Backpropagation Gradient Derivations

March 21, 2025

1 Loss Function

The Mean Squared Error (MSE) loss function is defined as:

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

2 Forward Propagation

The hidden layer pre-activation is given by:

$$H = XW^{(1)}. \tag{2}$$

Applying the sigmoid activation function:

$$Z = \sigma(H), \quad \text{where } \sigma(x) = \frac{1}{1 + e^{-x}}. \tag{3}$$

Appending a bias column to Z:

$$O = ZW^{(2)}. \tag{4}$$

Applying the sigmoid function to obtain the final output:

$$\hat{y} = \sigma(O). \tag{5}$$

3 Backpropagation

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

$$\frac{\partial L}{\partial W_{k,1}^{(2)}} = \sum_{i=1}^{N} \frac{\partial L}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial O_i} \cdot \frac{\partial O_i}{\partial W_{k,1}^{(2)}}. \tag{6}$$

Expanding each term:

- $\frac{\partial L}{\partial \hat{y}_i} = -\frac{1}{N}(Y_i - \hat{y}_i)$,
- $\frac{\partial \hat{y}_i}{\partial O_i} = \hat{y}_i(1 - \hat{y}_i)$,
- $\frac{\partial O_i}{\partial W_{k,1}^{(2)}} = Z_{i,k}$.

Substituting these values:

$$\frac{\partial L}{\partial W_{k,1}^{(2)}} = \sum_{i=1}^{N} -\frac{1}{N}(Y_i - \hat{y}_i)\hat{y}_i(1 - \hat{y}_i)Z_{i,k}. \tag{7}$$

The gradient for $W^{(1)}$ is computed as:

$$\frac{\partial L}{\partial W_{k,l}^{(1)}} = \sum_{i=1}^{N} \frac{\partial L}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial O_i} \cdot \frac{\partial O_i}{\partial Z_{i,l}} \cdot \frac{\partial Z_{i,l}}{\partial H_{i,l}} \cdot \frac{\partial H_{i,l}}{\partial W_{k,l}^{(1)}}. \tag{8}$$

Expanding each term:

- The first two terms remain the same as in the previous calculation,
- $\frac{\partial O_i}{\partial Z_{i,l}} = W_{l,1}^{(2)}$,
- $\frac{\partial Z_{i,l}}{\partial H_{i,l}} = \sigma(H_{i,l})(1 - \sigma(H_{i,l}))$,
- $\frac{\partial H_{i,l}}{\partial W_{k,l}^{(1)}} = X_{i,k}$.

Substituting these values:

$$\frac{\partial L}{\partial W_{k,l}^{(1)}} = \sum_{i=1}^{N} -\frac{1}{N}(Y_i - \hat{y}_i)\hat{y}_i(1 - \hat{y}_i)W_{l,1}^{(2)}\sigma(H_{i,l})(1 - \sigma(H_{i,l}))X_{i,k}. \tag{9}$$