Indian Institute of Technology Jodhpur
Machine Learning II, Major Exam

Date: January 31, 2021, Max Marks: 20, Max Time: 3 Hours

`Attempt all the questions. Best of luck` ☺

1. 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. **[1 Mark]**

2. Find the optimal values of $\mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k$ that minimize the cost function as defined below. **[1 Mark]**

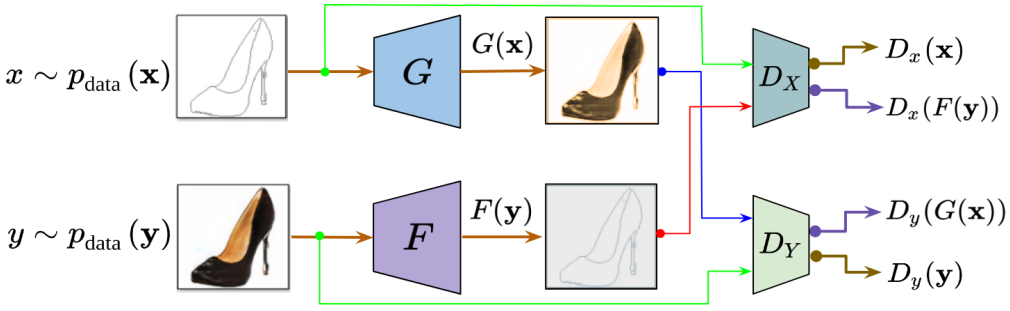
$$f(\mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k) = \frac{1}{2} \sum_{i=1}^{k} \left(\mu_i^2 + \sigma_i^2 - 1 - \log_e(\sigma_i^2)\right).$$

3. 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 $p_{\boldsymbol{\theta}}(\mathbf{x})$: $D_{\mathrm{KL}}(p(\mathbf{x})|p_{\boldsymbol{\theta}}(\mathbf{x}))$ and $\int_{\mathbf{x}}(p_{\boldsymbol{\theta}}(\mathbf{x}) - p(\mathbf{x}))^2 d\mathbf{x}$. Choose the best cost function out of these. **[1 Mark]**

4. 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:

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

Find the optimal discriminators $D_x^\star$ and $D_y^\star$ that maximize $\ell_{\mathrm{GAN}}(F, D_x)$ and $\ell_{\mathrm{GAN}}(G, D_y)$, respectively. Then, find the optimal generators F and G that minimize $\ell_{\mathrm{GAN}}(F, D_x^\star)$ and $\ell_{\mathrm{GAN}}(G, D_y^\star)$, respectively. **[2+2 Marks]**



5. 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

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

As directly maximizing the log-likelihood is intractable, we maximize the lower bound (ELBO) $\ell(\boldsymbol{\phi},\boldsymbol{\theta}) = -D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})|p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\right)$ 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

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

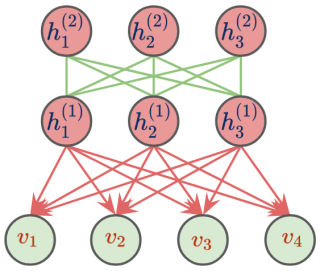
To find optimal $\boldsymbol{\phi}$, we need to find the gradient of $\ell(\boldsymbol{\phi},\boldsymbol{\theta})$. Can we find the gradient $\nabla_{\boldsymbol{\phi}}\ell(\boldsymbol{\phi},\boldsymbol{\theta})$ using the same approach as used for finding $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\phi},\boldsymbol{\theta})$? 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. **[0.5+2+0.5+0.5+0.5 Marks]**

6. Given a set of points $\{\mathbf{v}_i\}_{i=1}^n$, we model $p_{\boldsymbol{\theta}}(\mathbf{v})$ that is as close as possible to true distribution $p_{\mathrm{data}}(\mathbf{v})$. Here, all vectors $\mathbf{v}_i \in \{0,1\}^{n_v \times 1}$. To solve this generative modeling problem, we employed the restricted Boltzmann machine (RBM) framework. Here, we model the joint distribution of the visible variables $\mathbf{v}$ and hidden variables $\mathbf{h} \in \{0,1\}^{n_h \times 1}$ using the energy based model, i.e., $p(\mathbf{v},\mathbf{h}) = \frac{e^{-E(\mathbf{v},\mathbf{h})}}{Z}$. Here, the functions $E(\mathbf{v},h)$ denotes the energy function and defined as $E(\mathbf{v},\mathbf{h}) = -\mathbf{v}^\top\mathbf{W}\mathbf{h} - \mathbf{b}^\top\mathbf{v} - \mathbf{c}^\top\mathbf{h}$. Here, $\mathbf{W}$, $\mathbf{b}$, and $\mathbf{c}$ are the learnable parameters of the RBM. Explain why there is no terms in the energy function representing interactions among visible variables or among hidden variables. We denote $\boldsymbol{\theta} = \{\mathbf{W},\mathbf{b},\mathbf{c}\}$. In order to find the optimal parameters of the RBM, we maximize the log-likelihood function $\ell(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \log p_{\boldsymbol{\theta}}(\mathbf{v}_i)$. Show that

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \mathbf{W}_{ij}} &= \mathbb{E}_{\mathbf{v}\sim p_{\mathrm{data}}}[v_i h_j] - \mathbb{E}_{(\mathbf{v},\mathbf{h})\sim p_{\boldsymbol{\theta}}}[v_i h_j] \\
\frac{\partial \ell(\boldsymbol{\theta})}{\partial b_i} &= \mathbb{E}_{\mathbf{v}\sim p_{\mathrm{data}}}[v_i] - \mathbb{E}_{(\mathbf{v},\mathbf{h})\sim p_{\boldsymbol{\theta}}}[v_i] \\
\frac{\partial \ell(\boldsymbol{\theta})}{\partial c_j} &= \mathbb{E}_{\mathbf{v}\sim p_{\mathrm{data}}}[h_j] - \mathbb{E}_{(\mathbf{v},\mathbf{h})\sim p_{\boldsymbol{\theta}}}[h_j]
\end{aligned}$$

Now, explain what makes these gradients difficult to evaluate. Also, propose an approach to evaluate these gradient efficiently. **[2+0.75+0.75+0.75+0.75 Marks]**

7. Consider the below deep belief network with 2 hidden layers having weight matrices $\mathbf{W}^{(1)}$,$\mathbf{W}^{(2)}$ and bias vectors $\mathbf{b}^{(0)}$, $\mathbf{b}^{(1)}$, and $\mathbf{b}^{(2)}$ with $\mathbf{b}^{(0)}$ providing the biases for the visible layer. Then, find the probability distributions $p(h_1^{(2)},\mathbf{h}^{(1)})$ and $p(v_2|\mathbf{h}^{(1)})$. To train this DBN, we first train an RBM to



maximize $\mathbf{E}_{\mathbf{v}\sim p_{\mathrm{data}}}[\log p(v)]$ using the contrastive-divergence algorithm. This RBM consists of visible layer and the first hidden layer. Next, we train a second RBM with inputs as the output of the first layer and second hidden layer as the hidden layer of it. In order to train this second RBM, show that we have to maximize $\mathbf{E}_{\mathbf{v}\sim p_{\mathrm{data}}}\left[\mathbf{E}_{\mathbf{h}^{(1)}\sim p^{(1)}(\mathbf{h}^{(1)}|\mathbf{v})}\left[\log p^{(2)}(\mathbf{h}^{(1)})\right]\right]$. Here, $p^{(1)}$ is the probability distribution represented by the first RBM and $p^{(2)}$ is the probability distribution represented by the second RBM. Here, p_{data} is the true distribution from which input data is drawn and p is the distribution we want to find that is as close as possible to the true distribution. **[1+1+2 Marks]**