## Problem 1: Fréchet distance

Given $p_1(x) = \mathcal{N}(x; \mu_1, \Sigma_1)$ and $p_2(x) = \mathcal{N}(x; \mu_2, \Sigma_2)$, the Fréchet distance between $p_1$ and $p_2$ can be written in closed form as

$$\mathcal{W}_2^2(p_1, p_2) = \|\mu_1 - \mu_2\|_2^2 + \text{Tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right)$$

1. Find simpler expressions in the cases (1) $\Sigma_1 = \Sigma_2$, (2) $p$ is standard normal ($\mathcal{N}(0, I)$).
2. Show that $\mathcal{W}_2^2(p_1, p_2) = \mathcal{W}_2^2(p_2, p_1)$.
3. Fréchet inception distance (FID) is computed using Gaussians fit to features of an InceptionNet pretrained on ImageNet. Explain the motivation for using InceptionNet features instead of using Gaussians fit to raw pixel values.

## Problem 2: InfoNCE

For this problem, please read this paper. Given a dataset of i.i.d. samples $\{x_i\}_{i=1}^N$, we want learn a representation of $x_i$ given by $c_i = \text{Enc}(x_i)$. This problem can be solved by using (V)AE, i.e., $x \xrightarrow{\text{Enc}} c \xrightarrow{\text{Dec}} \hat{x} \approx x$. However, such a method requires training a reconstruction network Dec. Instead, one can try to maximise some measure of similarity between the distributions of $x_i$ and $c_i$. This is what InfoNCE loss allows us to do.

1. Following the paper, what are the disadvantages of training a reconstruction model Dec?
2. The paper proposes to use mutual information $I(x, c) = \text{KL}\Big(p(x, c) \| p(x) \otimes p(c)\Big)$ as a measure of similarity between distributions of $x_i$ and $c_i$.
   (a) Show that $I(x, c) = \mathbb{E}_{p(x,c)} \log \frac{p(x|c)}{p(x)}$
   (b) The InfoNCE loss for a tuple of $(x, x_{\text{neg}}, c)$ can be written as follows.

   $$\mathcal{L}(f) = -\mathbb{E}_{(x, x_{\text{neg}}, c)} \log \frac{f(x, c)}{f(x, c) + f(x_{\text{neg}}, c)}.$$

   Assuming that the optimum $f(x, c)$ is given by $\frac{p(x|c)}{p(x)}$ and $f(x_{\text{neg}}, c) = \frac{p(x_{\text{neg}}|c)}{p(x_{\text{neg}})} \approx 1$, show

   $$\mathcal{L}(f) \geq -\mathbb{E}_{(x,c)} \log \frac{p(x \mid c)}{p(x)} = -I(x, c).$$

   In other words, miminising $\mathcal{L}(f)$ allows to maximise mutual information between $x$ and $c$.

## Problem 3: Simple OT

Let $p(x) = \frac{1}{2}\mathbb{1}_{[0,2]}(x)$ and $q(x) = \frac{1}{2}\mathbb{1}_{[1,3]}(x)$, where $\mathbb{1}_A$ is the indicator function of $A$, i.e., $\mathbb{1}_A(x) = 1$ if $x \in A$ and $0$ otherwise.

1. Show that $f(x) = x + 1$ is the optimal transport map for the problem $\min_{f: f_{\#}p = q} \int_{\mathbb{R}} |x - f(x)|\, p(x)\, dx$. Notice that this problem is 1-OT (the absolute value of the distance is not squared), not 2-OT as used in FID.
2. Show that $f(x) = \begin{cases} x + 2 & \text{if } x \in [0, 1] \\ x & \text{if } x \in (1, 2] \end{cases}$ is another OT map with the same cost . What does this tell about the uniqueness of 1-OT maps?