1 Reversible Architectures [3pts]: In this section, we will investigate a variant for implementing reversible block with affine coupling layers. Consider the following reversible affline coupling block:

$$\begin{aligned} \boldsymbol{y}_1 &= \exp(\mathcal{G}(\boldsymbol{x}_2)) \circ \boldsymbol{x}_1 + \mathcal{F}(\boldsymbol{x}_2) \\ \boldsymbol{y}_2 &= \exp(\boldsymbol{s}) \circ \boldsymbol{x}_2 \end{aligned} \tag{1}$$

where $\circ$ denotes element-wise multiplication. The each inputs $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^{\frac{d}{2}}$. The functions $\mathcal{F}$ and $\mathcal{G}$ maps from $\mathbb{R}^{\frac{d}{2}} \to \mathbb{R}^{\frac{d}{2}}$. This modified block is identical to the ordinary reversible block, except that the inputs $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are multiplied element-wise by vectors $\exp(\mathcal{F}(\boldsymbol{x}_2))$ and $\exp(\boldsymbol{s})$.

1. **(1pt)** Give the equations for inverting this block, i.e. computing $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ from $\boldsymbol{y}_2$ and $\boldsymbol{y}_2$. You may use / to denote element-wise division.

$$x_2 = y_2 / \exp(s) \qquad \left(x_{2i} = \frac{y_{2i}}{\exp(s_i)} \right)$$

$$\begin{aligned} x_1 &= [y_1 - F(x_2)] / \exp(\mathcal{G}(x_2)) \\ &= [y_1 - F(y_2/\exp(s))] / \exp[\mathcal{G}(y_2/\exp(s))] \end{aligned}$$

2. **(1pt)** Give a formula for the Jacobian $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$, where $\boldsymbol{y}$ denotes the concatenation of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. You may denote the solution as a block matrix, as long as you clearly define what the matrix for each block corresponds to.

$$\frac{\partial y_1}{\partial x_1} = \exp(\mathcal{G}(x_2)), \quad \frac{\partial y_1}{\partial x_2} = F'(x_2) + \mathcal{G}'(x_2) \circ \exp(\mathcal{G}(x_2)) \circ x_1$$

$$\frac{\partial y_2}{\partial x_1} = 0, \quad \frac{\partial y_2}{\partial x_2} = \exp(s)$$

$$\text{So} \quad J = \left[\begin{array}{c|c} \exp(\mathcal{G}(x_2)) & F'(x_2) + \mathcal{G}'(x_2) \circ \exp(\mathcal{G}(x_2)) \circ x_1 \\ \hline 0 & \exp(s) \end{array}\right] \in \mathbb{R}^{D \times D}$$

Where all blocks are diagonal $D/2 \times D/2$ matrices with the stated vectors as their diagonals.

Because of the element-wise nature,

$$\text{all } \frac{\partial y_{ki}}{\partial x_{nj}} = 0 \text{ for } i \neq j$$

3. **(1pt)** Give a formula for the determinant of the Jacobian from previous part, i.e. compute $det\left(\frac{\partial y}{\partial x}\right)$. Is this a volume preserving transformation? Justify your answer.

J is upper triangular, so $|J| = \prod_{i=1}^{D} J_{ii}$

$$\Rightarrow \quad |J| = \prod_{i=1}^{D/2} \exp(g(x_{2,i}) + s_i)$$

$$= \exp\left(\sum_{i=1}^{D/2} g(x_{2,i}) + s_i\right) \neq 1$$

unless the sum is 0, which we can assume is not the case.

So since $|J| \neq 1$, this transformation is <u>not</u> volume preserving!

2 Variational Free Energy [6pts]: In this question you will derive some expressions related to variational free energy which is maximized to train a VAE. Recall that the VFE is defined as:

$$\mathcal{F}(q) = \mathbb{E}_q[\log p(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}(q(\boldsymbol{z})||p(\boldsymbol{z}))$$

where KL divergence is defined as

$$D_{KL}(q(\boldsymbol{z})||p(\boldsymbol{z})) = \mathbb{E}_q[\log q(\boldsymbol{z}) - \log p(\boldsymbol{z})]$$

We will assume that the prior $\boldsymbol{z}$ is a standard Gaussian:

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I}) = \prod_{i=1}^{D} p_i(z_i) = \prod_{i=1}^{D} \mathcal{N}(z_i; 0, 1)$$

Similarly we will assume that the variational approximation $q(\boldsymbol{z})$ is a fully factorized (i.e., diagonal) Gaussian:

$$q(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{D} q_i(z_i) = \prod_{i=1}^{D} \mathcal{N}(z_i; \mu_i, \sigma_i)$$

1. **(1pt)** Show that:

$$\mathcal{F}(q) = \log p(\boldsymbol{x}) - D_{KL}(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x}))$$

Haha, this one feels like a typical physics homework!

$$\begin{aligned}
F(q) &= E_q[\log(p(x|z))] - D_{KL}(q(z)||p(z)) \\
&= E_q[\log(p(x|z)) - \log(q(z)) + \log(p(z))] \\
&= E_q[\log(p(x|z)p(z)) - \log(q(z))] \\
&= E_q[\log(p(z|x)p(x)) - \log(z)] \\
&= E_q[\log(p(x)) + \log(p(z|x)) - \log q(z))] \\
&= \log(p(x)) - E_q[\log(p(z)) - \log(p(z|x))] \\
&= \log(p(x)) - D_{KL}(q(z)||p(z|x)) \quad \square
\end{aligned}$$

2. **(1pt)** Show that the KL term decomposes as a sum of KL terms for individual dimensions. In particular,

$$D_{KL}(q(\boldsymbol{z})||p(\boldsymbol{z})) = \sum_i D_{KL}(q_i(z_i)||p_i(z_i))$$

$$\begin{aligned}
D_{KL}(q(z)||p(z)) &= E_q[\log q(z) - \log p(z)] \\
&= E_q\left[\log\left(\prod_{i=1}^{D} q_i(z_i)\right) - \log\left(\prod_{i=1}^{D} p_i(z_i)\right)\right] \\
&= E_q\left[\sum_{i=1}^{D} (\log q_i(z_i) - \log p_i(z_i)\right. \\
&= \sum_{i=1}^{D} E_q[\log q_i(z_i) - \log p_i(z_i)] \\
&= \sum_{i=1}^{D} D_{KL}(q_i(z_i)||p_i(z_i)) \quad \square
\end{aligned}$$

3. **(2pts)** Give an explicit formula for the KL divergence $D_{KL}(q_i(z_i)||p_i(z_i))$. This should be a mathematical expression involving μ_i and σ_i.

$$D_{KL}(q_i(z_i)||p_i(z_i)) = E_q[\log(q_i(z_i)) - \log(p_i(z_i))]$$

$$= \int_{\mathbb{R}} dz\, q_i(z)\left(\log q_i(z) - \log p_i(z)\right)$$

$$= \int_{\mathbb{R}} dz\, q_i(z) \log \frac{q_i(z)}{p_i(z)}$$

$$= \int_{\mathbb{R}} dz\, q_i(z) \log\left[\frac{(\cancel{\sqrt{2\pi}}\sigma_i)^{-1} \exp\left(-\frac{(z-\mu_i)^2}{2\sigma_i^2}\right)}{\cancel{\sqrt{2\pi}}^{-1} \exp(-z^2/2)}\right]$$

$$= \int_{\mathbb{R}} dz\, q_i(z)\left[\log(\sigma_i^{-1}) + \log\left(\frac{\exp\left(-\frac{(z-\mu_i)^2}{2\sigma_i^2}\right)}{\exp(-z^2/2)}\right)\right]$$

$\int_{\mathbb{R}} dz\, q_i(z) = 1$

$$= \log(\sigma_i^{-1}) - \frac{1}{2\sigma^2}\int_{\mathbb{R}} dz\, q_i(z)(z-\mu_i)^2 + \frac{1}{2}\int_{\mathbb{R}} dz\, q_i(z)\, z^2$$

$$= \log(\sigma_i^{-1}) - \frac{Var_q[z_i]}{2\sigma^2} + \frac{1}{2}E_q[z_i^2]$$

$Var_q[z_i] = \sigma_i^2$

$E_q[z_i] = \mu_i$

$$= \log(\sigma_i^{-1}) - \frac{1}{2} + \frac{1}{2}(\sigma_i^2 + \mu_i^2)$$

$Var(x) = E(x^2) - E(x)^2$

$$= \frac{1}{2}\left(\mu_i^2 + \sigma_i^2 - \log(\sigma_i^2) - 1\right)$$

4. **(2pts)** One way to do gradient descent on the KL term is to apply the formula from above. Another approach is to compute stochastic gradients using the reparameterization trick:

$$\nabla_{\boldsymbol{\theta}} D_{KL}(q_i(z_i)||p_i(z_i)) = \mathbb{E}_{\epsilon}[\nabla_{\boldsymbol{\theta}} t_i]$$

, where

$$\boldsymbol{\theta} = \begin{bmatrix} \mu_i \\ \sigma_i \end{bmatrix}$$

and

$$\begin{aligned} z_i &= \mu_i + \sigma_i \epsilon_i \\ r_i &= \log q_i(z_i) \\ s_i &= \log p_i(z_i) \\ t_i &= r_i - s_i \end{aligned} \tag{2}$$

Show how to compute a stochastic estimate of $\nabla_{\boldsymbol{\theta}} D_{KL}(q_i(z_i)||p_i(z_i))$ by doing backpropagation on the above equations. You may find it helpful to draw the computation graph.

Let's start with $\bar{\theta}_i = \frac{\partial t_i}{\partial \theta_i} = [\bar{\mu}_i, \bar{\sigma}_i]$:

$\bar{t}_i = 1$, $\bar{r}_i = 1$, $\bar{s}_i = -1$

$$\bar{z}_i = \bar{r}_i \frac{\partial r_i}{\partial z_i} + \bar{s}_i \frac{\partial s_i}{\partial z_i} = \frac{\partial}{\partial z_i} \log q_i(z_i) - \frac{\partial}{\partial z_i} \log p_i(z_i)$$

$$= \frac{\partial}{\partial z_i}\left[\log q_i(z_i) - \log p_i(z_i)\right]$$

$$\frac{\partial}{\partial z_i} \log q_i(z) = \frac{\partial}{\partial z_i} \log\left[(\sqrt{2\pi}\,\sigma_i)^{-1} \exp\left(\frac{(z_i-\mu_i)^2}{2\sigma_i^2}\right)\right]$$

$$= \frac{\partial}{\partial z_i}\left[\log\left((\sqrt{2\pi}\,\sigma_i)^{-1}\right) - \frac{(z_i-\mu_i)^2}{2\sigma_i^2}\right]$$

$$= \frac{\partial}{\partial z_i}\left[-\frac{(z_i-\mu_i)^2}{2\sigma_i^2}\right] = \frac{\mu_i - z_i}{\sigma_i^2}$$

$$\frac{\partial}{\partial z_i} \log p_i(z_i) = \frac{\partial}{\partial z_i} \log\left[\sqrt{2\pi}^{-1} \exp(-z_i^2/2)\right]$$

$$= -\frac{\partial}{\partial z_i} \frac{z_i^2}{2} = -z_i$$

$$\Rightarrow \quad \bar{z}_i = \frac{\mu_i - z_i}{\sigma_i^2} + z_i = \bar{\mu}_i$$

$$\bar{\sigma}_i = \bar{z}_i \epsilon_i = \epsilon_i\left(\frac{\mu_i - z_i}{\sigma_i^2} + z_i\right)$$

We know that $z_i = \mu_i + \sigma_i \varepsilon_i$, so $\varepsilon_i = \frac{z_i - \mu_i}{\sigma_i}$
and thus $\varepsilon_i \sim \mathcal{N}(0,1)$

$$\Rightarrow \frac{\partial}{\partial \mu_i} D_{KL}(q_i(z_i) \| p_i(z_i)) = E_{\varepsilon_i}[\overline{\mu_i}]$$

$$= E_{\varepsilon_i}\left[\frac{\mu_i - z_i}{\sigma_i^2} + z_i \right]$$

$$= \underbrace{E_{\varepsilon_i}[-\varepsilon_i]}_{=0} + E_{\varepsilon_i}[\mu_i] + \sigma_i \underbrace{E_{\varepsilon_i}[\varepsilon_i]}_{=0}$$

$$= \mu_i$$

$$\frac{\partial}{\partial \sigma_i} D_{KL}(q_i(z_i) \| p_i(z_i)) = E_{\varepsilon_i}[\overline{\sigma_i}]$$

$$= E_{\varepsilon_i}\left[\varepsilon_i \left(\frac{\mu_i - z_i}{\sigma_i^2} + z_i \right) \right]$$

$$= E_{\varepsilon_i}[-\varepsilon_i^2 / \sigma_i] + E_{\varepsilon_i}[\varepsilon_i z_i]$$

$$= -\sigma_i^{-1} + \underbrace{E_{\varepsilon_i}[\varepsilon_i \mu_i]}_{= \mu_i E_{\varepsilon_i}[\varepsilon_i] = 0} + \underbrace{E_{\varepsilon_i}[\sigma_i \varepsilon_i^2]}_{= \sigma_i E_{\varepsilon_i}[\varepsilon_i^2] = \sigma_i}$$

$$= \sigma_i - \sigma_i^{-1}$$

So: $$\mathrm{grad}_{\Theta_i} D_{KL}(q_i(z_i) \| p_i(z_i)) = E_{\varepsilon_i}[\mathrm{grad}_{\Theta_i} t_i]$$

$$= \begin{pmatrix} \mu_i \\ \sigma_i - \sigma_i^{-1} \end{pmatrix}$$

Taking the derivative of the result from 3. gives the exact same gradient!