MAT 180 Theoretical Homework 4

Problem 1: Let G be the Softmax function so $G : \mathbb{R}^n \to \mathbb{R}^n$ is defined by

$$G_i(\mathbf{x}) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}.$$

Prove first that for any function $f : \mathbb{R}^n \to \mathbb{R}^n$ we have

$$\frac{\partial f_i(\mathbf{x})}{\partial x_j} = f_i(\mathbf{x}) \frac{\partial}{\partial x_j} \log(f_i(\mathbf{x})),$$

and use this to compute the Jacobian matrix

$$\frac{\partial G(\mathbf{x})}{\partial \mathbf{x}}.$$

The answer must be a matrix equation of the form

$$\mathrm{diag}(\mathbf{u}) - \mathbf{v}\mathbf{w}^\top$$

for suitable vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$.

Problem 2: Compute the derivatives for the functions $\mathrm{ReLU}, \mathrm{Linear}, \mathrm{Squared}, \mathrm{Sigmoid} : \mathbb{R}^n \to \mathbb{R}^n$ as you did for Softmax in Problem 1. These are all diagonal maps of the form $G(\mathbf{x}) = (g(x_1), \ldots, g(x_n))$. Of course ReLU is not differentiable when any of the inputs are zero, so just give the derivative where it is differentiable.

Problem 3: Let $\mathcal{N}$ be a neural network with architecture $P(n_0, \ldots, n_D)$ and depth D (so layer l has n_l neurons and every possible connection to adjacent layers). Suppose we are using $\mathcal{N}$ to fit a supervised dataset $\mathbf{X}, \mathbf{Y}$ where $\mathbf{X} \in \mathbb{R}^{m \times n}, \mathbf{Y} \in \mathbb{R}^{m \times k}$. To introduce regularization into our optimization algorithm, it is useful to consider a cost function of the form

$$J(W, B) = \frac{1}{m} \sum_{i=1}^{m} L(\mathcal{N}\mathbf{X}_{i,*}, \mathbf{Y}_{i,*}) + \lambda \mathrm{Reg}(W).$$

Let us consider the regularization term

$$\mathrm{Reg}(W) = \sum_{l=0}^{D-1} ||\mathbf{W}^{(l)}||^2_{\mathrm{Frob}}.$$

Compute $\frac{\partial \mathrm{Reg}(W)}{\partial \mathbf{W}^{(l)}}$ for $0 \le l \le D-1$. Write down formulas for updating the values of $\mathbf{W}^{(l)}$ and $\mathbf{B}^{(l)}$ in a stochastic gradient descent step by modifying the unregularized formulas given in lecture. Also write down the formulas in the case of the momentum method.