CMSC 720 - Final Exam 2024

Name:

UID:

This exam is for the internal use in CMSC 720 Spring 2024 class. Please do NOT share.

Question 1:

Consider a latent variable Z and an observed variable X that depends on Z. Let $P(X,Z)$ be a joint distribution between these variables. Also, consider a separate distribution $Q(Z)$.

(a) (10 points) The KL Divergence between $Q(Z)$ and $P(Z|X)$ is defined as:

$$KL(Q(.)||P(.|X) = \mathbb{E}_{Z\sim Q}\left[\log\left(\frac{Q(Z)}{P(Z|X)}\right)\right]$$

Use Jensen's Inequality to show that $KL(Q(.)||P(.|X)) \geq 0$.

Note: Given a convex function $f(.)$, Jensen's Inequality states that $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$

$\log(.)$ is a concave function, making $-\log(.)$ a convex function. Therefore, we write the negative KL divergence as,

$$\begin{aligned} -\mathbb{E}_{Z\sim Q}\left[\log\left(\frac{Q(Z)}{P(Z|X)}\right)\right] &= \mathbb{E}_{Z\sim Q}\left[\log\left(\frac{P(Z|X)}{Q(Z)}\right)\right] \\ -KL(Q(.)||P(.|X)) &\leq \log\left(\mathbb{E}_{Z\sim Q}\left[\frac{P(Z|X)}{Q(Z)}\right]\right) && \text{(Jensen's Inequality)} \\ &\leq \log\left(\sum_Z Q(Z)\frac{P(Z|X)}{Q(Z)}\right) \\ &\leq \log(1) \\ &\leq 0 \end{aligned}$$

(b) (15 points) The Variational Lower Bound L is defined as:

$$L(X) = \mathbb{E}_{Z\sim Q}\left[\log P(X,Z)\right] - \mathbb{E}_{Z\sim Q}\left[\log Q(Z)\right]$$

Use part (a) to show that $\log P(X) \geq L$. Under what condition is equality achieved?

$$\begin{aligned} L(X) &= \mathbb{E}_{Z\sim Q}\left[\log P(X,Z)\right] - \mathbb{E}_{Z\sim Q}\left[\log Q(Z)\right] \\ &= \mathbb{E}_{Z\sim Q}\left[\log P(Z|X)P(X)\right] - \mathbb{E}_{Z\sim Q}\left[\log Q(Z)\right] \\ &= \mathbb{E}_{Z\sim Q}[\log P(Z|X)] + \log P(X) - \mathbb{E}_{Z\sim Q}\left[\log Q(Z)\right] \\ &= \log P(X) + \mathbb{E}_{Z\sim Q}[\log P(Z|X) - \log Q(Z)] \\ &= \log P(X) - \mathbb{E}_{Z\sim Q}\left[\log\frac{Q(Z)}{P(Z|X)}\right] && \text{(second term shown to be greater than 0 in (a))} \end{aligned}$$

$\log P(X) \geq L$

(c) (5 points) Let $Q(Z) = \mathcal{N}(\mu, I)$. Our goal is to maximize the variational lower bound on $P(X)$, i.e.,

$$\max_{\mu} \mathbb{E}_{Z\sim Q}[\log P(X,Z)] - \mathbb{E}_{Z\sim Q}[\log Q(Z)]$$

In order to do this, we plan to use gradient descent. Can we write the gradient descent update as follows? Why or why not?

$$\mu^{(t+1)} := \mu^{(t)} + \alpha \mathbb{E}_{Z\sim Q} \nabla_{\mu} \left[\log P(X,Z) - \log Q(Z)\right]$$

No. We cannot bring the expectation $\mathbb{E}_{Z\sim Q}$ outside the gradient operation because it depends on Q which is parameterized by μ.

Question 2:

Consider two discrete distributions P and Q defined as the following:

$$P = \begin{cases} 1/3, X = -1 \\ 1/3, X = 0 \\ 1/3, X = 1 \end{cases}$$

$$Q = \begin{cases} 1/2, X = 1 \\ 1/2, X = -1 \end{cases}$$

Using the transportation cost between two points x and z as $|x - z|$, answer the following questions:

(a) (10 points) Find a valid transport plan between P and Q? What is its average transport cost (show your calculation)?

Sol: Let $\Pi(P,Q)$ be the set of valid transport plans from P to Q (i.e. $\Pi(P,Q)$ contains all distributions over $P \times Q$ whose marginals are respectively P and Q), then a valid transport plan $\pi \in \Pi(P,Q)$ is as follows:

$$\pi(p,q) = \begin{cases} 1/3 & (p,q) = (-1,-1) \\ 1/3 & (p,q) = (1,1) \\ 1/6 & (p,q) = (0,-1) \\ 1/6 & (p,q) = (0,1) \\ 0 & otherwise \end{cases}$$

Its average transport cost is

$$1/3 \cdot |-1-(-1)| + 1/3 \cdot |1-1| + 1/6 \cdot |0-(-1)| + 1/6 \cdot |0-1| = 1/3$$

(b) (10 points) What is the optimal transport cost between P and Q? Show your proof.

Sol: It is 1/3. From (a), we already show there is a transport plan with a transport cost of 1/3, now we show that any valid transport plan has a cost of at least 1/3.

Let $\pi \in \Pi(P,Q)$ be any valid transport plan. Since the marginals for π are respectively P and Q, we have

$$\begin{cases} \sum_q \pi(0,q) = P(0) \\ \pi(0,q) = 0 \qquad \text{when } q \notin \{-1,1\} \end{cases}$$

Thus let c be its transport cost, we have

$$\begin{aligned}
c \geq & \sum_{q} \pi(0, q) \cdot |0 - q| \\
= & \sum_{q \in \{1, -1\}} \pi(0, q) \cdot |0 - q| \\
\geq & \sum_{q \in \{1, -1\}} \pi(0, q) \cdot \min\left(|0 - 1|, |0 - (-1)|\right) \\
= & \sum_{q \in \{1, -1\}} \pi(0, q) \\
= & P(0) = 1/3
\end{aligned}$$

Thus $1/3$ is the optimal transport cost.

Question 3:

In logistic regression, given a training set $S = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), ..., (\mathbf{x}^{(m)}, y^{(m)})\}$, the objective is to minimize the following loss function:

$$L(\theta) = -\left[\sum_{i=1}^{m} y^{(i)} \log\left[\sigma(\theta^T \mathbf{x}^{(i)})\right] + (1 - y^{(i)}) \log\left[1 - \sigma(\theta^T \mathbf{x}^{(i)})\right]\right]$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$. In this problem, we instead consider a regularized loss function defined as follows:

$$L(\theta) = -\left[\sum_{i=1}^{m} y^{(i)} \log\left[\sigma(\theta^T \mathbf{x}^{(i)})\right] + (1 - y^{(i)}) \log\left[1 - \sigma(\theta^T \mathbf{x}^{(i)})\right]\right] + \frac{\lambda}{2}\|\theta\|^2$$

The term $\frac{\lambda}{2}\|\theta\|^2$ is the regularized term added to the original loss. λ is a fixed and positive number.

(a) (10 points) Compute the gradient of $L(\theta)$ with respect to θ, i.e., $\nabla_\theta L(\theta)$.

$$\begin{aligned}
\frac{\partial}{\partial \theta} L(\theta) = & -\left[\sum_{i=1}^{m} y^{(i)} \frac{1}{\sigma(\theta^T \mathbf{x}^{(i)})} \sigma(\theta^T \mathbf{x}^{(i)})(1 - \sigma(\theta^T \mathbf{x}^{(i)}))\mathbf{x}^{(i)}\right. \\
& \left. + (1 - y^{(i)}) \frac{1}{1 - \sigma(\theta^T \mathbf{x}^{(i)})} (-\sigma(\theta^T \mathbf{x}^{(i)})(1 - \sigma(\theta^T \mathbf{x}^{(i)})))\mathbf{x}^{(i)}\right] + \frac{\lambda}{2} 2\theta \\
= & -\left[\sum_{i=1}^{m} y^{(i)}(1 - \sigma(\theta^T \mathbf{x}^{(i)}))\mathbf{x}^{(i)} + (1 - y^{(i)})(-\sigma(\theta^T \mathbf{x}^{(i)}))\mathbf{x}^{(i)}\right] + \lambda\theta \\
= & -\left[\sum_{i=1}^{m} y^{(i)}\mathbf{x}^{(i)} - y^{(i)}(\sigma(\theta^T \mathbf{x}^{(i)}))\mathbf{x}^{(i)} - (\sigma(\theta^T \mathbf{x}^{(i)}))\mathbf{x}^{(i)} + y^{(i)}(\sigma(\theta^T \mathbf{x}^{(i)}))\mathbf{x}^{(i)}\right] + \lambda\theta \\
= & -\left[\sum_{i=1}^{m} y^{(i)}\mathbf{x}^{(i)} - (\sigma(\theta^T \mathbf{x}^{(i)}))\mathbf{x}^{(i)}\right] + \lambda\theta \\
= & \left[\sum_{i=1}^{m} (\sigma(\theta^T \mathbf{x}^{(i)}) - y^{(i)})\mathbf{x}^{(i)}\right] + \lambda\theta
\end{aligned}$$

Therefore, $\nabla_\theta L(\theta) = \sum_{i=1}^{m} (\sigma(\theta^T \mathbf{x}^{(i)}) - y^{(i)})\mathbf{x}^{(i)} + \lambda\theta$

(b) (10 points) Compute the Hessian of $L(\theta)$ with respect to θ, i.e., $\nabla_\theta^2 L(\theta)$

$$\nabla_\theta^2 L(\theta) = \sum_{i=1}^{m} \sigma(\theta^T \mathbf{x}^{(i)})(1 - \sigma(\theta^T \mathbf{x}^{(i)}))\mathbf{x}^{(i)}\mathbf{x}^{(i)T} + \lambda I$$

(c) (5 points) Is $L(\theta)$ a convex function? Why?

$\mathbf{x}^{(i)}\mathbf{x}^{(i)T}$ is an outer product so it is Positive Semi-Definite (PSD). $\sigma(\theta^T\mathbf{x}^{(i)})(1-\sigma(\theta^T\mathbf{x}^{(i)}))$ is also PSD because the range of the sigmoid function is $(0,1)$. We also know that $\lambda > 0$. Therefore, since the Hessian $\nabla^2_\theta L(\theta)$ is PSD, that implies $L(\theta)$ is convex.

(d) (5 points) Write down the gradient descent algorithm for optimizing θ.

Using the gradient $\nabla_\theta L(\theta)$ we computed in part (a), we can update the weights at step t to be,

```
for t=1 to T,
    $\theta_{t+1} = \theta_t - \eta\left[\sum_{i=1}^m(\sigma(\theta_t^T\mathbf{x}^{(i)}) - y^{(i)})\mathbf{x}^{(i)} + \lambda\theta_t\right]$
```

where η is the learning rate.

(e) (5 points) Write down the stochastic gradient descent (SGD) with batch size of 1 for optimizing θ.

```
for t=1 to T,
    Sample an index $k \in \{1, 2, ...m\}$
    Compute the gradient $\nabla_\theta L(\theta)$ for sample $\mathbf{x}^{(k)}$ and use that to update the weights.
    $\theta_{t+1} = \theta_t - \eta\left[(\sigma(\theta_t^T\mathbf{x}^{(k)}) - y^{(k)})\mathbf{x}^{(k)} + \lambda\theta_t\right]$
```

where η is the learning rate.

Question 4:

(10 points) What is LPIPS? Describe the steps needed to compute LPIPS.

LPIPS (Learned Perceptual Image Patch Similarity) is a metric that uses deep features (from a neural network) of a pair of images to compute their perceptual distance.

Let $g(.)$ be a CNN of L layers and $g_l(.)$ represent the output of the l^{th} layer. Given an image $\mathbf{x}$, we take the channel-normalized outputs $\hat{g}_l(\mathbf{x})$ for all l and normalize by the layer size (height, width)

$$\left\{\frac{\hat{g}_1(\mathbf{x})}{\sqrt{w_1h_1}}, \frac{\hat{g}_2(\mathbf{x})}{\sqrt{w_2h_2}}, \ldots \frac{\hat{g}_L(\mathbf{x})}{\sqrt{w_Lh_L}}\right\} = \phi(\mathbf{x})$$

The LPIPS distance between two images $\mathbf{x}^{(i)}$, $\mathbf{x}^{(j)}$ is defined as,

$$d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \|\phi(\mathbf{x}^{(i)}) - \phi(\mathbf{x}^{(j)})\|_2$$

Question 5:

(20 points) Score function is defined as,

$$\delta(X) = \nabla_X \log q(X)$$

where $q(X)$ is a probability density function defining the underlying distribution. Suppose we draw N samples in an i.i.d manner from a normal distribution $\mathcal{N}(0, I)$ as $\{X_1^{(0)}, X_2^{(0)}, ...X_N^{(0)}\} = X^{(0)}$. At each step, we update these samples following the score function,

$$X_i^{(t)} = X_i^{(t-1)} + \beta\delta(X_i^{(t-1)})$$

for a sufficiently small β. If $t \to \infty$, would the distribution of X_i^∞ converge to $q(X)$? Why?

At each step of the updation, we have a $\delta(X_i^{(t-1)})$ that points towards the steepest ascent with respect to the underlying distribution q.

Since q is the normal distribution here, we have

$$\log q(X) \propto -\|X\|^2$$
$$\implies \nabla_X \log q(X) = -c_0 X$$

where $c_0 > 0$ is some constant. This means that at initial iteration, the update is

$$X^{(1)} = X^{(0)} + \beta c_0(-X^{(0)}).$$

Note that here $X^{(0)}$ and $\beta c_0(-X^{(0)})$ both are normal distributions with mean $\mathbf{0}$ since the latter distribution is a linearly scaled version of the former. Therefore, $X^{(1)}$ can be defined using a normal distribution of mean $\mathbf{0}$ as the sum of two normal distributions is another normal distribution.
Similarly in the next iterations, since the underlying distribution remains to be a normal distribution with mean $\mathbf{0}$,

$$\begin{aligned} X^{(i+1)} &= X^{(i)} - c_i\beta X^{(i)} \\ &= X^{(i)}(1 - c_i\beta) \\ &= X^{(0)}\Pi_{j=0}^{i}(1 - c_j\beta) \end{aligned}$$

Since $c_j > 0 \ \forall j$ and β is sufficiently small, $\Pi_{j=0}^{i}(1 - c_j\beta)$ will converge to 0 as $i \to \infty$ (products of positive numbers less than 1). Therefore, $X^{(i)}$ will converge to $\mathbf{0}$ or the mode of the distribution and not converge to $q(X)$.

Question 6:

In each step of the forward pass of a diffusion process, Gaussian noise is being added to the input as,

$$q(X_t|X_{t-1}) \sim \mathcal{N}\left(\sqrt{1-\beta_t}X_{t-1}, \beta_t I\right)$$

(a) (10 points) What is the distribution of X_t given X_0?

Given the distribution $q(X_t|X_{t-1}) \sim \mathcal{N}\left(\sqrt{1-\beta_t}X_{t-1}, \beta_t I\right)$, we can write

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\epsilon$$

where $\epsilon \sim \mathcal{N}(0, I)$, such that we scale down X_{t-1} by a factor and add noise. Let us write the above computation for each t.

$$\begin{aligned} X_1 &= \sqrt{1-\beta_1}X_0 + \sqrt{\beta_1}\epsilon_0 \\ X_2 &= \sqrt{1-\beta_2}X_1 + \sqrt{\beta_2}\epsilon_1 \\ X_2 &= \sqrt{1-\beta_2}\sqrt{1-\beta_1}X_0 + \sqrt{1-\beta_2}\sqrt{\beta_1}\epsilon_0 + \sqrt{\beta_2}\epsilon_1 \end{aligned}$$

Let $\bar{\alpha_2} = (1-\beta_2)(1-\beta_1)$, then the terms $\sqrt{1-\beta_2}\sqrt{\beta_1}\epsilon_0 + \sqrt{\beta_2}\epsilon_1$ can be written as a scaled gaussian $\sim \mathcal{N}(0, ((1-\beta_2)\beta_1 + \beta_2)I) \sim \mathcal{N}(0, (\beta_1 - \beta_2\beta_1 + \beta_2)I) \sim \mathcal{N}(0, (1-\bar{\alpha_2})I)$. Therefore,

$$X_2 = \sqrt{\bar{\alpha_2}}X_0 + \sqrt{1-\bar{\alpha_2}}\epsilon$$

where $\epsilon \sim \mathcal{N}(0, I)$. Therefore, the distribution X_t is given by,

$$X_t = \sqrt{\bar{\alpha_t}}X_0 + \sqrt{1-\bar{\alpha_t}}\epsilon$$

where $\bar{\alpha_t} = (1-\beta_1)(1-\beta_2)...(1-\beta_t)$.

(b) (10 points) If $t \to \infty$, what is the limit distribution?

Let $(1-\beta_t) = \alpha_t$. We choose $0 < \alpha_t < 1$, say $\alpha = \frac{1}{2}$. Then suppose $t = 1000$, then,

$$\bar{\alpha_t} = \left(\frac{1}{2}\right)^{1000} \sim 0$$

Therefore, if $t \to \infty$, from part (a) X_t will only depend on $\epsilon \sim \mathcal{N}(0, I)$, therefore,

$$q(X_t|X_{t-1}) \to \mathcal{N}(0, I)$$

Question 7:

(15 points) If a function $f(\cdot)$ is L-Lipschitz, show that it translates to provable robustness with respect to the underlying metric.
Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz in terms of L_2 norm.

$$\begin{aligned} \implies |f(\mathbf{x} + \epsilon) - f(\mathbf{x})| &\leq L\|\mathbf{x} + \epsilon - \mathbf{x}\|_2 && \text{(L-Lipschitz)} \\ &= L\|\epsilon\|_2 \end{aligned}$$

This implies that for small perturbations within an L_2 radius of $\|\epsilon\|_2$ in the input space, $f(\cdot)$ would only differ by a value of $L\|\epsilon\|_2$.