CSI 5138 Homework Exercise 2

Question 1 (25 points) *Suppose that $\mathcal{L}(x)$ is a scalar-valued function with variable $x \in \mathbb{R}^K$. The function is defined via the following sequence of function compositions, where A and B are both $K \times K$ matrices:*

$$
\begin{aligned}
y &:= Ax && (1)\\
u &:= \sigma(y) && (2)\\
v &:= Bx && (3)\\
z &:= A\,(u \odot v) && (4)\\
w &:= Az && (5)\\
\mathcal{L} &:= \|w\|^2 && (6)
\end{aligned}
$$

where σ is the sigmoid function and the operation $\odot$ denotes element-wise product.

For a given input vector x and a configuration of (A, B), give the steps of the back-propagation algorithm that computes the gradients $\frac{\partial \mathcal{L}}{\partial A}$ and $\frac{\partial \mathcal{L}}{\partial B}$ evaluated at (x, A, B). Based on these steps, write a program in Python that finds

$$(\widehat{A}, \widehat{B}) := \arg\min_{(A,B)} \sum_{i=1}^{N} \mathcal{L}(x_i; A, B)$$

using gradient descent implemented via back-propagation, for any N points $x_1, x_2, \ldots, x_N$ in $\mathbb{R}^K$. **Note: You are not allowed to use the auto-differentiation libraries in your program. That is, you must implement back-propagation manually.**

Question 2 (15 points) *For a K-class classification problem, with $X \in \mathbb{R}^m$, suppose that there are two models $\mathcal{H}_1$ and $\mathcal{H}_2$, given below. Each member hypothesis in $\mathcal{H}_1$ and in $\mathcal{H}_2$ specfies a $p_{Y|X}$.*

$$
\begin{aligned}
\mathcal{H}_1 &:= \{\mathbf{softmax}(Wx) : W \in \mathbb{R}^{K\times m}\}\\
\mathcal{H}_2 &:= \{\mathbf{softmax}((A+B)Cx) : A \in \mathbb{R}^{K\times K}, B \in \mathbb{R}^{K\times K}, C \in \mathbb{R}^{K\times m}\}
\end{aligned}
$$

Prove that $\mathcal{H}_1 = \mathcal{H}_2$.

Question 3 (60 points) *MNIST dataset is a simple and popular dataset for image classification. You can download the dataset from `http://yann.lecun.com/exdb/mnist/` and get more information about the dataset. In this exercise, you will need to do the following.*

- *develop three classifiers for this dataset using three different models: soft-max regression, MLP, and CNN.*
- *For each model, investigated its behaviour with and without dropout.*
- *For each model, investigated its behaviour with and without batch normalization.*

You may freely explore any design freedom in each model (e.g., width/depth of MLP, kernel size/number of kernels/depth in CNN). You need to submit your code together with a report documenting your observations. In your report, you may feel free to include anything interesting you observe and remark on the lessons learned.