are :

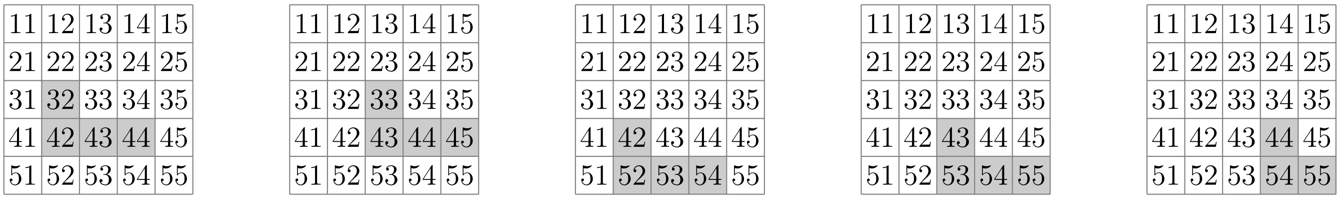

FIGURE 3 – Receptive field of pixel 33, 34, 43, 44, 45 for mask A

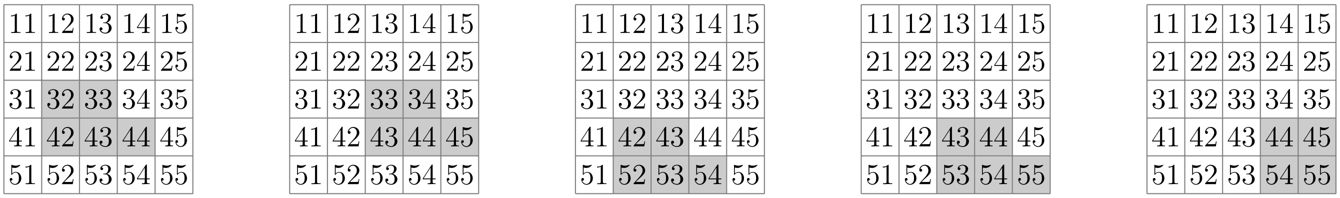

FIGURE 4 – Receptive field of pixel 33, 34, 43, 44, 45 for mask B

1. If we use $\mathbf{M}^A$ for the first layer and $\mathbf{M}^A$ for the second layer we combine the receptive fields of pixels 33, 43, 44, and 45 from Figure 3.

2. If we use $\mathbf{M}^A$ for the first layer and $\mathbf{M}^B$ for the second layer we combine the receptive fields of pixels 33, 34, 43, 44, and 45 from Figure 3.

The same logic applies for the other subquestions. Thus, the answers are :

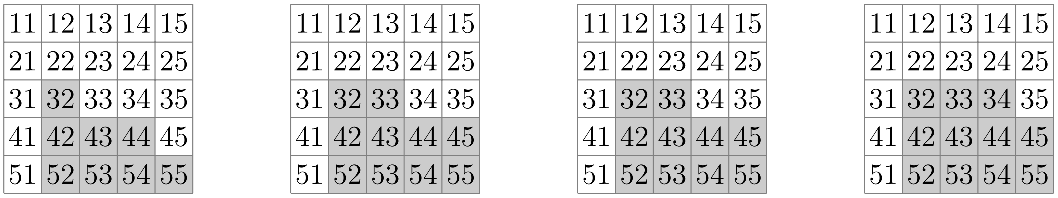

FIGURE 5 – Receptive fields for answer 1, 2, 3, and 4

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

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

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 $\boldsymbol{g}$:

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

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 $\mathcal{N}(\mu, \text{diag}(\sigma^2))$ 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.

1. $\boldsymbol{g}(\boldsymbol{z}_0) = \mu + \sigma \odot \boldsymbol{z}_0$

$$\implies \left|\det\left(\tfrac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0}\right)\right|^{-1} = \left|\det\left(\tfrac{\partial}{\partial \boldsymbol{z}_0}(\mu + \sigma \odot \boldsymbol{z}_0)\right)\right|^{-1} = \left|\det\left(\text{diag}(\sigma)\right)\right|^{-1} = \tfrac{1}{\sigma^K}$$

Now we can find the form of $\boldsymbol{g}(z_o)$;

$$\begin{aligned} q(z_o) &= N(0, I_K) \\ &= \frac{1}{\sqrt{(2\pi)^K}} \left|I_K\right|^{-\frac{1}{2}} \exp(-\frac{1}{2} z_0^T z_0) \end{aligned}$$

For every element of z_0 ;

$$\begin{aligned} z_0 &= \frac{\boldsymbol{g}(z_0) - \mu}{\sigma} \\ &= \text{diag}(\sigma)^{-1}(\boldsymbol{g}(z_0) - \mu) \end{aligned}$$

Also,

$$z_0^T z_0 = (\boldsymbol{g}(z_0) - \mu)^T \, \text{diag}(\sigma^2)^{-1}(\boldsymbol{g}(z_0) - \mu)$$

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

Combining everything we get ;

$$
\begin{aligned}
q(g(z_0)) &= \frac{1}{\sqrt{(2\pi)^K}} \left|I_K\right|^{-\frac{1}{2}} \cdot \frac{1}{\sigma^K} \cdot \exp(-\frac{1}{2}(g(z_0)-\mu)^T \operatorname{diag}(\sigma^2)^{-1}(g(z_0)-\mu)) \\
&= \frac{1}{\sqrt{(2\pi)^K}} \left|\operatorname{diag}(\sigma^2)\right|^{-\frac{1}{2}} \cdot \exp(-\frac{1}{2}(g(z_0)-\mu)^T \operatorname{diag}(\sigma^2)^{-1}(g(z_0)-\mu)) \\
&= N(\mu, \operatorname{diag}(\sigma^2))
\end{aligned}
$$

2. As shown in the previous question, the jacobian matrix is a K x K diagonal matrix of σ and its determinant was equal to σ^K, i.e., evaluating the trace of the jacobian. Therefore, its time complexity is $O(K)$.

3. $\boldsymbol{g}(z_0) = \mu + \boldsymbol{S}z_0$
$\implies \left|\det\left(\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0}\right)\right|^{-1} = \left|\det\left(\frac{\partial}{\partial \boldsymbol{z}_0}(\mu + \boldsymbol{S}z_0)\right)\right|^{-1} = |S|^{-1} = \frac{1}{\det(S)}$

Now we can find the form of $g(z_0)$;

$$q(z_o) = \frac{1}{\sqrt{(2\pi)^K}} \left|I_K\right|^{-\frac{1}{2}} \exp(-\frac{1}{2} z_0^T z_0)$$

For every element of z_0 ;

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

$$z_0^T z_0 = (g(z_0)-\mu)^T (SS^T)^{-1}(g(z_0)-\mu)$$

Combining everything we get ;

$$
\begin{aligned}
q(g(z_0)) &= \frac{1}{\sqrt{(2\pi)^K}} \left|I_K\right|^{-\frac{1}{2}} \cdot \frac{1}{\det(S)} \cdot \exp(-\frac{1}{2}(g(z_0)-\mu)^T (SS^T)^{-1}(g(z_0)-\mu)) \\
&= \frac{1}{\sqrt{(2\pi)^K}} \left|SS^T\right|^{-\frac{1}{2}} \cdot \exp(-\frac{1}{2}(g(z_0)-\mu)^T (SS^T)^{-1}(g(z_0)-\mu)) \\
&= N(\mu, SS^T)
\end{aligned}
$$

4. As shown in the previous question, the jacobian matrix is a K x K matrix of σ, which was $\det(S)$. Given that $\boldsymbol{S}$ is a lower triangular matrix, its determinant is the multiplication of its diagonal elements. Therefore, the time complexity of the determinant 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

[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.

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 $p(\boldsymbol{z}|\boldsymbol{x})$.
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.

1. Since $q(z|x)$ and $p(z)$ do not depend on θ, we can utilize this in the following way (also note that as the problem state, the derivation holds if $p(z|x)$ perfectly matches $q(z|x)$) ;

$$\begin{aligned}
&\arg\max_\theta\{\mathbb{E}_{q(z|x)}[\log p_\theta(x|z)p(z)]\}\\
&= \arg\max_\theta\left\{\mathbb{E}_{q(z|x)}\left[\log \frac{p_\theta(x|z)p(z)}{p_\theta(z|x)}\frac{p_\theta(z|x)}{q(z|x)}\right]\right\}\\
&= \arg\max_\theta\left\{\mathbb{E}_{q(z|x)}\left[\log \frac{p_\theta(x|z)p(z)}{q(z|x)}\right]\right\} \qquad \text{Because } [\frac{p_\theta(z|x)}{p_\theta(z|x)}]\\
&= \arg\max_\theta\left\{\mathbb{E}_{q(z|x)}\left[\log p_\theta(x|z)p(z) - \log q(z|x)\right]\right\}\\
&= \arg\max_\theta\left\{\mathbb{E}_{q(z|x)}\left[\log p_\theta(x,z) - \log q(z|x)\right]\right\} \qquad \text{Because } [p(x,z) = p(x|z)p(z)]\\
&= \arg\max_\theta\left\{\mathbb{E}_{q(z|x)}\left[\log q(z|x)p(x) - \log q(z|x)\right]\right\} \quad [p(x,z) = p(z|x)p(x) \text{ and } p(z|x) = q(z|x)]\\
&= \arg\max_\theta\{\mathbb{E}_{q(z|x)}[\log p_\theta(x)] - \mathbb{E}_{q(z|x)}[\log q(z|x) - \log p_\theta(z|x)]\}\\
&= \arg\max_\theta\left\{\log p_\theta(x) - \mathbb{E}_{q(z|x)}\left[\log \frac{q(z|x)}{p_\theta(z|x)}\right]\right\}\\
&= \arg\max_\theta\{\log p_\theta(x) - D_{KL}(q(z|x)||p_\theta(z|x))\}
\end{aligned}$$

2. As seen in the lecture notes on VAE, the inference gap of q_ϕ^* can be written as ;

$$\log p_\theta(x_i) - \mathcal{L}[q_{\phi^*}(z|x_i) = \log p_\theta(x_i) - \mathcal{L}[q_i^*(z)] + \mathcal{L}[q_i^*(z)] - \mathcal{L}[q_{\phi^*}(z|x_i)]$$

Using the equality seen at page 37 of the lecture notes on VAE, we can write the following ;

$$\log p_\theta(x) - D_{KL}(q(z|x)||p_\theta(z|x)) = \mathcal{L}[q(z|x)]$$
$$\log p_\theta(x) - \mathcal{L}[q(z|x)] = D_{KL}(q(z|x)||p_\theta(z|x))$$

Replacing this equality in the inference gap formula. we get ;

$$D_{KL}(q_{\phi^*}(z|x_i)||p_\theta(z|x_i)) = D_{KL}(q_i^*(z)||p_\theta(z|x_i)) + (\mathcal{L}[q_i^*(z)] - \mathcal{L}[q_{\phi^*}(z|x_i)])$$

Because q_i^* is the maximizer amongst Q, $(\mathcal{L}[q_i^*(z)] - \mathcal{L}[q_{\phi^*}(z|x_i)])$ is positive.

Thus, $D_{KL}(q_{\phi^*}(z|x_i)||p_\theta(z|x_i)) \geq D_{KL}(q_i^*(z)||p_\theta(z|x_i))$

3. a) There is a bias due to the KL divergence because the expected data log likelihood is the only term in the ELBO that depends on θ.

3. b) In terms of computational cost, approximating $q_\phi^*(z|x_i)$ is more efficient since we don't have to update q_i for each data point x_i.

3. c) In terms of memory, approximating $q_i^*(z)$ needs more memory in order to store the variational parameters for each training example individually, meaning that there are n times as many parameters than in the amortized case.

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.)

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

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.

1. First, let's define an equivalence to the unnormalized density

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

We can see that it is equivalent by the following ;

$$\begin{aligned}
\tilde{q}(z|x, z_2, ..., z_K) &= \frac{\frac{p(x,z)}{q(z|x)}}{\frac{1}{k}\sum_{j=1}^{K} \frac{p(x,z_j)}{q(z_j|x)}} q(z|x) \\
&= \left[\frac{p(x,z)}{q(z|x)} \cdot k \sum_{j=1}^{K} \frac{q(z_j|x)}{p(x,z_j)}\right] q(z|x) \\
&= p(x,z) \cdot \left[k \sum_{j=1}^{K} \frac{q(z_j|x)}{p(x,z_j)}\right] \\
&= p(x,z) \cdot k \sum_{j=1}^{K} \frac{q(z_j|x)}{p(x,z_j)} \\
&= p(x,z) \div \frac{1}{k}\left(\frac{p(x,z)}{q(z|x)} + \sum_{j=2}^{K} \frac{p(x,z)}{q(z_j|x)}\right) \\
&= \frac{p(x,z)}{\frac{1}{k}\left(\frac{p(x,z)}{q(z|x)} + \sum_{j=2}^{K} \frac{p(x,z)}{q(z_j|x)}\right)}
\end{aligned}$$

Using this we get ;

$$
\begin{aligned}
&\mathbb{E}_{z_2:z_k}[\mathcal{L}_1[\tilde{q}(z|x, z_2, ..., z_k)]] \\
&= \mathbb{E}_{z_2:z_k}\left[\int_z \tilde{q}(z|x, z_{2:k}) \log\left(\frac{p(x,z)}{\tilde{q}(z|x, z_{2:k})}\right) dz\right] \\
&= \mathbb{E}_{z_2:z_k}\left[\int_z \tilde{q}(z|x, z_{2:k}) \log\left(\frac{p(x,z)}{\frac{\frac{p(x,z)}{q(z|x)}}{\frac{1}{k}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}}} q(z|x)\right) dz\right] \\
&= \mathbb{E}_{z_2:z_k}\left[\int_z \tilde{q}(z|x, z_{2:k}) \log\left(\frac{1}{k}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}\right) dz\right] \\
&= \mathbb{E}_{z_2:z_k}\left[\int_z k\frac{\frac{p(x,z)}{q(z|x)}}{\frac{1}{k}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}} q(z|x) \log\left(\frac{1}{k}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}\right) dz\right] \\
&= \mathbb{E}_{z_2:z_k}\left[\int_{z_1} k\frac{\frac{p(x,z_1)}{q(z_1|x)}}{\frac{1}{k}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}} q(z|x) \log\left(\frac{1}{k}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}\right) dz\right] \\
&= \mathbb{E}_{z_2:z_k}\left[k\frac{\frac{p(x,z_1)}{q(z_1|x)}}{\frac{1}{k}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}} q(z|x) \log\left(\frac{1}{k}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}\right)\right]
\end{aligned}
$$

Since z_i has the same expectation as z_1, we can write ;

$$
\begin{aligned}
&= \mathbb{E}_{z_2:z_k}\left[\frac{\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}}{\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}} \log\left(\frac{1}{k}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}\right)\right] \\
&= \mathbb{E}_{z_2:z_k}\left[\log\left(\frac{1}{k}\sum_{j=1}^{K}\frac{p(x,z_j)}{q(z_j|x)}\right)\right] \\
&= \mathcal{L}_K[q(z|x)]
\end{aligned}
$$

2. To show that $q_k(z|x)$ is a probability density function ;

$$
\begin{aligned}
\int_z q_k(z|x)dz &= \int_z \mathbb{E}_{z_{2:k}}[\tilde{q}(z|x, z_2, ..., z_k)]dz \\
&= \int_z \frac{q(z|x)}{q(z|x)} \mathbb{E}_{z_{2:k}}\left[\frac{p(x,z)}{\frac{1}{K}\left(\frac{p(x,z)}{q(z|x)} + \sum_{j=2}^{K} \frac{p(x,z_j)}{q(z_j|x)}\right)}\right] dz \\
&= \mathbb{E}_z \mathbb{E}_{z_{2:k}}\left[\frac{\frac{p(x,z)}{q(z|x)}}{\frac{1}{k}\left(\frac{p(x,z)}{q(x,z)} + \sum_{j=2}^{K} \frac{p(x,z_j)}{q(z_j|x)}\right)}\right] \\
&= \mathbb{E}_{z_1,...,z_k}\left[\frac{\frac{p(x,z_1)}{q(z_1|x)}}{\frac{1}{k}\left(\sum_{j=1}^{K} \frac{p(x,z_j)}{q(z_j|x)}\right)}\right] \qquad \text{(change of notation } z = z_1\text{)} \\
&= k \cdot \mathbb{E}_{z_1,...,z_k}\left[\frac{\frac{p(x,z_1)}{q(z_1|x)}}{\left(\sum_{j=1}^{K} \frac{p(x,z_j)}{q(z_j|x)}\right)}\right]
\end{aligned}
$$

Since z_i has the same expectation as z_1, we can replace k with the sum of k terms

$$
\begin{aligned}
&= \sum_{i=1}^{K} \mathbb{E}_{z_1,...,z_k}\left[\frac{\frac{p(x,z_1)}{q(z_1|x)}}{\left(\sum_{j=1}^{K} \frac{p(x,z_j)}{q(z_j|x)}\right)}\right] \\
&= \mathbb{E}_{z_1,...,z_k}\left[\frac{\sum_{i=1}^{K} \frac{p(x,z_1)}{q(z_1|x)}}{\left(\sum_{j=1}^{K} \frac{p(x,z_j)}{q(z_j|x)}\right)}\right] \qquad \text{(Linearity of expectation)} \\
&= \mathbb{E}_{z_1:z_k} 1 \\
&= 1
\end{aligned}
$$

To show that $\mathcal{L}_1[q_k(z|x)]$ is an even tighter lower bound than $\mathcal{L}_k[q(z|x)]$;

Remember from the previous question that $q_k(z|x) = \mathbb{E}_{z_{2:k}}[\tilde{q}(z|x, z_2, ..., z_k)]$ and from the definition $\tilde{q}(z|x, z_2 : k) = \frac{p(x,z)}{\frac{1}{K}\left(\frac{p(x,z)}{q(z|x)} + \sum_{j=2}^{K} \frac{p(x,z_j)}{q(z_j|x)}\right)}$

We have ;

$$
\begin{aligned}
L_1[q_k(z|x)] &= \mathbb{E}_z[\log\left(\frac{p(x,z)}{q_k(z|x)}\right)] \\
&= \mathbb{E}_z\left[\log\left(\frac{p(x,z)}{\mathbb{E}_{z_{2:k}}\left[\frac{p(x,z)}{\frac{1}{K}\left(\frac{p(x,z)}{q(z|x)}+\sum_{j=2}^{K}\frac{p(x,z_j)}{q(z_j|x)}\right)}\right]}\right)\right] \\
&= \mathbb{E}_z\left[\log\left(\frac{1}{\mathbb{E}_{z_{2:k}}\left[\frac{1}{\frac{1}{K}\left(\frac{p(x,z)}{q(z|x)}+\sum_{j=2}^{K}\frac{p(x,z_j)}{q(z_j|x)}\right)}\right]}\right)\right] \\
&= \mathbb{E}_z\left[-\log\left(\mathbb{E}_{z_{2:k}}\left[\frac{1}{\frac{1}{K}\left(\frac{p(x,z)}{q(z|x)}+\sum_{j=2}^{K}\frac{p(x,z_j)}{q(z_j|x)}\right)}\right]\right)\right] \\
&= -\mathbb{E}_z\left[\log\left(\mathbb{E}_{z_{2:k}}\left[\frac{1}{K}\left(\frac{p(x,z)}{q(z|x)}+\sum_{j=2}^{K}\frac{p(x,z_j)}{q(z_j|x)}\right)\right]\right)\right] \\
&\text{Let } \hat{p}(x|z) = \frac{1}{K}\left(\frac{p(x,z)}{q(z|x)}+\sum_{j=2}^{K}\frac{p(x,z_j)}{q(z_j|x)}\right) \\
\\
&= -\mathbb{E}_z[\log(\mathbb{E}_{z_{2:k}}[\hat{p}(x|z)^{-1}])] \\
&= -\int_z p(x,z)\mathbb{E}_{z_{2:k}}[\hat{p}(x|z)^{-1}]\log(\mathbb{E}_{z_{2:k}}[\hat{p}(x|z)^{-1}])dz
\end{aligned}
$$

Given the hint that $f(x) = -x\log x$ is concave for $x > 0$ and using Jensen's inequality for concave transformations, i.e, $f(\mathbb{E}[x]) \geq \mathbb{E}[f(x)]$, then $f(\mathbb{E}[x]) = -\mathbb{E}[x]\log\mathbb{E}[x] \geq \mathbb{E}[-x\log x]$

We can write the following ;

$$
\begin{aligned}
L_1[q_k(z|x)] &\geq -\int_z p(x,z)\mathbb{E}_{z_2:k}[\hat{p}(x|z)^{-1}\log(\hat{p}(x|z)^{-1})]dz \\
&= -\int_z p(x,z)\int_{z_{2:k}} q_k(z|x)\hat{p}(x|z)^{-1}\log(\hat{p}(x|z)^{-1})dz \\
&= -\int_{z_{1:k}} \frac{q(z_1|x)}{q(z_1|x)} p(x,z_1)q_k(z|x)\hat{p}(x|z)^{-1}\log(\hat{p}(x|z)^{-1})dz \\
&= -\int_{z_{1:k}} \frac{p(x|z_1)}{q(z_1|x)} q(z_1|x)\hat{p}(x|z)^{-1}\log(\hat{p}(x|z)^{-1})dz \\
&= \int_{z_{1:k}} \frac{p(x,z_1)}{\frac{q(z_1|x)}{\hat{p}(x|z)}} q(z_1|x)\log(\hat{p}(x|z))dz
\end{aligned}
$$

Substituting $\hat{p}(x|z)$, we get ;

$$
= k\int_{z_{1:k}} \frac{\frac{p(x,z_1)}{q(z_1|x)}}{\sum_{j=1}^K \frac{p(x,z_j)}{q(z_j|x)}} q(z|x)\log\left(\frac{1}{K}\left(\sum_{j=1}^K \frac{p(x,z_j)}{q(z_j|x)}\right)\right)dz
$$

Since z_i has the same expectation as z_1, we can replace k with the sum of k terms ;

$$
\begin{aligned}
&= \sum_{i=1}^k \int_{z_{1:k}} \frac{\frac{p(x,z_i)}{q(z_i|x)}}{\sum_{j=1}^K \frac{p(x,z_j)}{q(z_j|x)}} q(z|x)\log\left(\frac{1}{K}\left(\sum_{j=1}^K \frac{p(x,z_j)}{q(z_j|x)}\right)\right)dz \\
&= \int_{z_{1:k}} \frac{\sum_{i=1}^k \frac{p(x,z_i)}{q(z_i|x)}}{\sum_{j=1}^K \frac{p(x,z_j)}{q(z_j|x)}} q(z|x)\log\left(\frac{1}{K}\left(\sum_{j=1}^K \frac{p(x,z_j)}{q(z_j|x)}\right)\right)dz \\
&= \int_{z_{1:k}} q(z|x)\log\left(\frac{1}{K}\left(\sum_{j=1}^K \frac{p(x,z_j)}{q(z_j|x)}\right)\right)dz \\
&= \mathbb{E}_{z_{1:k}}\left[\log\left(\frac{1}{K}\left(\sum_{j=1}^K \frac{p(x,z_j)}{q(z_j|x)}\right)\right)\right] \\
&= L_k[q(z|x)]
\end{aligned}
$$

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.

1. Let's define

$$g(y) = af(y) \to y = af(y) \to g^{-1}(y) = \frac{1}{a} f^{-1}(y) \to f^{-1}(y) = \max(0, y)'$$

We can see that $f(y) = \max(0, y)'$ is non-invertible with the following proof by example ;

$$f(-5) = 0 \qquad f^{-1}(0) = \beta \quad , \beta \in [-\infty, 0]$$

Therefore, $f^{-1}(y)$ is surjective, hence $g(z)$ is non-invertible.

2. In order to show that a function is strictly monotonically increasing on its domain, we have to show that $f'(x) > 0 \quad \forall x \in (a, b)$.

Let's define σ^{-1} ;

$$\begin{aligned}\sigma^{-1}(x) &= \log\left(\frac{x}{1-x}\right) \\ &= \log(x) - \log(1-x)\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial z}\sigma^{-1}(x(z)) &= \frac{\partial}{\partial z}\log(x(z)) - \log(1 - x(z)) \\ &= \frac{1}{x(z)}\frac{\partial}{\partial z}x(z) - \frac{1}{(1 - x(z))}\frac{\partial}{\partial z}(1 - x(z)) \\ &= \frac{1}{x(z)}\frac{\partial}{\partial z}x(z) + \frac{1}{(1 - x(z))}\frac{\partial}{\partial z}x(z)\end{aligned}$$

Keeping in mind that $x(z) = \sum_{i=1}^{N} w_i \sigma(a_i z + b_i)$, we have ;

$$\begin{aligned}\frac{\partial}{\partial z}x(z) &= \sum_{i=1}^{N} w_i \sigma(a_i z + b_i) \\ &= \sum_{i=1}^{N} w_i \frac{\partial}{\partial z}\sigma(a_i z + b_i)\end{aligned}$$

Knowing that

$$\sigma(a_i z + b_i) = \frac{1}{1 + e^{-(a_i z + b_i)}}$$

We can do the following ;

$$\begin{aligned}
\frac{\partial}{\partial z}\sigma(a_i z + b_i) &= \frac{\partial}{\partial z}(1+e^{-(a_i z+b_i)})^{-1} \\
&= -(1+e^{-(a_i z+b_i)})^{-2}\frac{\partial}{\partial z}(1+e^{-(a_i z+b_i)}) \\
&= -(1+e^{-(a_i z+b_i)})^{-2}e^{-(a_i z+b_i)} \cdot -a_i \\
&= (1+e^{-(a_i z+b_i)})^{-2}e^{-(a_i z+b_i)}a_i \\
&= \frac{e^{-(a_i z+b_i)}a_i}{(1+e^{-(a_i z+b_i)})^2} \\
&= \frac{e^{-(a_i z+b_i)}}{1+e^{-(a_i z+b_i)}} \cdot \frac{a_i}{1+e^{-(a_i z+b_i)}} \\
&= \frac{e^{-(a_i z+b_i)}+1-1}{1+e^{-(a_i z+b_i)}} \cdot \frac{a_i}{1+e^{-(a_i z+b_i)}} \\
&= \left(\frac{1+e^{-(a_i z+b_i)}}{1+e^{-(a_i z+b_i)}} - \frac{1}{1+e^{-(a_i z+b_i)}}\right)\frac{a_i}{1+e^{-(a_i z+b_i)}} \\
&= \frac{a_i}{1+e^{-(a_i z+b_i)}} \cdot \left(1 - \frac{1}{1+e^{-(a_i z+b_i)}}\right) \\
&= a_i\sigma(a_i z + b_i)(1-\sigma(a_i z+b_i))
\end{aligned}$$

Coming back to what we had and substituting ;

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

Given the following ;

$$0 < \sigma(z) < 1 \quad \text{(definition of the sigmoid)}, \quad 0 < w_i < 1, \quad a_i > 0$$

We can say that $0 < x(z) < 1$ and $\frac{\partial}{\partial z}x(z) > 0$. Therefore, we previously had ;

$$\frac{\partial}{\partial z}\sigma^{-1}(x(z)) = \frac{1}{x(z)}\frac{\partial}{\partial z}x(z) + \frac{1}{(1-x(z))}\frac{\partial}{\partial z}x(z)$$

Where every term > 0. Thus, $\frac{\partial}{\partial z}\sigma^{-1}(x(z)) > 0 \quad \forall z \in (-\infty, \infty)$, which implies invertibility.

3. As stated in the previous question, a strictly monotonically increasing function on its domain implies invertibility. Also, for a function to be strictly monotonically increasing, its derivative needs to be positive over its domain. Hence, to be strictly monotonically increasing, thus invertible ;

$$\frac{d}{dz}g(z) = \frac{d}{dz}(z + f(z)) > 0$$

$$1 + \frac{d}{dz}f(z) > 0 \Rightarrow \frac{d}{dz}f(z) > -1$$

4.1) Let $g(z) = z + \beta h(\alpha, r)(z - z_0)$

$$\begin{aligned}
y &= z + \frac{\beta}{\alpha + r}(z - z_0) \\
y - z_0 &= (z - z_0) + \frac{\beta}{\alpha + r}(z - z_0) \\
y &= (z - z_0) + \frac{\beta}{\alpha + r}(z - z_0) + z_0 \\
&= \frac{(z - z_0)(\alpha + r)}{\alpha + r} + \frac{(z - z_0)\beta}{\alpha + r} + z_0 \\
&= \frac{(z - z_0)(\alpha + r) + (z - z_0)\beta}{\alpha + r} + z_0 \\
&= \left(\frac{\alpha + r + \beta}{\alpha + r}\right)(z - z_0) + z_0
\end{aligned}$$

Which is equivalent to ;

$$||y - z_0||_2^2 = \left(\frac{\alpha + r + \beta}{\alpha + r}\right)^2 r^2$$

With $\beta + \alpha \geq 0$, we only need to make sure that $\frac{\alpha + r + \beta}{\alpha + r} > 0$ and we can remove the square. Thus getting ;

$$\begin{aligned}
||y - z_0||_2 &= \left(\frac{\alpha + r + \beta}{\alpha + r}\right) r \\
(\alpha + r)||y - z_0||_2 &= (\alpha + r + \beta) r \\
\alpha||y - z_0||_2 + r||y - z_0||_2 &= \alpha r + r^2 + \beta r \\
r^2 + (\alpha + \beta - ||y - z_0||_2) r - \alpha||y - z_0||_2 &= 0
\end{aligned}$$

Renaming to $f(r)$;

$$f(r) = r^2 + (\alpha + \beta - ||y - z_0||_2) r - \alpha||y - z_0||_2 = 0$$

With the quadratic equation ; $\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$, we get ;

$$r = \frac{-(\alpha + \beta - ||y - z_0||_2) \pm \sqrt{\delta}}{2}$$

Where $\delta = (\alpha + \beta - ||y - z_0||_2)^2 + 4\alpha||y - z_0||_2$ and is > 0. Since $r = ||z - z_0||_2$, we know that $r > 0$ and we can see that r has an unique solution.

4.2)

$$\begin{aligned}
g(z) &= z + \beta h(\alpha, r)(z - z_0) \\
y &= z + \beta h(\alpha, r)(z - z_0) \\
&\text{Substituting } z = z_0 + r\tilde{z} \\
y &= z_0 + r\tilde{z} + \beta h(\alpha, r)(z_0 + r\tilde{z} - z_0) \\
y - z_0 &= r\tilde{z} + \beta h(\alpha, r)(r\tilde{z}) \\
y - z_0 &= r\tilde{z}(1 + \beta h(\alpha, r)) \\
\frac{y - z_0}{1 + \beta h(\alpha, r)} &= r\tilde{z} \\
\tilde{z} &= \frac{y - z_0}{r(1 + \beta h(\alpha, r))} \\
&\text{Substituting } h(\alpha, r) = \frac{1}{\alpha + r} \\
\tilde{z} &= \frac{y - z_0}{r(1 + \frac{\beta}{\alpha + r})}
\end{aligned}$$

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.

1. Doing gradient ascent on $V(d, g)$ w.r.t d (discriminator) has the following form ;

$$d_{k+1} \leftarrow d_k + \alpha \frac{\partial}{\partial d}(V(d_k, g_k)) = d_k + \alpha g_k$$

Doing gradient descent on $V(d, g)$ w.r.t g (generator) has the following form ;

$$g_{k+1} \leftarrow g_k - \alpha \frac{\partial}{\partial g}(V(d_k, g_k)) = g_k - \alpha d_k$$

[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

Rewriting our answer in the asked form : $[d_{k+1}, g_{k+1}]^T$ gives us ;

$$\boldsymbol{A}[d_k, g_k]^T$$

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

2. Stationary points are points on the surface of the graph where all partial derivatives are zero (equivalently, the gradient is zero). Thus,

$$\begin{aligned} \nabla V(d,g) &= \begin{bmatrix} \frac{\partial}{\partial d}(V(d,g)) \\ \frac{\partial}{\partial g}(V(d,g)) \end{bmatrix} \\ &= \begin{bmatrix} g \\ d \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow g = 0, \quad d = 0 \end{aligned}$$

Thus, the stationary point is at (0,0).

3. To find the eigenvalues of A ;

$$\det(\boldsymbol{A} - \lambda I) = 0$$

Thus,

$$\begin{aligned} &\det\left(\begin{bmatrix} 1 & \alpha \\ -\alpha & 1 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) = 0 \\ &\det\left(\begin{bmatrix} 1-\lambda & \alpha \\ -\alpha & 1-\lambda \end{bmatrix}\right) = 0 \\ &(1-\lambda)^2 + \alpha^2 = 0 \\ &(1-\lambda)^2 = -\alpha^2 \\ &\sqrt{(\lambda-1)^2} = \sqrt{-\alpha^2} \qquad \text{Because } [(1-\lambda)^2 = (\lambda-1)^2] \\ &\text{Let } i = \sqrt{-1} \\ &\lambda - 1 \pm i\alpha \\ &\lambda = 1 \pm i\alpha \end{aligned}$$

Therefore, $\lambda_1 = 1 + i\alpha$ and $\lambda_2 = 1 - i\alpha$.

Since eigenvalues with complex numbers imply rotation, the update rule is such that the updates will be rotating in the parameter space and will never converge to an equilibrium as $k \to \infty$.