Due Date: April 29th 23:59, 2020

Student:
Gustavo Alonso Patino

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 5 (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.

   Solution

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

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

FIGURE 2 – (Left) $\boldsymbol{M}^A$ first layer. (Right) $\boldsymbol{M}^A$ second layer

2. If we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^B$ for the second layer.
   Solution

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

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 3 – (Left) $\boldsymbol{M}^A$ first layer. (Right) $\boldsymbol{M}^B$ second layer

3. If we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^A$ for the second layer.
Solution

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 4 – (Left) $\boldsymbol{M}^B$ first layer. (Right) $\boldsymbol{M}^A$ second layer

4. If we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^B$ for the second layer.

Solution

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 5 – (Left) $\boldsymbol{M}^B$ first layer. (Right) $\boldsymbol{M}^B$ second layer

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 $q(\boldsymbol{z}_0) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_K)$ 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).

Solution

$$\frac{\partial \boldsymbol{g}}{\partial z_0} = \frac{\partial}{\partial z_0}(\mu + \sigma \odot \boldsymbol{z}_0) = \text{diag}(\sigma) \tag{2}$$

so

$$\left|\det\left(\frac{\partial \boldsymbol{g}}{\partial z_0}\right)\right|^{-1} = \frac{1}{|\text{diag}(\sigma)|} \tag{3}$$

we know that $q(\boldsymbol{z}_0) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_K) = \frac{1}{\sqrt{(2\pi)^k}} \exp\left(\frac{-1}{2} z_0^T z_0\right)$ but $\boldsymbol{g}(z_0) = \mu + \sigma \odot \boldsymbol{z}_0 \rightarrow z_0 = \frac{\boldsymbol{g}(z_0) - \mu}{\sigma}$ (using elementwise division) replacing in the definition of normal distribution of $q(z_0)$ we arrive to

$$q(z_0) = \frac{1}{\sqrt{(2\pi)^k}} \exp\left(\frac{-1}{2}\left(\frac{\boldsymbol{g}(z_0) - \mu}{\sigma}\right)^T \left(\frac{\boldsymbol{g}(z_0) - \mu}{\sigma}\right)\right) \tag{4}$$

$$= \frac{1}{\sqrt{(2\pi)^k}} \exp\left(\frac{-1}{2}\left(\boldsymbol{g}(z_0) - \mu\right)^T \left(\text{diag}(\sigma^2)\right)^{-1} \left(\boldsymbol{g}(z_0) - \mu\right)\right) \tag{5}$$

using Eq.3 and replacing on the Eq.1 we achieve

$$q(g(z_0)) = \frac{1}{\sqrt{(2\pi)^k |\text{diag}(\sigma^2)|}} \exp\left(\frac{-1}{2}\left(\boldsymbol{g}(z_0) - \mu\right)^T \left(\text{diag}(\sigma^2)\right)^{-1} \left(\boldsymbol{g}(z_0) - \mu\right)\right) \tag{6}$$

$$= \mathcal{N}(\mu, \text{diag}(\sigma^2)) \tag{7}$$

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.

Solution
$|\det \boldsymbol{J}_{\boldsymbol{z}_0} \boldsymbol{g}(\boldsymbol{z}_0)|$ when $\boldsymbol{g}(\boldsymbol{z}_0) = \mu + \sigma \odot \boldsymbol{z}_0$ is the determinant of a diagonal matrix which has a time complexity of $\mathcal{O}(K)$

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).

Solution

$$\frac{\partial \boldsymbol{g}}{\partial z_0} = \frac{\partial}{\partial z_0}(\mu + \boldsymbol{S} z_0) = \boldsymbol{S} \tag{8}$$

so

$$\left|\det\left(\frac{\partial \boldsymbol{g}}{\partial z_0}\right)\right|^{-1} = \frac{1}{|\boldsymbol{S}|} \tag{9}$$

we proceed in a similar way to the exercise in item 1: we know that $q(\boldsymbol{z}_0) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_K) = \frac{1}{\sqrt{(2\pi)^k}} \exp\left(\frac{-1}{2} z_0^T z_0\right)$ but $\boldsymbol{g}(z_0) = \mu + \boldsymbol{S}\boldsymbol{z}_0 \rightarrow z_0 = \boldsymbol{S}^{-1}(\boldsymbol{g}(z_0) - \mu)$, replacing in the definition of

$q(z_0)$ we arrive to the expression

$$q(z_0) = \frac{1}{\sqrt{(2\pi)^k}} \exp\left(\frac{-1}{2}\left(\boldsymbol{S}^{-1}(g(z_0)-\mu)\right)^T\left(\boldsymbol{S}^{-1}(g(z_0)-\mu)\right)\right) \tag{10}$$

$$\frac{1}{\sqrt{(2\pi)^k}} \exp\left(\frac{-1}{2}\left(g(z_0)-\mu\right)^T\left(\boldsymbol{S}\boldsymbol{S}^T\right)^{-1}\left(g(z_0)-\mu\right)\right) \tag{11}$$

and using the definition for $q(g(z_0))$ in Eq.1 we obtain

$$q(g(z_0)) = \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{S}|}} \exp\left(\frac{-1}{2}\left(g(z_0)-\mu\right)^T\left(\boldsymbol{S}\boldsymbol{S}^T\right)^{-1}\left(g(z_0)-\mu\right)\right) \tag{12}$$

$$= \mathcal{N}(\mu, \boldsymbol{S}\boldsymbol{S}^T) \tag{13}$$

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 $|\det \boldsymbol{J}_{\boldsymbol{z}_0}\boldsymbol{g}(\boldsymbol{z}_0)|$?

Solution
For a trinagular matrix we have again a time complexity of $\mathcal{O}(K)$

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

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

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

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.

Solution

$$\begin{aligned}
\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})] &= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x},\boldsymbol{z})] && (14)\\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{z},\boldsymbol{x})] && (15)\\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})] && (16)\\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{z}|\boldsymbol{x}) - \log q(\boldsymbol{z}|\boldsymbol{x})] + \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\log q(\boldsymbol{z}|\boldsymbol{x}) + \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\log p(\boldsymbol{x}) && (17)\\
&= -\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\log\frac{q(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} + \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\log p(\boldsymbol{x}) + \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\log q(\boldsymbol{z}|\boldsymbol{x}) && (18)\\
&= \log p(x) - D_{KL}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})) + \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\log q(\boldsymbol{z}|\boldsymbol{x}) && (19)
\end{aligned}$$

If we do a maximization in relation to the parameters θ we can consider the last term in Eq.19 as a residual constant term and we can guarantee that

$$\arg\max_\theta\{\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})]\} = \arg\max_\theta\{\log p(x) - D_{KL}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))\} \quad (20)$$

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 ?

Solution Consider the maximal ELBO for an instance $\boldsymbol{x}_i$: $\max\ \mathcal{L}(\theta,\phi,\boldsymbol{x}_i)$, and the maximum value of the sum of the ELBO for all the instances: $\max\ \sum_{i=1}^n \mathcal{L}(\theta,\phi,\boldsymbol{x}_i)$, clearly we can guarantee that

$$\sum_{i=1}^n \max\mathcal{L}(\theta,\phi,\boldsymbol{x}_i) \geq \max\sum_{i=1}^n \mathcal{L}(\theta,\phi,\boldsymbol{x}_i) \quad (21)$$

By definition

$$\mathcal{L}(\theta,\phi;\boldsymbol{x}) = \mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x}\mid\boldsymbol{z})] - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}\mid\boldsymbol{x})||p(\boldsymbol{z}))$$

therefore

$$\begin{aligned}
&\sum_{i=1}^n \arg\max_\phi \left\{\mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x}\mid\boldsymbol{z})] - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}\mid\boldsymbol{x})||p(\boldsymbol{z}))\right\} \geq && (22)\\
&\arg\max_\phi \left\{\sum_{i=1}^n \mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x}\mid\boldsymbol{z})] - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}\mid\boldsymbol{x})||p(\boldsymbol{z}))\right\} && (23)
\end{aligned}$$

using the definitions of ϕ^* and q^* in the last equation we find

$$\begin{aligned}
&\mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x}\mid\boldsymbol{z})] - D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)) \geq && (24)\\
&\mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x}\mid\boldsymbol{z})] - D_{\mathrm{KL}}(q_\phi^*(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)) && (25)
\end{aligned}$$

so we conclude that

$$D_{\mathrm{KL}}(q_i^*(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)) \leq D_{\mathrm{KL}}(q_\phi^*(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)) \quad (26)$$

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)

Solution The bias in estimating the marginal likelihood using ELBO is just the KL divergence, so the same result shown above holds, i.e

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

(b) from the computational point of view (efficiency)
Solution
Per iteration there is no difference between the number of calculations to be made for both algorithms. However q^* may require more times for each iteration to reach the optimum. So q^* is less computationally efficient than q_ϕ^*

(c) in terms of memory (storage of parameters)
Solution
For q_i^* we must bound each data point $\boldsymbol{x}_i$ with a particular ELBO $\mathcal{L}_i$ so the storage is linear in the number of samples n of the training set, i.e $\mathcal{O}(n)$, so this operation should be more expensive than q_ϕ^*.

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.

Solution

$$\mathbb{E}_{z_{2:K}}[\mathcal{L}_1[\tilde{q}(z|x, z_2 : z_K)]] = \mathbb{E}_{z_{2:K}}\left[\int_z \tilde{q}(z|x, z_2 : z_K) \log \left(\frac{p(x, z)}{\tilde{q}(z|x, z_2 : z_K)} \right) dz \right] \tag{27}$$

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

Using the equation

$$\tilde{q}(z|x, z_2 : z_K) = \frac{p(x,z)}{\frac{1}{k}\sum_{i=1}^{k}\frac{p(x,z_i)}{q(z_i|x)}} \tag{28}$$

in Eq.27 we obtain

$$\mathbb{E}_{z_{2:K}}[\mathcal{L}_1[\tilde{q}(z|x, z_2 : z_K)]] = \mathbb{E}_{z_{2:K}}\left[\int_z \tilde{q}(z|x, z_2 : z_K)\log\left(\frac{p(x,z)}{\frac{1}{k}\sum_{i=1}^{k}\frac{p(x,z_i)}{q(z_i|x)}}\right)dz\right] \tag{29}$$

$$\mathbb{E}_{z_{2:K}}\left[\int_z \tilde{q}(z|x, z_2 : z_K)\log\left(\frac{1}{k}\sum_{i=1}^{k}\frac{p(x,z_i)}{q(z_i|x)}\right)\right] \tag{30}$$

Using again Eq.28 in Eq.30 and multiplying by the factor $\frac{q(z|x)}{q(z|x)}$ we obtain

$$\mathbb{E}_{z_{2:K}}[\mathcal{L}_1[\tilde{q}(z|x, z_2 : z_K)]] = \mathbb{E}_{z_{2:K}}\left[\int_z k\frac{\frac{p(x,z)}{q(z|x)}}{\sum_{i=1}^{k}\frac{p(x,z_i)}{q(z_i|x)}}q(z|x)\log\left(\frac{1}{k}\sum_{i=1}^{k}\frac{p(x,z_i)}{q(z_i|x)}\right)dz\right] \tag{31}$$

Considering that

$$\mathbb{E}_{z_{2:K}}\left[\int_z k\frac{\frac{p(x,z)}{q(z|x)}}{\sum_{i=1}^{k}\frac{p(x,z_i)}{q(z_i|x)}}q(z|x)\log\left(\frac{1}{k}\sum_{i=1}^{k}\frac{p(x,z_i)}{q(z_i|x)}\right)dz\right] = \mathbb{E}_{z_{1:K}}\left[k\frac{\frac{p(x,z_1)}{q(z_1|x)}}{\sum_{i=1}^{k}\frac{p(x,z_i)}{q(z_i|x)}}\log\left(\frac{1}{k}\sum_{i=1}^{k}\frac{p(x,z_i)}{q(z_i|x)}\right)\right] \tag{32}$$

and using the property $k\frac{p(x,z_1)}{q(z_1|x)} = \sum_{i=1}^{k}\frac{p(x,z_i)}{q(z_i|x)}$ in Eq.32 we conclude that

$$\mathbb{E}_{z_{2:K}}[\mathcal{L}_1[\tilde{q}(z|x, z_2 : z_K)]] = \mathbb{E}_{z_{1:K}}\left[\log\left(\frac{1}{k}\sum_{i=1}^{k}\frac{p(x,z_i)}{q(z_i|x)}\right)\right] = \tag{33}$$

$$= \mathcal{L}_k(q(z|x)) \tag{34}$$

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.)

Solution

Showing 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, implies to demonstrate that it is a normalized function, therefore using the Eq.28 and multiplying by the factor $\frac{q(z|x)}{q(z|x)}$ we can say that

$$\int_z q_K(z|x)\,dz = \int_z \mathbb{E}_{z_{2:K}}[\tilde{q}(z|x, z_2, ..., z_K)]\,dz \tag{35}$$

$$= \int_z \mathbb{E}_{z_{2:K}}\left[\frac{p(x,z)}{\frac{1}{k}\sum_{i=1}^{k}\frac{p(x,z_i)}{q(z_i|x)}}\,dz\right] \tag{36}$$

$$= \int_z q(z|x)\mathbb{E}_{z_{2:K}}\left[\frac{\frac{p(x,z)}{q(z|x)}}{\frac{1}{k}\sum_{i=1}^{k}\frac{p(x,z_i)}{q(z_i|x)}}\,dz\right] \tag{37}$$

Considering again that

$$\int_z q(z|x)\mathbb{E}_{z_{2:K}}\left[\frac{\frac{p(x,z)}{q(z|x)}}{\frac{1}{k}\sum_{i=1}^k \frac{p(x,z_i)}{q(z_i|x)}}\right] dz = \mathbb{E}_{z_{1:K}}\left[\frac{\frac{p(x,z_1)}{q(z_1|x)}}{\frac{1}{k}\sum_{i=1}^k \frac{p(x,z_i)}{q(z_i|x)}}\right] \tag{38}$$

and using the property $k\frac{p(x,z_1)}{q(z_1|x)} = \sum_{i=1}^k \frac{p(x,z_i)}{q(z_i|x)}$ in Eq.38 we arrive to

$$\int_z q_K(z|x)\, dz = \mathbb{E}_{z_{1:K}}\left[\frac{\sum_{i=1}^k \frac{p(x,z_i)}{q(z_i|x)}}{\sum_{i=1}^k \frac{p(x,z_i)}{q(z_i|x)}}\right] \tag{39}$$

$$= \mathbb{E}_{z_{1:K}}[1] \tag{40}$$

$$= 1 \tag{41}$$

Now we will show it is a tighter bound, it means we will show that $\mathcal{L}_K[q(z|x)] \geq \mathcal{L}_1[q(z|x)]$.

$$\mathcal{L}_K[q(z|x)] = \mathbb{E}_{z\sim q(z|x)}\left[\log \frac{p(x,z)}{q(z|x)}\right] \tag{42}$$

but

$$q_K(z|x) := \mathbb{E}_{q(z_{2:K}|x)} \frac{p(x,z)}{\hat{p}(x|z_{1:k})}$$

so replacing in Eq.42

$$\mathcal{L}_K[q(z|x)] = \mathbb{E}_{z\sim q(z|x)}\left[\log\left(\frac{p(x,z)}{\mathbb{E}_{q(z_{2:K}|x)}\frac{p(x,z)}{\hat{p}(x|z_{1:k})}}\right)\right] \tag{43}$$

$$= \mathbb{E}_{z\sim q(z|x)}\left[\log\left(\frac{1}{\mathbb{E}_{q(z_{2:K}|x)}\frac{1}{\hat{p}(x|z_{1:k})}}\right)\right] \tag{44}$$

$$= \mathbb{E}_{z\sim q(z|x)}\left[-\log(\mathbb{E}_{q(z_{2:K}|x)}[\hat{p}(x|z_{1:k})^{-1}])\right] \tag{45}$$

Remembering the definition of expectation $\mathbb{E}_{z\sim q(z|x)}$, and the property $f(\mathbb{E}[x]) = -\mathbb{E}[x]\log\mathbb{E}[x] \geq \mathbb{E}[-x\log x]$ we find the integral form:

$$\mathcal{L}_K[q(z|x)] = -\int_z p(x,z)\mathbb{E}_{q(z_{2:K}|x)}[\hat{p}(x|z_{1:k})^{-1}]\log(\mathbb{E}_{q(z_{2:K}|x)}[\hat{p}(x|z_{1:k})^{-1}])\, dz \tag{46}$$

$$\geq -\int_z p(x,z)\mathbb{E}_{q(z_{2:K}|x)}[\hat{p}(x|z_{1:k})^{-1}]\log(\hat{p}(x|z_{1:k})^{-1})\, dz \tag{47}$$

using again the definition of expectation for $\mathbb{E}_{q(z_{2:K}|x)}$ we transform the last equation as

$$\mathcal{L}_K[q(z|x)] \geq -\int_z p(x,z)\int_{z2:k} q(z_{2:K}|x)\hat{p}(x|z_{1:k})^{-1}\log(\hat{p}(x|z_{1:k})^{-1})\, dz \tag{48}$$

Considering that $\int z(), \int z_{2:k}()$ can be contained in an integral of the form $\int z_{1:k}$ and multiplying by the factor $\frac{q(z_1|x)}{q(z_1|x)}$ we arrive to

$$\mathcal{L}_K[q(z|x)] \geq -\int_{z_{1:k}} \frac{q(z_1|x)}{q(z_1|x)} p(x,z_1)q(z_{2:K}|x)\hat{p}(x|z_{1:k})^{-1}\log(\hat{p}(x|z_{1:k})^{-1})\, dz \tag{49}$$

but $q(z_1|x)q(z_{2:K}|x) = q(z_{1:K}|x)$ therefore we have

$$\mathcal{L}_K[q(z|x)] \geq -\int_{z_{1:k}} \frac{p(x,z_1)}{q(z_1|x)} q(z_{1:K}|x)\hat{p}(x|z_{1:k})^{-1}\log(\hat{p}(x|z_{1:k})^{-1})\,dz \tag{50}$$

$$\geq \int_{z_{1:k}} \frac{\frac{p(x,z_1)}{q(z_1|x)}}{\hat{p}(x|z_{1:k})} q(z_{1:k}|x)\log\hat{p}(x|z_{1:k})\,dz \tag{51}$$

using the property $k\frac{p(x,z_1)}{q(z_1|x)} = \sum_{i=1}^k \frac{p(x,z_i)}{q(z_i|x)}$ in the equation above we obtain

$$\mathcal{L}_K[q(z|x)] \geq k\int_{z_{1:k}} \frac{\frac{p(x,z_1)}{q(z_1|x)}}{\sum_{i=1}^k \frac{p(x,z_i)}{q(z_i|x)}} q(z_{1:k}|x)\log\left(\frac{1}{k}\sum_{i=1}^k \frac{p(x,z_i)}{q(z_i|x)}\right)\,dz \tag{52}$$

$$\geq \sum_{i=1}^k \int_{z_{1:k}} \frac{\frac{p(x,z_i)}{q(z_i|x)}}{\sum_{i=1}^k \frac{p(x,z_i)}{q(z_i|x)}} q(z_{1:k}|x)\log\left(\frac{1}{k}\sum_{i=1}^k \frac{p(x,z_i)}{q(z_i|x)}\right)\,dz \tag{53}$$

$$\geq \int_{z_{1:k}} q(z_{1:k}|x)\log\left(\frac{1}{k}\sum_{i=1}^k \frac{p(x,z_i)}{q(z_i|x)}\right)\,dz \tag{54}$$

$$\geq \mathbb{E}_{q(z_{1:k})}\left[\log\left(\frac{1}{k}\sum_{i=1}^k \frac{p(x,z_i)}{q(z_i|x)}\right)\right] \tag{55}$$

$$\geq \mathcal{L}_1[q(z|x)] \tag{56}$$

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.

   Solution
   $g(z) = af(bz+c)$ so $g(z) = a\,\max\{0, bz+c\}$. We can see here that $g(z) = 0$ for every $bz + c \leq 0$, i.e a value of $g(z)$ can be mapped to multiple values in the input set, making the function non invertible

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.

   Solution
   Considering that $\sigma(x) = \frac{1}{1+\exp(-x)}$ we know that

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

   Assuming $f(z) \equiv \sum_{i=1}^n w_i\sigma(a_iz + b_i)$ we can write

   $$g(z) = \log\left(\frac{f(z)}{1-f(z)}\right) \tag{57}$$

$$g'(z) = \frac{1-f(z)}{f(z)} \frac{d}{dz}\left(\frac{f(z)}{1-f(z)}\right) \tag{58}$$

$$= \frac{1-f(z)}{f(z)} \frac{f'(z)}{(1-f(z))^2} \tag{59}$$

but $f'(z) = \sum_{i=1}^{n} w_i \sigma'(a_i z + b_i) a_i$, $\sigma'(a_i z + b_i)$ is positive in all the domain, $a_i > 0$ and $0 < w_i < 1$ so we conclude that $f'(z) > 0$ and consequently $\frac{f'(z)}{(1-f(z))^2} > 0$. On the other hand, $\sigma(a_i z + b_i)$ is bound between 0 and 1 and $\sum_{i=1}^{n} w_i = 1$ so we conclude that $f(z) = \sum_{i=1}^{n} w_i \sigma(a_i z + b_i)$ has the boundaries $0 \leq f(z) \leq 1$, so from Eq.59 we conclude that $g'(z) \geq 0$ and the function is strictly monotonic

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

Solution

$$g(z) = z + f(z) \rightarrow \frac{dg}{dz} = 1 + \frac{df}{dz} \tag{60}$$

if $\frac{df}{dz} > -1$ consequently $\frac{dg}{dz} > 0$ and g is invertible.

4. Consider the following transformation:

$$g(\boldsymbol{z}) = \boldsymbol{z} + \beta h(\alpha, r)(\boldsymbol{z} - \boldsymbol{z}_0) \tag{61}$$

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 (61). (ii) Given r and $\boldsymbol{y}$, show that equation (61) has a unique solution $\tilde{\boldsymbol{z}}$.

Solution

— Using the transformation $\boldsymbol{z} = \boldsymbol{z}_0 + r\tilde{\boldsymbol{z}}$ we have

$$g(\boldsymbol{z}) = \boldsymbol{z} + \beta h(\alpha, r)(\boldsymbol{z} - \boldsymbol{z}_0) \tag{62}$$

$$= \boldsymbol{z}_0 + r\tilde{\boldsymbol{z}} + \beta h(\alpha, r) r\tilde{\boldsymbol{z}} \tag{63}$$

therefore, considering that $\tilde{\boldsymbol{z}} = \frac{\boldsymbol{z} - \boldsymbol{z}_0}{r} \rightarrow |\tilde{\boldsymbol{z}}| = 1$ we have

$$|g(\boldsymbol{z}) - \boldsymbol{z}_0| = r + \frac{r\beta}{\alpha + r} \rightarrow \tag{64}$$

$$\frac{\partial}{\partial r}|g(\boldsymbol{z}) - \boldsymbol{z}_0| = \frac{1 + \beta\alpha}{(\alpha + r)^2} \tag{65}$$

A sufficient condition to derive the unique r should be

$$\frac{1 + \beta\alpha}{(\alpha + r)^2} \geq 0 \rightarrow \tag{66}$$

$$\beta \geq -\frac{(\alpha + r)^2}{\alpha} \tag{67}$$

but $r \geq 0$ because $r = ||\boldsymbol{z} - \boldsymbol{z}_0||_2$ so we conclude that a sufficient condition is $\beta \geq -\alpha$

— Using Eq.63 we can obtain the solution for $\tilde{\boldsymbol{z}}$ as

$$\tilde{\boldsymbol{z}} = \frac{g(\boldsymbol{z}) - \boldsymbol{z}_0}{r\left(1 + \frac{\beta}{\alpha + r}\right)} \tag{68}$$

so it has a unique solution.

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

$$V(d, g) = dg \tag{69}$$

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.

Solution
For minimize V w.r.t g and maximize w.r.t d we can use the gradient descent/ascent algorithm

$$d^{k+1} = d^k + \eta \frac{\partial V}{\partial d} \tag{70}$$

$$g^{k+1} = g^k - \eta \frac{\partial V}{\partial g} \tag{71}$$

with η the learning rate. In matricial form we can write this equations as

$$\begin{pmatrix} d^{k+1} \\ g^{k+1} \end{pmatrix} = \begin{pmatrix} 1 & \eta \\ -\eta & 1 \end{pmatrix} \begin{pmatrix} d^k \\ g^k \end{pmatrix} \tag{72}$$

therefore we can define

$$\mathbf{A} = \begin{pmatrix} 1 & \eta \\ -\eta & 1 \end{pmatrix} \tag{73}$$

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?

Solution
The stationary points are defined by $\frac{\partial V}{\partial d} = 0$ and $\frac{\partial V}{\partial g} = 0$ so the coordinates of the stationary points are $g = 0$, $d = 0$

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

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 $k \to \infty$.

Solution
The eigenvalues of **A** are

$$(1-\lambda)^2 + \eta^2 = 0 \to \tag{74}$$
$$\lambda = 1 \pm i\eta \tag{75}$$

we can see that $|\lambda| > 1$ so the system should not converge. In order to see the dynamic of the system we could consider that the simultaneous gradient "game" can be understood as applying the Euler-method to the ordinary differential equation

$$\frac{d}{dt}[d(t), g(t)]^T = \left[\frac{\partial}{\partial d}L_d, \frac{\partial}{\partial g}L_g\right]^T \tag{76}$$

where L_d and L_g are the loss functions of the players. For the player "d" $L_d = dg$ and for the player "g" the loss is $L_g = -dg$ replacing on Eq.76 we obtain

$$[\dot{d}(t), \dot{g}(t)]^T = [g, -d]^T \tag{77}$$

so we have the differential equation

$$\ddot{d} + d = 0$$

and the solution of the system is

$$[d(t), g(t)]^T \propto [cos(t), sin(t)]^T \tag{78}$$

i.e the solution moves in a circle around the stationary point $(0, 0)$