1. ANN contains three layers: input layer ($\mathrm{X}^{N\times 1}$), a hidden layer ($\mathrm{H}^{N_1\times 1}$), and an output layer ($\hat{\mathrm{Y}}^{K\times 1}$). Let W^1 and W^1_0 respectiely represent the weight matrix and weight bias vector for the hidden layer, and W^2 and W^2_o be the weight matrix and weight bias vector for the output layer respectively. Assuming the ReLU activation function for the hidden layer and softmax function for the ouput layer, derive the expression for $\frac{\partial\hat{\mathrm{Y}}}{\partial\mathrm{W}^1}$ and $\frac{\partial\hat{\mathrm{Y}}}{\partial\mathrm{W}^1_{i,j}}$, where $\mathrm{W}^1_{i,j}$ represents ith row and jth column element of W^1. Show the derivation process and intermediate results.

1. For $\frac{\partial\hat{\mathrm{Y}}}{\partial\mathrm{W}^1}$

Let $\mathrm{Z}^i \in \mathbb{R}^{N_i\times 1}$ the output of $(\mathrm{W}^i)^\top \mathrm{sth}$. sth can be either X or H.

The forward model:

1. $Z^1 = (\mathrm{W}^1)^\top X + W^1_0$, where $Z^1 \in \mathbb{R}^{N_1\times 1}, \mathrm{W}^1 \in \mathbb{R}^{N\times N_1}, \mathrm{W}^1_0 \in \mathbb{R}^{N_1\times 1}$
2. $H = \max(0, Z^1)$, where $H \in \mathbb{R}^{N_1\times 1}$
3. $Z^2 = (\mathrm{W}^2)^\top H + W^2_0$, where $Z^2 \in \mathbb{R}^{K\times 1}, \mathrm{W}^2 \in \mathbb{R}^{N_1\times K}, \mathrm{W}^2_0 \in \mathbb{R}^{K\times 1}$
4. $\hat{\mathrm{Y}} = \sigma_M(Z^2) = \begin{bmatrix} \exp(Z^2[1])/\sum_{k'=1}^{K}\exp(Z^2[k']) \\ \dots \\ \exp(Z^2[K])/\sum_{k'=1}^{K}\exp(Z^2[k']) \end{bmatrix}$, where $\hat{\mathrm{Y}} \in \mathbb{R}^{K\times 1}$

The backward process:

$\frac{\partial\hat{\mathrm{Y}}}{\partial\mathrm{W}^1} = \frac{\partial Z^1}{\partial W^1}\frac{\partial H}{\partial Z^1}\frac{\partial Z^2}{\partial H}\frac{\partial \hat{Y}}{\partial Z^2}$

1. $\frac{\partial \hat{Y}}{\partial Z^2}$

   This is a $K \times K$ matrix: $\begin{bmatrix} \frac{\partial \hat{Y}[1]}{\partial Z^2} & \dots & \frac{\partial \hat{Y}[K]}{\partial Z^2} \end{bmatrix}^{1 \times K}$

   For $\frac{\partial \hat{Y}[i]}{\partial Z^2} = \begin{bmatrix} \frac{\partial \hat{Y}[i]}{\partial Z^2[1]} \\ \dots \\ \frac{\partial \hat{Y}[i]}{\partial Z^2[K]} \end{bmatrix}^{K \times 1}$

   For $\frac{\partial \hat{Y}[i]}{\partial Z^2[j]}$, if $i = j:\ \frac{\partial \hat{Y}[i]}{\partial Z^2[j]} = \sigma_M(z[i])(1 - \sigma_M(z[i]))$, else: $-\sigma_M(z[i])\sigma_M(z[j])$

   So the $K \times K$ derivative matrix is：

$$\begin{bmatrix} \sigma_M(Z^2[1])(1 - \sigma_M(Z^2[1])) & -\sigma_M(Z^2[1])\sigma_M(Z^2[2])) & \dots & -\sigma_M(Z^2[K])\sigma_M(Z^2[K])) \\ -\sigma_M(Z^2[2])\sigma_M(Z^2[1])) & \sigma_M(Z^2[2])(1 - \sigma_M(Z^2[2])) & \dots & -\sigma_M(Z^2[K])\sigma_M(Z^2[K])) \\ \dots & & & \\ -\sigma_M(Z^2[K])\sigma_M(Z^2[1])) & -\sigma_M(Z^2[K])\sigma_M(Z^2[2])) & \dots & \sigma_M(Z^2[K])(1 - \sigma_M(Z^2[K])) \end{bmatrix}^{K \times K}$$

1. $\frac{\partial Z^2}{\partial H}$

   This is a $N_1 \times K$ matrix: W^2

1. $\frac{\partial H}{\partial Z^1}$

   This is a $N_1 \times 1$ vector: $\begin{bmatrix} \frac{\partial H[1]}{\partial Z^1[1]} \\ \dots \\ \frac{\partial H[N_1]}{\partial Z^1[N_1]} \end{bmatrix}^{N_1 \times 1}$

   For each $\frac{\partial H[i]}{\partial Z^1[i]}$, if $Z^i > 0$, $\frac{\partial H[i]}{\partial Z^1[i]} = 1$, else $\frac{\partial H[i]}{\partial Z^1[i]} = 0$

1. $\frac{\partial Z^1}{\partial W^1}$

   This is a $N \times N_1 \times N_1$ tensor: $\begin{bmatrix} \frac{\partial Z^1[1]}{\partial W^1} & \dots & \frac{\partial Z^1[N_1]}{\partial W^1} \end{bmatrix}^{1 \times N_1}$

   At $\frac{\partial Z^1[i]}{\partial W^1} = \begin{bmatrix} \frac{\partial Z^1[i]}{\partial W^1[1]} & \dots & \frac{\partial Z^1[i]}{\partial W^1[N_1]} \end{bmatrix}^{1 \times N_1}$

   For $\frac{\partial Z^1[i]}{\partial W^1[j]}$ if i = j, $\frac{\partial Z^1[i]}{\partial W^1[j]} = X^{N \times 1}$, else $\frac{\partial Z^1[i]}{\partial W^1[j]} = 0^{N \times 1}$

   So the $N \times N_1 \times N_1$ tensor is:

$$\Big[\begin{bmatrix} X^{N \times 1} & 0^{N \times 1} & \dots & 0^{N \times 1} \end{bmatrix}^{N \times N_1}, \begin{bmatrix} 0^{N \times 1} & X^{N \times 1} & \dots & 0^{N \times 1} \end{bmatrix}^{N \times N_1}, \begin{bmatrix} 0^{N \times 1} & 0^{N \times 1} & \dots & X^{N \times 1} \end{bmatrix}^{N \times N_1} \Big]^{N \times N_1 \times N_1}$$

To multiply them:

$$\frac{\partial \hat{Y}}{\partial W^1} = \frac{\partial Z^1}{\partial W^1}\frac{\partial H}{\partial Z^1}W^2\frac{\partial \hat{Y}}{\partial Z^2}$$

The dim is:

$$\begin{aligned}&(N \times N_1 \times N_1) \times (N_1 \times 1)(N_1 \times K)(K \times K)\\=&(N \times N_1)(N_1 \times K)(K \times K)\\=&N \times K\end{aligned}$$

2. For $\frac{\partial \hat{\mathrm{Y}}}{\partial \mathrm{W}^1_{i,j}}$

Forward model

1. $z^1[j] = w^1[i][j]x[i] + w^1_0[j]$, where $z^1[j] \in \mathbb{R}$
2. $h^1[j] = \max(0, z^1[j])$, where $h^1[j] \in \mathbb{R}$
3. $Z^2 = w^2[j][]h^1[j] + w^2_0[]$, where $Z^2_k \in \mathbb{R}^{K\times 1}$
4. $\hat{Y}[k] = \sigma_M(Z^2)$

The backward process:

because this W^1_{ij} is a scalar, as we mentioned in the class, we should change the sequence of chain rule

$\frac{\partial \hat{\mathrm{Y}}}{\partial \mathrm{W}^1_{i,j}} = \frac{\partial \hat{Y}}{\partial Z^2}\frac{\partial Z^2}{\partial H_j}\frac{\partial H_j}{\partial Z^1_j}\frac{\partial Z^1_j}{\partial W^1_{i,j}}$

1. $\frac{\partial \hat{Y}}{\partial Z^2}$ the same with what we mentioned, $K \times K$

1. $\frac{\partial Z^2}{\partial H_j}$ is a $K \times 1$ vector $(W^2_j)^\top$ which is the transpose of row of matrix W^2

1. $\frac{\partial H_j}{\partial Z^1_j}$, if $Z_j > 0$, $\frac{\partial H_j}{\partial Z^1_j} = 1$, else $\frac{\partial H_j}{\partial Z^1_j} = 0$

1. $\frac{\partial Z^1_j}{\partial W^1_{i,j}} = x_i$

To multiply them:

$\frac{\partial \hat{\mathrm{Y}}}{\partial \mathrm{W}^1_{i,j}} = \frac{\partial \hat{Y}}{\partial Z^2}(W^2_j)^\top\frac{\partial H_j}{\partial Z^1_j}x_i$

The dim: $(K \times K) \times (K \times 1) \times 1 \times 1 = (K \times 1)$

This is consistent with our vector by scalar derivative convention.

Problem 2

The structure of a Neural Network is given below The input value $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and the desired output value is $Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. The initial weight matrix for the first layer W^1:

$$W^1 = \begin{bmatrix} W_1^1 & W_2^1 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 4 & 5 \end{bmatrix}$$

The bias weight matrix $W_0^1 = \begin{bmatrix} 1 \\ -6 \end{bmatrix}$. The initial weight matrix for the second layer W^2:

$$W^2 = \begin{bmatrix} W_1^2 & W_2^2 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 4 & 3 \end{bmatrix}$$

The bias weight matrix $W_0^2 = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$

Perform the following tasks:

Forward propagation: calculate the value of the hidden nodes using sigmoid function as the activation function and obtain the predicted value for output nodes $\hat{Y}$ using softmax as the ouput function. Compute the gradient of the output $\bigtriangledown \hat{Y}$ using cross-entropy loss function.

Forward propagation:

1. $Z^1 = (\mathrm{W}^1)^\top X + W_0^1$, where $Z^1 \in \mathbb{R}^{N_1 \times 1}, \mathrm{W}^1 \in \mathbb{R}^{N \times N_1}, \mathrm{W}_0^1 \in \mathbb{R}^{N_1 \times 1}$

   $$Z^1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix} + \begin{bmatrix} 1 \\ -6 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$$

2. $H = \sigma(Z^1) = \exp(Z^1)/(1 + \exp(Z^1))$, where $H \in \mathbb{R}^{N_1 \times 1}$

   $$H = \begin{bmatrix} \exp(4)/(1 + \exp(4)) \\ \exp(0)/(1 + \exp(0) \end{bmatrix} = \begin{bmatrix} .982 \\ .500 \end{bmatrix}$$

3. $Z^2 = (\mathrm{W}^2)^\top H + W_0^2$, where $Z^2 \in \mathbb{R}^{K \times 1}, \mathrm{W}^2 \in \mathbb{R}^{N_1 \times K}, \mathrm{W}_0^2 \in \mathbb{R}^{K \times 1}$

   $$Z^2 = \begin{bmatrix} 3.964 \\ 4.446 \end{bmatrix} + \begin{bmatrix} -1 \\ -2 \end{bmatrix} = \begin{bmatrix} 2.964 \\ 2.446 \end{bmatrix}$$

1. $\hat{\mathrm{Y}} = \sigma_M(Z^2) = \begin{bmatrix} \exp(Z^2[1]) / \sum_{k'=1}^K \exp(Z^2[k']) \\ \dots \\ \exp(Z^2[K]) / \sum_{k'=1}^K \exp(Z^2[k']) \end{bmatrix}$, where $\hat{\mathrm{Y}} \in \mathbb{R}^{K \times 1}$

   $$\hat{\mathrm{Y}} = \begin{bmatrix} .627 \\ .373 \end{bmatrix}$$

2. $\mathcal{L}(Y, \hat{Y}) = -Y^\top \log \hat{Y}$

   $$\mathcal{L}(Y, \hat{Y}) = .985$$

Gradient of $\hat{Y}$: $\bigtriangledown\hat{Y}$

$$\bigtriangledown\hat{Y} = -\frac{\partial\mathcal{L}(Y,\hat{Y})}{\partial\hat{Y}} = -\frac{Y}{\hat{Y}} = \begin{bmatrix} 0 \\ -2.679 \end{bmatrix}$$

Backpropagation:

Gradient of H: $\bigtriangledown H$

$$\bigtriangledown H = \frac{\partial Z^2}{\partial H}\frac{\partial\sigma(Z^2)}{\partial Z^2}\bigtriangledown\hat{Y} = W^2\begin{bmatrix} \hat{Y}[1](1-\hat{Y}[1]) & -\hat{Y}[1]\hat{Y}[2] \\ -\hat{Y}[2]\hat{Y}[1] & \hat{Y}[2](1-\hat{Y}[2]) \end{bmatrix}\begin{bmatrix} -Y[1]/\hat{Y}[1] \\ -Y[2]/\hat{Y}[2] \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 4 & 3 \end{bmatrix}\begin{bmatrix} .627 \\ -.627 \end{bmatrix}$$
$$= \begin{bmatrix} -.627 \\ .627 \end{bmatrix}$$

Gradient of $\bigtriangledown W_0^2$

$$\bigtriangledown W_0^2 = \frac{\partial Z^2}{\partial W_0^2}\frac{\partial\sigma(Z^2)}{\partial Z^2}\bigtriangledown\hat{Y} = 1*\begin{bmatrix} \hat{Y}[1](1-\hat{Y}[1]) & -\hat{Y}[1]\hat{Y}[2] \\ -\hat{Y}[2]\hat{Y}[1] & \hat{Y}[2](1-\hat{Y}[2]) \end{bmatrix}\begin{bmatrix} -Y[1]/\hat{Y}[1] \\ -Y[2]/\hat{Y}[2] \end{bmatrix} = \begin{bmatrix} .627 \\ -.627 \end{bmatrix}$$

Gradient of $\bigtriangledown W^2$

$$\bigtriangledown W^2 = \frac{\partial Z^2}{\partial W^2}\frac{\partial\sigma(Z^2)}{\partial Z^2}\bigtriangledown\hat{Y} = H(\begin{bmatrix} \hat{Y}[1](1-\hat{Y}[1]) & -\hat{Y}[1]\hat{Y}[2] \\ -\hat{Y}[2]\hat{Y}[1] & \hat{Y}[2](1-\hat{Y}[2]) \end{bmatrix}\begin{bmatrix} -Y[1]/\hat{Y}[1] \\ -Y[2]/\hat{Y}[2] \end{bmatrix})^\top$$
$$= \begin{bmatrix} .982 \\ .5 \end{bmatrix}\begin{bmatrix} .627 & -.627 \end{bmatrix} = \begin{bmatrix} .616 & -.616 \\ .314 & -.314 \end{bmatrix}$$

Gradient of $\bigtriangledown W_0^1$

$$\bigtriangledown W_0^1 = \frac{\partial Z^1}{\partial W_0^1}\frac{\partial\sigma(Z^1)}{\partial Z^1}\bigtriangledown H = 1*\begin{bmatrix} H[1](1-H[1]) \\ H[2](1-H[2]) \end{bmatrix}\bigtriangledown H = \begin{bmatrix} -.011 \\ .157 \end{bmatrix}$$

Gradient of $\bigtriangledown W^1$

$$\bigtriangledown W^1 = \frac{\partial Z^1}{\partial W^1}\frac{\partial\sigma(Z^1)}{\partial Z^1}\bigtriangledown H = X(\begin{bmatrix} H[1](1-H[1]) \\ H[2](1-H[2]) \end{bmatrix}\bigtriangledown H)^\top = \begin{bmatrix} 1 \\ 0 \end{bmatrix}\begin{bmatrix} -.627 & .627 \end{bmatrix} = \begin{bmatrix} -.011 & .157 \\ 0 & 0 \end{bmatrix}$$

Update

$$W^1 = W^1 - .5*\bigtriangledown W^1 = \begin{bmatrix} 3 & 6 \\ 4 & 5 \end{bmatrix} - .5*\begin{bmatrix} -.627 & .627 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 3.006 & 5.922 \\ 4 & 5 \end{bmatrix}$$

$$W_0^1 = W_0^1 - .5*\bigtriangledown W_0^1 = \begin{bmatrix} 1 \\ -6 \end{bmatrix} - .5*\begin{bmatrix} -.011 \\ .157 \end{bmatrix} = \begin{bmatrix} 1.006 \\ -6.079 \end{bmatrix}$$

$$W^2 = W^2 - .5 * \bigtriangledown W^2 = \begin{bmatrix} 2 & 3 \\ 4 & 3 \end{bmatrix} - .5 * \begin{bmatrix} .616 & -.616 \\ .314 & -.314 \end{bmatrix} = \begin{bmatrix} 1.692 & 3.308 \\ 3.843 & 3.157 \end{bmatrix}$$

$$W_0^2 = W_0^2 - .5 * \bigtriangledown W_0^2 = \begin{bmatrix} -1 \\ -2 \end{bmatrix} - .5 * \begin{bmatrix} .627 \\ -.627 \end{bmatrix} = \begin{bmatrix} -1.627 \\ -1.373 \end{bmatrix}$$

New Loss

$\mathcal{L}_{\text{new}} = .1 < .985$ This means the gradient work.