ML-As-4

Problem 1: Neural Networks (**20 pts**)

Consider a 3-layer fully connected neural network with the following architecture:

- Input layer: n = 4 neurons
- Hidden layer: m = 3 neurons using a custom activation function $f(x) = ReLU(x) + sin(x)$
- Output layer: k = 2 neurons using a softmax activation function $\sigma(z_i) = \frac{e_i^z}{\sum_j e^{z_j}}$

The network parameters (**weights and biases**) are given as:

- $W_1 \in \mathbb{R}^{3\times4}$ and $b_1 \in \mathbb{R}^3$ for the hidden layer.
- $W_2 \in \mathbb{R}^{2\times3}$ and $b_2 \in \mathbb{R}^2$ for the output layer.

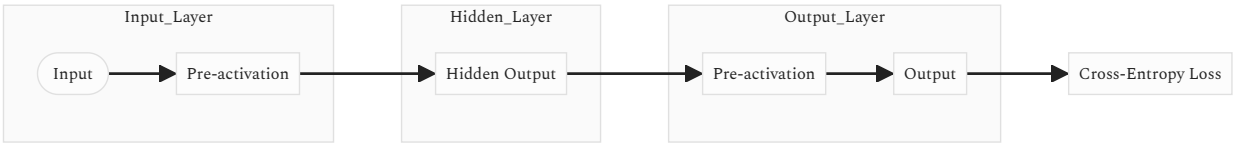
Given the input vector $x \in \mathbb{R}^4$ and target output $y \in \mathbb{R}^2$. Define the loss function as cross-entropy loss:

$$Loss = -\sum_{i=1}^{k} y_i \log(\hat{y}_i)$$

where $\hat{y}$ is the output after the softmax activation.

Q1

1. Derive the equations for the forward pass through the network, including both the hidden and output layers. (**3 pts**)



Forward pass through the network:

- Hidden layer:

$$z^{(1)} = W_1 x + b_1$$
$$h = f(z^{(1)}) = ReLU(z^{(1)}) + sin(z^{(1)})$$

- Output layer:

$$z^{(2)} = W_2 h + b_2$$
$$\hat{y} = \sigma(z^{(2)}) = \frac{e^{z^{(2)}}}{\sum_j e^{z_j^{(2)}}}$$

- Loss:

$$\hat{y}i = \frac{e^{z_i^{(2)}}}{\sum j = 1^2 e^{z_j^{(2)}}}$$

2. Calculate the outputs $\mathbf{Z_1}$, $\mathbf{H}$, $\mathbf{Z_2}$, and $\hat{\mathbf{y}}$ explicitly for a given input $\mathbf{x} = [1, -1, 0.5, 2]^T$ and the following initial weights and biases:

$$\mathbf{W}_1 = \begin{pmatrix} 0.1 & -0.2 & 0.3 & 0.4 \\ 0.5 & -0.3 & 0.1 & -0.2 \\ 0.4 & 0.2 & -0.5 & 0.3 \end{pmatrix}, \quad \mathbf{b}_1 = \begin{pmatrix} 0.1 \\ -0.1 \\ 0.05 \end{pmatrix}$$

$$\mathbf{W}_2 = \begin{pmatrix} -0.3 & 0.2 & 0.1 \\ 0.4 & -0.5 & 0.3 \end{pmatrix}, \quad \mathbf{b}_2 = \begin{pmatrix} 0.05 \\ -0.05 \end{pmatrix}$$

- Note that Z_1 is the net input to the hidden layer, H is the activation output of the hidden layer, and Z_2 is the net input to the output layer. (**3 pts**)

$$Z_1 = W_1 x + b_1 = \begin{pmatrix} 0.1 & -0.2 & 0.3 & 0.4 \\ 0.5 & -0.3 & 0.1 & -0.2 \\ 0.4 & 0.2 & -0.5 & 0.3 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 0.5 \\ 2 \end{pmatrix} + \begin{pmatrix} 0.1 \\ -0.1 \\ 0.05 \end{pmatrix} = \begin{pmatrix} 1.35 \\ 0.35 \\ 0.6 \end{pmatrix}$$

$$H = ReLU(Z_1) + sin(Z_1) = \begin{pmatrix} 2.325 \\ 0.692 \\ 1.164 \end{pmatrix}$$

$$Z_2 = W_2 H + b_2 = \begin{pmatrix} -0.3 & 0.2 & 0.1 \\ 0.4 & -0.5 & 0.3 \end{pmatrix} \begin{pmatrix} 2.33 \\ 0.69 \\ 1.64 \end{pmatrix} + \begin{pmatrix} 0.05 \\ -0.05 \end{pmatrix} = \begin{pmatrix} -0.3927 \\ 0.8832 \end{pmatrix}$$

$$\hat{y} = \sigma(Z_2) = \begin{pmatrix} 0.2181 \\ 0.7819 \end{pmatrix}$$

Q2

Derive the gradient of the loss with respect to each parameter (W_1, b_1, W_2, b_2) in the network and obtain the gradient values using results from the first question. Use matrix calculus to express the gradients.
Hint: You can first calculate the error terms δ_2 and δ_1 for each layer and use them to express the gradients. (**10 pts**)

Error term δ_2

$$\delta_2 = \hat{y} - y$$

Gradient of the loss with respect to W_2

$$\frac{\partial Loss}{\partial W_2} = \frac{\partial Loss}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial Z_2} \frac{\partial Z_2}{\partial W_2} = \delta_2 H^T$$

Gradient of the loss with respect to b_2

$$\frac{\partial Loss}{\partial b_2} = \frac{\partial Loss}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial Z_2} \frac{\partial Z_2}{\partial b_2} = \delta_2$$

Error term δ_1

$$\delta_1 = \frac{\partial Loss}{\partial H} \frac{\partial H}{\partial Z_1} = W_2^T \delta_2 \odot \frac{\partial H}{\partial Z_1}$$

Gradient of the loss with respect to W_1

$$\frac{\partial Loss}{\partial W_1} = \frac{\partial Loss}{\partial H} \frac{\partial H}{\partial Z_1} \frac{\partial Z_1}{\partial W_1} = \delta_1 x^T$$

Gradient of the loss with respect to b_1

$$\frac{\partial Loss}{\partial b_1} = \frac{\partial Loss}{\partial H} \frac{\partial H}{\partial Z_1} \frac{\partial Z_1}{\partial b_1} = \delta_1$$

Q3

Suppose the learning rate $\alpha = 0.001$. Please calculate the updated parameter values after one back propagation process. (**4 pts**)

The general update rule for each parameter θ with gradient descent is:

$$\theta = \theta - \alpha \frac{\partial \text{Loss}}{\partial \theta}$$

where α is the learning rate.

Updating W_2:

$$W_2^{\text{new}} = W_2 - \alpha \frac{\partial \text{Loss}}{\partial W_2} = W_2 - 0.001 \delta_2 H^T$$

Updating b_2:

$$b_2^{\text{new}} = b_2 - \alpha \frac{\partial \text{Loss}}{\partial b_2} = b_2 - 0.001 \delta_2$$

Updating W_1:

$$W_1^{\text{new}} = W_1 - \alpha \frac{\partial \text{Loss}}{\partial W_1} = W_1 - 0.001 \delta_1 x^T$$

Updating b_1:

$$b_1^{\text{new}} = b_1 - \alpha \frac{\partial \text{Loss}}{\partial b_1} = b_1 - 0.001 \delta_1$$