# Gradient Derivation for Neural Network

## 1 1. Mean Squared Error (MSE) Loss Function

The loss function used is the Mean Squared Error (MSE):

$$L(Y, \hat{Y}) = \frac{1}{2N} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$

where:

- $N$ is the number of training samples.
- $Y_i$ is the actual output (true label).
- $\hat{Y}_i$ is the predicted output.

## 2 2. Gradient for $W^{(2)}$ (Output Layer Weights)

Using the chain rule, the gradient of the loss function with respect to $W^{(2)}$ is:

$$\frac{\partial L}{\partial W_{k,l}^{(2)}} = \sum_{i=1}^{N} \frac{\partial L}{\partial \hat{Y}_i} \times \frac{\partial \hat{Y}_i}{\partial O_i} \times \frac{\partial O_i}{\partial W_{k,l}^{(2)}}$$

Breaking it down:

1. Derivative of Loss w.r.t. $\hat{Y}$:

$$\frac{\partial L}{\partial \hat{Y}_i} = -\frac{1}{N}(Y_i - \hat{Y}_i)$$

2. Derivative of $\hat{Y}$ w.r.t. $O$ (Sigmoid Activation Derivative):

$$\frac{\partial \hat{Y}_i}{\partial O_i} = \hat{Y}_i(1 - \hat{Y}_i)$$

3. Derivative of $O$ w.r.t. $W^{(2)}$:

$$\frac{\partial O_i}{\partial W_{k,l}^{(2)}} = Z_{i,l}$$

Thus, the full gradient for $W^{(2)}$ including $\frac{1}{N}$ is:

$$\frac{\partial L}{\partial W_j^{(2)}} = -\frac{1}{N}\sum_{i=1}^{N} Z_{i,j} \times (Y_i - \hat{Y}_i) \times \hat{Y}_i \times (1 - \hat{Y}_i)$$

# 3 3. Gradient for $W^{(1)}$ (Hidden Layer Weights)

Applying the chain rule again:

$$\frac{\partial L}{\partial W_{k,l}^{(1)}} = \sum_{i=1}^{N} \frac{\partial L}{\partial \hat{Y}_i} \times \frac{\partial \hat{Y}_i}{\partial O_i} \times \frac{\partial O_i}{\partial Z_{i,l}} \times \frac{\partial Z_{i,l}}{\partial H_{i,l}} \times \frac{\partial H_{i,l}}{\partial W_{k,l}^{(1)}}$$

Breaking it down:

1. First three terms are already computed:

$$(Y_i - \hat{Y}_i) \times \hat{Y}_i \times (1 - \hat{Y}_i)$$

2. Derivative of $O$ w.r.t. $Z$:

$$\frac{\partial O_i}{\partial Z_{i,l}} = W_l^{(2)}$$

3. Derivative of $Z$ w.r.t. $H$ (Sigmoid Activation Derivative):

$$\frac{\partial Z_{i,l}}{\partial H_{i,l}} = Z_{i,l} \times (1 - Z_{i,l})$$

4. Derivative of $H$ w.r.t. $W^{(1)}$:

$$\frac{\partial H_{i,l}}{\partial W_{k,l}^{(1)}} = X_{i,k}$$

Thus, the full gradient for $W^{(1)}$ including $\frac{1}{N}$ is:

$$\frac{\partial L}{\partial W_{k,l}^{(1)}} = -\frac{1}{N}\sum_{i=1}^{N} X_{i,k} \times (Y_i - \hat{Y}_i) \times \hat{Y}_i \times (1 - \hat{Y}_i) \times W_l^{(2)} \times Z_{i,l} \times (1 - Z_{i,l})$$

# 4 4. Gradient Descent Updates

After computing the gradients, we update the weights using gradient descent:

$$W_j^{(2)} \leftarrow W_j^{(2)} - \gamma \frac{\partial L}{\partial W_j^{(2)}}$$

$$W_{k,l}^{(1)} \leftarrow W_{k,l}^{(1)} - \gamma \frac{\partial L}{\partial W_{k,l}^{(1)}}$$

where $\gamma$ is the learning rate.