HW 3 Solution

1 Problem 1 (20 points)

We need to apply the backward propagation from the output layer toward the input layer. From the dimensions given in the problem, we note that $\boldsymbol{W}^1 \in \mathbb{R}^{N\times N_1}$, $\boldsymbol{W}_0^1 \in \mathbb{R}^{N_1}$, $\boldsymbol{W}^2 \in \mathbb{R}^{N_1\times K}$, $\boldsymbol{W}_0^2 \in \mathbb{R}^{K}$, and $\hat{\boldsymbol{y}} \in \mathbb{R}^{K}$. Since the loss function is not given, I consider a general cost function defined as $\ell(\boldsymbol{y}, \hat{\boldsymbol{y}})$. For the output layer we have

$$\nabla \boldsymbol{W}^2 = \frac{\partial \ell(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\partial \boldsymbol{W}^2} = \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{W}^2}\frac{\partial \ell(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\partial \hat{\boldsymbol{y}}} = \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{W}^2}\nabla \hat{\boldsymbol{y}} = \sum_{k=1}^{K} \frac{\partial \hat{\boldsymbol{y}}[k]}{\partial \boldsymbol{W}^2}\nabla \hat{\boldsymbol{y}}[k] \ . \tag{1}$$

We know that $\hat{\boldsymbol{y}}[k] = \sigma_K((\boldsymbol{W}_k^2)^T + \boldsymbol{W}_0^2)$ where $\boldsymbol{W}_k^2$ is the k-th column of $\boldsymbol{W}_k^2$, and also

$$\frac{\partial \hat{\boldsymbol{y}}[k]}{\partial \boldsymbol{W}^2} = \left[\frac{\partial \hat{\boldsymbol{y}}[k]}{\partial \boldsymbol{W}_1^2}, \frac{\partial \hat{\boldsymbol{y}}[k]}{\partial \boldsymbol{W}_2^2}, \ldots, \frac{\partial \hat{\boldsymbol{y}}[k]}{\partial \boldsymbol{W}_K^2}\right] \ . \tag{2}$$

On the other hand, from the previous homework, we have the partial derivative of softmax function. Therefore,

$$\frac{\partial \hat{\boldsymbol{y}}[k]}{\partial \boldsymbol{W}_i^2} = \begin{cases} \hat{\boldsymbol{y}}[k](1-\hat{\boldsymbol{y}}[k])\boldsymbol{H} & \text{if } i = k \\ -\hat{\boldsymbol{y}}[k]\hat{\boldsymbol{y}}[i]\boldsymbol{H} & \text{if } i \neq k \end{cases} \ . \tag{3}$$

For the bias vector $\boldsymbol{W}_0^2$ the same procedure should be performed where it can be observed the the only difference in that the bias is not multiplied by $\boldsymbol{H}$ and therefore $\boldsymbol{H}$ will not appear in the derivatives. Hence,

$$\nabla \boldsymbol{W}_0^2 = \sum_{k=1}^{K} \frac{\partial \hat{\boldsymbol{y}}[k]}{\partial \boldsymbol{W}_0^2}\nabla \hat{\boldsymbol{y}}[k] = \sum_{k=1}^{K} \left[\frac{\partial \hat{\boldsymbol{y}}[k]}{\partial \boldsymbol{W}_0^2[1]}, \frac{\partial \hat{\boldsymbol{y}}[k]}{\partial \boldsymbol{W}_0^2[2]}, \ldots, \frac{\partial \hat{\boldsymbol{y}}[k]}{\partial \boldsymbol{W}_0^2[K]}\right] \nabla \hat{\boldsymbol{y}}[k] \ , \tag{4}$$

where

$$\frac{\partial \hat{\boldsymbol{y}}[k]}{\partial \boldsymbol{W}_0^2[i]} = \begin{cases} \hat{\boldsymbol{y}}[k](1-\hat{\boldsymbol{y}}[k]) & \text{if } i=k \\ -\hat{\boldsymbol{y}}[k]\hat{\boldsymbol{y}}[i] & \text{if } i \neq k \end{cases} . \tag{5}$$

In order to calculate the partial derivatives for the weights of the hidden layer, first we need to calculate $\nabla \boldsymbol{H}$.

$$\nabla \boldsymbol{H} = \sum_{k=1}^{K} \frac{\partial \hat{\boldsymbol{y}}[k]}{\partial \boldsymbol{H}} \nabla \hat{\boldsymbol{y}}[k] = \sum_{k=1}^{K} \Big[\boldsymbol{W}_k^2 \hat{\boldsymbol{y}}[k] - \sum_{j=1}^{K} \boldsymbol{W}_j^2 \hat{\boldsymbol{y}}[k]\hat{\boldsymbol{y}}[j] \Big] \nabla \hat{\boldsymbol{y}}[k] . \tag{6}$$

By applying the chain rule, we have

$$\nabla \boldsymbol{W}^1 = \frac{\partial \boldsymbol{H}}{\partial \boldsymbol{W}^1} \nabla \boldsymbol{H} = \sum_{k=1}^{K} \frac{\partial \boldsymbol{H}[k]}{\partial \boldsymbol{W}^1} \nabla \boldsymbol{H}[k] \tag{7}$$

$$= \sum_{k=1}^{K} \Big[\frac{\partial \boldsymbol{H}[k]}{\partial \boldsymbol{W}_i^1}, \frac{\partial \boldsymbol{H}[k]}{\partial \boldsymbol{W}_2^1} \cdots, \frac{\partial \boldsymbol{H}[k]}{\partial \boldsymbol{W}_{N_1}^1} \Big] \nabla \boldsymbol{H}[k] , \tag{8}$$

where

$$\frac{\partial \mathrm{ReLU}((\boldsymbol{W}_k^1)^T \boldsymbol{X} + \boldsymbol{W}_0^1)}{\partial \boldsymbol{W}_i^1} = \begin{cases} \boldsymbol{X}[k] & \text{if } i=k \text{ and } (\boldsymbol{W}_i^1)^T \boldsymbol{X} + \boldsymbol{W}_0^1[i] > 0 \\ 0 & \text{Otherwise} \end{cases} . \tag{9}$$

For the bias term, we have

$$\nabla \boldsymbol{W}_0^1 = \sum_{k=1}^{K} \Big[\frac{\partial \boldsymbol{H}[k]}{\partial \boldsymbol{W}_0^1[1]}, \frac{\partial \boldsymbol{H}[k]}{\partial \boldsymbol{W}_0^1[2]} \cdots, \frac{\partial \boldsymbol{H}[k]}{\partial \boldsymbol{W}_0^1[N_1]} \Big]^T \nabla \boldsymbol{H}[k] , \tag{10}$$

where

$$\frac{\partial \mathrm{ReLU}((\boldsymbol{W}_k^1)^T \boldsymbol{X} + \boldsymbol{W}_0^1)}{\partial \boldsymbol{W}_0^1[i]} = \begin{cases} 1 & \text{if } i=k \text{ and } (\boldsymbol{W}_i^1)^T \boldsymbol{X} + \boldsymbol{W}_0^1[i] > 0 \\ 0 & \text{Otherwise} \end{cases} . \tag{11}$$

Correctly calculating each of the gradient of $\boldsymbol{W}^1$ and $\boldsymbol{W}^2$ has **5 points**, each of the bias terms will earn you **3 points**, and you will get **4 points** for correctly using the chain rule and having the right dimensions for the gradients.

2 Problem 2 (30 points)

2.1

Given the input value $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, and the initialization of the weight matrix for the first layer W^1:

$$W^1 = \begin{bmatrix} 3 & 6 \\ 4 & 5 \end{bmatrix}; W_0^1 = \begin{bmatrix} 1 \\ -6 \end{bmatrix}. \tag{12}$$

The value of the hidden nodes can be calculated as:

$$\boldsymbol{H} = \phi((W^1)^t \boldsymbol{x} + W_0^1) \tag{13}$$

while $\phi(z) = \frac{1}{1+\exp(-z)}$.

Thus, the value of the hidden nodes are:

$$\begin{aligned} \boldsymbol{H} &= \phi(\begin{bmatrix} 3 & 4 \\ 6 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ -6 \end{bmatrix}) \\ &= \begin{bmatrix} 0.982 \\ 0.5 \end{bmatrix} \end{aligned} \tag{14}$$

.

Using multi-class sigmoid function as the output function, the predicted output value $\hat{\boldsymbol{y}}$ can be calculated as:

$$\hat{\boldsymbol{y}} = \sigma_M((W^2)^t \boldsymbol{H} + W_0^2) \tag{15}$$

while $\sigma_M(z_k) = \frac{\exp((W_k^2)^T \boldsymbol{H} + W_{0,k}^2)}{\sum_k \exp((W_k^2)^T \boldsymbol{H} + W_{0,k}^2)}$.

Thus, the value of the predicted output nodes are:

$$\begin{aligned} \hat{\boldsymbol{y}} &= \sigma_M(\begin{bmatrix} 2 & 4 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} 0.982 \\ 0.5 \end{bmatrix} + \begin{bmatrix} -1 \\ -2 \end{bmatrix}) \\ &= \begin{bmatrix} 0.6267 \\ 0.3733 \end{bmatrix} \end{aligned} \tag{16}$$

.

When loss function is chosen as squared loss function, the gradient of the output $\nabla \hat{\boldsymbol{y}}$ is:

$$\begin{aligned} \nabla \hat{\boldsymbol{y}} &= \frac{\partial l(\boldsymbol{y}[m], \hat{\boldsymbol{y}}[m])}{\partial \hat{\boldsymbol{y}}[m]} \\ &= \frac{1}{2} \frac{\partial (\boldsymbol{y}[m] - \hat{\boldsymbol{y}}[m])^t (\boldsymbol{y}[m] - \hat{\boldsymbol{y}}[m])}{\partial \hat{\boldsymbol{y}}[m]} \\ &= -(\boldsymbol{y}[m] - \hat{\boldsymbol{y}}[m]) \end{aligned} \tag{17}$$

which is $-\begin{bmatrix} 0-0.6267 \\ 1-0.3733 \end{bmatrix} = \begin{bmatrix} 0.6267 \\ -0.6267 \end{bmatrix}$ in this case.

2.2

Output Layer For the output layer, the gradient of the weight matrix can be calculated as:

$$\begin{aligned} \nabla W^2 &= \frac{\partial \hat{\boldsymbol{y}}}{\partial W^2} \nabla \boldsymbol{y} \\ &= \sum_{i=1}^{2} \frac{\partial \hat{y}[i]}{\partial W^2} \nabla y[i] \\ &= \sum_{i=1}^{2} \begin{bmatrix} \frac{\partial \hat{y}[i]}{\partial W_1^2} & \frac{\partial \hat{y}[i]}{\partial W_2^2} \end{bmatrix} \nabla y[i] \end{aligned} \tag{18}$$

Thus, the gradient of W^2 is:

$$\begin{aligned} &\nabla W^2 = \begin{bmatrix} \hat{y}[1](1-\hat{y}[1])\boldsymbol{H} & -\hat{y}[1]\hat{y}[2]\boldsymbol{H} \end{bmatrix} \nabla \hat{y}[1] + \begin{bmatrix} -\hat{y}[1]\hat{y}[2]\boldsymbol{H} & \hat{y}[2](1-\hat{y}[2])\boldsymbol{H} \end{bmatrix} \nabla y[2] \\ &= \begin{bmatrix} 0.6267 * 0.3733 \begin{bmatrix} 0.982 \\ 0.5 \end{bmatrix} & -0.6267 * 0.3733 \begin{bmatrix} 0.982 \\ 0.5 \end{bmatrix} \end{bmatrix} * 0.6267 \\ &+ \begin{bmatrix} -0.6267 * 0.3733 \begin{bmatrix} 0.982 \\ 0.5 \end{bmatrix} & 0.6267 * 0.3733 \begin{bmatrix} 0.982 \\ 0.5 \end{bmatrix} \end{bmatrix} * (-0.6267) \\ &= \begin{bmatrix} 0.2880 & -0.2880 \\ 0.1466 & -0.1466 \end{bmatrix} \end{aligned} \tag{19}$$

The gradient of the weight bias vector:

$$\nabla W_0^2 = \begin{bmatrix} 0.2932 \\ -0.2932 \end{bmatrix} \tag{20}$$

So, the gradient of the Hidden nodes $\nabla\boldsymbol{H}$ is:

$$\begin{aligned}
\nabla\boldsymbol{H} &= \frac{\partial\sigma_M(\boldsymbol{z})}{\partial\boldsymbol{H}}\nabla\boldsymbol{y} \\
&= \sum_{i=1}^{2}\frac{\partial\hat{y}[i]}{\partial\boldsymbol{H}}\nabla y[i] \\
&= \hat{y}[1][W_1^2 - \sum_{j=1}^{2}\hat{y}[j]W_j^2]\nabla y[1] + \hat{y}[2][W_2^2 - \sum_{j=1}^{2}\hat{y}[j]W_j^2]\nabla y[2] \\
&= 0.6267 * [\begin{bmatrix}2\\4\end{bmatrix} - (0.6267 * \begin{bmatrix}2\\4\end{bmatrix} + 0.3733 * \begin{bmatrix}3\\3\end{bmatrix})] \\
&+ 0.3733 * [\begin{bmatrix}3\\3\end{bmatrix} - (0.6267 * \begin{bmatrix}2\\4\end{bmatrix} + 0.3733 * \begin{bmatrix}3\\3\end{bmatrix})] \\
&= \begin{bmatrix}-0.2932\\0.2932\end{bmatrix}
\end{aligned} \tag{21}$$

Hidden Layer Given $\nabla\boldsymbol{H}$, the gradient of the first weight matrix is:

$$\begin{aligned}
\nabla W^1 &= \frac{\partial\boldsymbol{H}}{\partial W^1}\nabla\boldsymbol{H} \\
&= \sum_{i=1}^{2}\frac{\partial H[i]}{\partial W^1}\nabla H[i] \\
&= \sum_{i=1}^{2}\begin{bmatrix}\frac{\partial H[i]}{\partial W_1^1} & \frac{\partial H[i]}{\partial W_2^1}\end{bmatrix}\nabla H[i] \\
&= \begin{bmatrix}H[1](1-H[1])\begin{bmatrix}1\\0\end{bmatrix} & 0\end{bmatrix} * (-0.2932) + \begin{bmatrix}0 & H[2](1-H[2])\begin{bmatrix}1\\0\end{bmatrix}\end{bmatrix} * 0.2932 \\
&= \begin{bmatrix}-0.0052 & 0.0733\\0 & 0\end{bmatrix}
\end{aligned} \tag{22}$$

The gradient of the bias weight vector of the first layer is:

$$\nabla W_0^1 = \begin{bmatrix}-0.0052\\0.0733\end{bmatrix} \tag{23}$$

2.3

Weight Update The weight update equation is:

$$W = W - r * \nabla W \tag{24}$$

Update the weight matrix accordingly:

$$
\begin{aligned}
W^2_{new} &= W^2 - r * \nabla W^2 \\
&= \begin{bmatrix} 2 & 3 \\ 4 & 3 \end{bmatrix} - 0.5 * \begin{bmatrix} 0.2880 & -0.2880 \\ 0.1466 & -0.1466 \end{bmatrix} \\
&= \begin{bmatrix} 1.8560 & 3.1440 \\ 3.9267 & 3.0733 \end{bmatrix}
\end{aligned} \tag{25}
$$

The updated weight bias vector is:

$$
\begin{aligned}
W^2_{0,new} &= W^2_0 - r * \nabla W^2_0 \\
&= \begin{bmatrix} -1 \\ -2 \end{bmatrix} - 0.5 * \begin{bmatrix} 0.2932 \\ -0.2932 \end{bmatrix} \\
&= \begin{bmatrix} -1.1466 \\ -1.8534 \end{bmatrix}
\end{aligned} \tag{26}
$$

The updated weight matrix for the first layer is:

$$
\begin{aligned}
W^1_{new} &= W^1 - r * \nabla W^1 \\
&= \begin{bmatrix} 3 & 6 \\ 4 & 5 \end{bmatrix} - 0.5 * \begin{bmatrix} -0.0052 & 0.0733 \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} 3.0026 & 5.9633 \\ 4 & 5 \end{bmatrix}
\end{aligned} \tag{27}
$$

Weight bias vector of the first layer:

$$
\begin{aligned}
W^1_{0,new} &= W^1_0 - r * \nabla W^1_0 \\
&= \begin{bmatrix} 1 \\ -6 \end{bmatrix} - 0.5 * \begin{bmatrix} -0.0052 \\ 0.0733 \end{bmatrix} \\
&= \begin{bmatrix} 1.0026 \\ -6.0367 \end{bmatrix}
\end{aligned} \tag{28}
$$

Update Output Value

Case 1 After the updating of the weight matrix of the first layer, the value of hidden nodes are:

$$
\begin{aligned}
\boldsymbol{H} &= \phi((W^1_{new})^t \boldsymbol{x} + W^1_{0,new}) \\
&= \begin{bmatrix} 0.9821 \\ 0.4817 \end{bmatrix}
\end{aligned} \tag{29}
$$

.

The updated predicted output value $\hat{\boldsymbol{y}}$ is:

$$\begin{aligned}\hat{\boldsymbol{y}} &= \sigma_M((W^2_{new})^t \boldsymbol{H} + W^2_{0,new}) \\ &= \begin{bmatrix} 0.4633 \\ 0.5367 \end{bmatrix}\end{aligned} \tag{30}$$

.

The loss function value before updating is:

$$\begin{aligned}&\frac{1}{2}(\boldsymbol{y}[m] - \hat{\boldsymbol{y}}[m])^t(\boldsymbol{y}[m] - \hat{\boldsymbol{y}}[m]) \\ &= \frac{1}{2}\sum_{i=1}^{2} \nabla y[i]^2 \\ &= 0.3927\end{aligned} \tag{31}$$

.

After updating, the loss function value is:

$$\begin{aligned}&\frac{1}{2}\sum_{i=1}^{2} \nabla y[i]^2 \\ &= 0.2146\end{aligned} \tag{32}$$

It is clear that the loss function value has been reduced.

Case 2 If the weight matrix of the first layer is fixed, the value of the hidden nodes are also fixed, the updated predicted output value should be:

$$\begin{aligned}\hat{\boldsymbol{y}} &= \sigma_M((W^2_{new})^t \boldsymbol{H} + W^2_{0,new}) \\ &= \begin{bmatrix} 0.4672 \\ 0.5328 \end{bmatrix}\end{aligned} \tag{33}$$

.

From Case 1, we know the loss function value before updating is 0.3927, After updating weight matrix in the second layer, the loss function value has been changed to:

$$\begin{aligned}&\frac{1}{2}\sum_{i=1}^{2} \nabla y[i]^2 \\ &= 0.2183\end{aligned} \tag{34}$$

Also, the loss function value has been reduced.