Indian Institute of Technology Jodhpur
Machine Learning II: Fractal 3
Practice Problems

1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function that is $f(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) \leq tf(\mathbf{x}_1 + (1-t)f(\mathbf{x}_2)$. Let $\mathbf{x}$ be a random vector with joint PDF $p(\mathbf{x})$. If f is a convex function, then show that
$$\mathbb{E}_{\mathbf{x}\sim p}[f(\mathbf{x})] \geq f(\mathbb{E}_{\mathbf{x}\sim p}[\mathbf{x}]).$$

2. Consider that $p = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $q = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Here $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k)$. Then, shown that the KL divergence between p and q is defined as below.
$$D_{\text{KL}}(p|q) = \frac{1}{2}\sum_{i=1}^{k}\left(\sigma_i^2 + \mu_i^2 - 1 - \log_e(\sigma_i^2)\right).$$

3. Jensen-Shannon Divergence between two distributions f and g is defined as
$$\text{D}_{JS}(f\|g) = \frac{1}{2}\text{D}_{\text{KL}}\left(f\|\frac{f+g}{2}\right) + \frac{1}{2}\text{D}_{\text{KL}}\left(g\|\frac{f+g}{2}\right)$$
. Show that it is symmetric and zero only of $p = q$.

4. Consider the BiGAN and its loss function. Show that, in order to fool a perfect discriminator D, BiGAN encoder E and generator G must invert each other. That is, $G(E(\mathbf{x})) = \mathbf{x}$ and $E(G(\mathbf{z})) = \mathbf{z}$.

5. If generator G and discriminator D have enough capacity, and at each step of the training algorithm, the discriminator is allowed to reach its optimum given G, and p_G is updated so as to improve the criterion
$$\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\log D^\star(\mathbf{x})] + \mathbb{E}_{\mathbf{x}\sim p_G}[\log(1 - D^\star(\mathbf{x}))]$$
then p_G converges to p_{data}.

6. Show that the global minimum of $C(G) = \underset{D}{\max} V(G, D)$ is achieved if and only if $p_G = p_{\text{data}}$. At that point, show that $C(G)$ achieves the value $-\log 4$.

7. Consider a coin with head probability equal to r. Let X be a random variable such that $X = -1$, if head appears and $X = 1$, if tail appears. Consider the probability mass function (PMF) p_X of X when $r = 0.2$ and the PMF q_X of X when $r = 0.8$. Find the KL divergence between p_X and q_X.

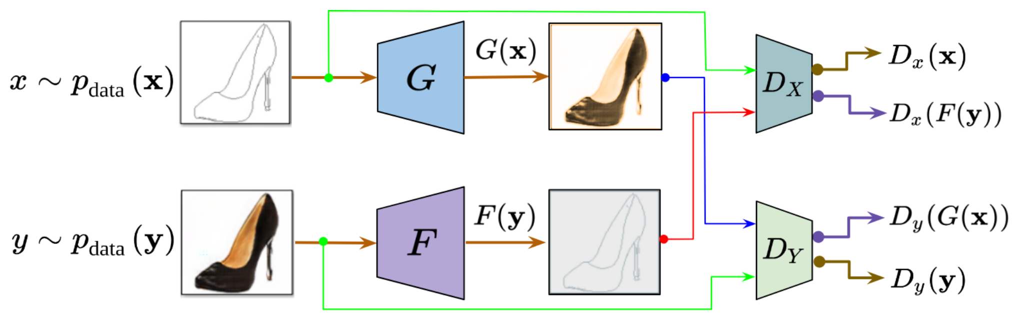
8. Find the optimal values of $\mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k$ that minimize the cost function as defined below.
$$f(\mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k) = \frac{1}{2}\sum_{i=1}^{k}\left(\mu_i^2 + \sigma_i^2 - 1 - \log_e(\sigma_i^2)\right).$$

9. Given a dataset $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ containing samples drawn from an unknown data distribution $p(\mathbf{x})$, we want to learn a distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$ that is as close as possible to the true distribution $p(\mathbf{x})$. Consider two cost functions to find the optimal $p_{\boldsymbol{\theta}}(\mathbf{x})$: $D_{\text{KL}}(p(\mathbf{x})|p_{\boldsymbol{\theta}}(\mathbf{x}))$ and $\int_{\mathbf{x}}(p_{\boldsymbol{\theta}}(\mathbf{x}) - p(\mathbf{x}))^2 d\mathbf{x}$. Choose the best cost function out of these.

10. Consider Cycle-GAN using which we solve the problem of Image-to-image translation where we want to learn a mapping that translate X into Y and Y into X. Consider trying to train this network using only the GAN losses that are defined as:

$$\begin{aligned}\ell_{\text{GAN}}(F, D_x) &= \mathbb{E}_{\mathbf{x}}[\log D_x(\mathbf{x})] + \mathbb{E}_{\mathbf{y}}[\log(1 - D_x(F(\mathbf{y})))] \\ \ell_{\text{GAN}}(G, D_y) &= \mathbb{E}_{\mathbf{y}}[\log D_y(\mathbf{y})] + \mathbb{E}_{\mathbf{x}}[\log(1 - D_y(G(\mathbf{x})))]\end{aligned}$$

Find the optimal discriminators $D_x^\star$ and $D_y^\star$ that maximize $\ell_{\text{GAN}}(F, D_x)$ and $\ell_{\text{GAN}}(G, D_y)$, respectively. Then, find the optimal generators F and G that minimize $\ell_{\text{GAN}}(F, D_x^\star)$ and $\ell_{\text{GAN}}(G, D_y^\star)$, respectively.



11. Given a dataset $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ containing samples drawn from an unknown data distribution $p(\mathbf{x})$, we want to learn a distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$ that is as close as possible to the true distribution $p(\mathbf{x})$. We have to minimize the KL divergence between the distributions $p(\mathbf{x})$ and $p_{\boldsymbol{\theta}}(\mathbf{x})$ with respect to $\boldsymbol{\theta}$ to find the optimal values of the parameters $\boldsymbol{\theta}$. Show that minimizing the KL divergence is equivalent to maximize the log likelihood $\log p_{\boldsymbol{\theta}}(\mathbf{x})$ with the parameters $\boldsymbol{\theta}$. To maximize this log-likelihood function, we used VAE as one of the approach where we use latent variables $\mathbf{z}$ that semantically represent the input dataset. Then, we learn the conditional distribution $p(\mathbf{x}|\mathbf{z})$ and use it to sample new points from the distribution $p(\mathbf{x})$. Now, consider a distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ parametrized by the parameters $\boldsymbol{\phi}$. Then, show that the log-likelihood $\log p_{\boldsymbol{\theta}}(\mathbf{x})$ is equal to

$$-D_{\text{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})|p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\right) + D_{\text{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})|p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})\right).$$

As directly maximizing the log-likelihood is intractable, we maximize the lower bound (ELBO) $\ell(\boldsymbol{\phi}, \boldsymbol{\theta}) = -D_{\text{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})|p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\right)$ on it. How can you ensure that the gap between the log likelihood and the ELBO is zero? Now, to find optimal parameters $\boldsymbol{\theta}$, we need to find the gradient of $\ell(\boldsymbol{\phi}, \boldsymbol{\theta})$. Show that

$$\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\phi}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}(\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})).$$

To find optimal $\boldsymbol{\phi}$, we need to find the gradient of $\ell(\boldsymbol{\phi}, \boldsymbol{\theta})$. Can we find the gradient $\nabla_{\boldsymbol{\phi}}\ell(\boldsymbol{\phi}, \boldsymbol{\theta})$ using the same approach as used for finding $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\phi}, \boldsymbol{\theta})$? If yes, find an expression for $\nabla_{\boldsymbol{\phi}}\ell(\boldsymbol{\phi}, \boldsymbol{\theta})$, if not, then suggest what difficulty you face and an approach to solve this issue.