# CS280 Fall 2019 Assignment 1
# Part A

Basic Neural Networks

Due on October 6, 2019

**Name:**

**Student ID:**

**1. Hessian in Logistic Regression (10 points)**

Let $\sigma(a) = \frac{1}{1+e^{-a}}$ be an activation function, the loss function of LR is

$$f(\mathbf{w}) = -\sum_{i=1}^{n}[y_i \log(\mu_i) + (1 - y_i)\log(1 - \mu_i)],$$

where $\mu_i = \sigma(\mathbf{w}^\intercal \mathbf{x}_i)$. (Assume $\mathbf{w}, \mathbf{x}_i \in \mathbb{R}^d, X \in \mathbb{R}^{n\times d}, X_i \in \mathbb{R}^{1\times d}$ )

- Show that the Hessian of $f$ can be written as $H = X^\intercal S X$, where $S = diag(\mu_1(1 - \mu_1), \cdots, \mu_n(1 - \mu_n))$ and $X = [X_1, \cdots, X_n]^\intercal$

**2. Linear Regression (5 points)**

Linear regression has the form

$$f(x) = E[y|x] = b + \mathbf{w}^\intercal \mathbf{x}.$$

- It is possible to solve for $\mathbf{w}$ and $b$ separately. Show that

$$b = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\intercal \mathbf{w} = \bar{y} - \bar{\mathbf{x}}^\intercal \mathbf{w}$$

**3. Gradient descent for fitting GMM (15 points)**

Consider the Gaussian mixture model

$$p(\mathbf{x}|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)$$

where $[\cdots, \pi_k, \cdots] \sim \text{Multinomial}(\phi), \phi_k \geq 0, \sum_{j=1}^{k} \phi_j = 1$. (Assume $\mathbf{x}, \mu_k \in \mathbb{R}^d$, $\mathbf{\Sigma}_k \in \mathbb{R}^{d\times d}$)

Define the log likelihood as

$$l(\theta) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\theta)$$

Denote the posterior responsibility that cluster $k$ has for datapoint $n$ as follows:

$$r_{nk} := p(z_n = k|\mathbf{x}_n, \theta) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \mathbf{\Sigma}_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(\mathbf{x}_n|\mu_{k'}, \mathbf{\Sigma}_{k'})}$$

- (5 points) Show that the gradient of the log-likelihood wrt $\mu_k$ is

  $$\frac{d}{d\mu_k} l(\theta) = \sum_n r_{nk} \mathbf{\Sigma}_k^{-1}(\mathbf{x}_n - \mu_k)$$

- (5 points) Derive the gradient of the log-likelihood wrt $\pi_k$, if without considering the constraint on $\pi_k$.

  **Bonus** (2 points): what if with the constraint $\sum_k \pi_k = 1$. (hint: reparameterization using the softmax function)

- (5 points) Derive the gradient of the log-likelihood wrt $\mathbf{\Sigma}_k$.

  **Bonus** (3 points): what if with the constraint that $\mathbf{\Sigma}_k$ is symmetric positive definitive. (hint: reparameterization using Cholesky Decomposition)