Problem 1

Consider the following multi-layer neural network (Figure 1) which includes an input layer, one hidden layer, and an output layer, containing d, n, q neurons respectively. The parameters between the input layer and the hidden layer are $\boldsymbol{W}_1 \in \mathbb{R}^{d\times n}$, $\boldsymbol{b}_1 \in \mathbb{R}^n$, and the parameters between the hidden layer and the output layer are $\boldsymbol{W}_2 \in \mathbb{R}^{n\times q}$, $\boldsymbol{b}_2 \in \mathbb{R}^q$. Where $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ are the weight matrices and $\boldsymbol{b}_1$, $\boldsymbol{b}_2$ are the bias vectors.

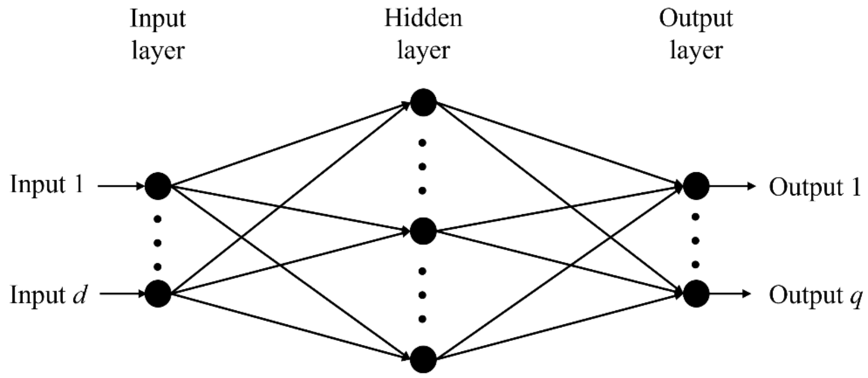

Figure 1: Neural Network

Let us first compute the forward propagation. Let $\boldsymbol{x} \in \mathbb{R}^d$ be the input. The hidden layer is computed as follows:

$$\boldsymbol{h} = \boldsymbol{W}_1^{\top}\boldsymbol{x} + \boldsymbol{b}_1 \in \mathbb{R}^n \tag{1-1}$$

Then the ReLU activation function is applied to Eq.(1-1):

$$\boldsymbol{a} = ReLU(\boldsymbol{h}) \in \mathbb{R}^n \tag{1-2}$$

The output layers' activation is obtained using the following transformation

$$\boldsymbol{z} = \boldsymbol{W}_2^{\top}\boldsymbol{a} + \boldsymbol{b}_2 \in \mathbb{R}^q \tag{1-3}$$

Finally, the soft-max function is applied to Eq.(1-3) to obtain the probability for each category.

$$\hat{\boldsymbol{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_q \end{bmatrix} = Softmax(\boldsymbol{z}) = \begin{bmatrix} Softmax(z_1) \\ Softmax(z_2) \\ \vdots \\ Softmax(z_q) \end{bmatrix} = \begin{bmatrix} \dfrac{\exp(z_1)}{\sum_{i=1}^{q}\exp(z_i)} \\ \dfrac{\exp(z_2)}{\sum_{i=1}^{q}\exp(z_i)} \\ \vdots \\ \dfrac{\exp(z_q)}{\sum_{i=1}^{q}\exp(z_i)} \end{bmatrix} \in \mathbb{R}^q \tag{1-4}$$

Here, $\hat{y}_i$, z_i is the i-th element of $\hat{\boldsymbol{y}}$, $\boldsymbol{z}$, $\hat{\boldsymbol{y}}$ is the predicted output by the feed-forward neural network.

Your task is to compute the derivatives of the Cross-Entropy loss function given in Eq.(1-5) with respect to $\boldsymbol{W}_1,\ \boldsymbol{W}_2,\ \boldsymbol{b}_1,\ \boldsymbol{b}_2$ by hand, i.e.,

$$\frac{\partial Loss}{\partial \boldsymbol{W}_1},\ \frac{\partial Loss}{\partial \boldsymbol{W}_2},\ \frac{\partial Loss}{\partial \boldsymbol{b}_1},\ \frac{\partial Loss}{\partial \boldsymbol{b}_2}.$$

$$Loss = -\sum_{i=1}^{q} y_i \ln(\hat{y}_i) \tag{1-5}$$

Here, $y_i \in \{0,\ 1\}$ is the ground-truth indicator, $\hat{y}_i$ is the i-th element of $\hat{\boldsymbol{y}}$.

Show all the intermediate derivative computation steps. You might benefit from making a rough schematic of the back-propagation process.