Neural Network Backpropagation with Generic Activation Functions

Architecture

- **Input Layer:** 4 neurons
- **Hidden Layer 1:** 2 neurons

- **Hidden Layer 2:** 4 neurons
- **Hidden Layer 3:** 2 neurons
- **Output Layer:** 2 neurons

Notation

- L: Loss function
- $z_j^{(l)}$: Pre-activation for neuron j in layer l
- $a_j^{(l)}$: Activation output for neuron j in layer l
- $W_{jk}^{(l)}$: Weight connecting neuron k in layer $l-1$ to neuron j in layer l
- $b_j^{(l)}$: Bias for neuron j in layer l
- $g(z)$: Generic activation function for hidden layers
- $g'(z)$: Derivative of activation function

Forward Pass

Input Layer

$$a_k^{(0)} = x_k \text{ for } k = 1, \dots, 4$$

Hidden Layer 1

$$z_j^{(1)} = \sum_{k=1}^{4} W_{jk}^{(1)} a_k^{(0)} + b_j^{(1)}$$

$$a_j^{(1)} = g(z_j^{(1)})$$

Hidden Layer 2

$$z_j^{(2)} = \sum_{k=1}^{2} W_{jk}^{(2)} a_k^{(1)} + b_j^{(2)}$$

$$a_j^{(2)} = g(z_j^{(2)})$$

Hidden Layer 3

$$z_j^{(3)} = \sum_{k=1}^{4} W_{jk}^{(3)} a_k^{(2)} + b_j^{(3)}$$

$$a_j^{(3)} = g(z_j^{(3)})$$

Output Layer

$$z_j^{(4)} = \sum_{k=1}^{2} W_{jk}^{(4)} a_k^{(3)} + b_j^{(4)}$$

$$a_j^{(4)} = \text{softmax}(z_j^{(4)}) = \frac{e^{z_j^{(4)}}}{\sum_{p=1}^{2} e^{z_p^{(4)}}}$$

Loss Function

$$L = -\sum_{j=1}^{2} t_j \log(a_j^{(4)})$$

Backpropagation

Output Layer (L=4)

Error Term:

$$\delta_j^{(4)} = a_j^{(4)} - t_j$$

Weight Gradients:

$$\frac{\partial L}{\partial W_{jk}^{(4)}} = \delta_j^{(4)} a_k^{(3)}$$

Bias Gradients:

$$\frac{\partial L}{\partial b_j^{(4)}} = \delta_j^{(4)}$$

Hidden Layer 3 (L=3)

Error Term:

$$\delta_k^{(3)} = \left(\sum_{j=1}^{2} W_{jk}^{(4)} \delta_j^{(4)} \right) g'(z_k^{(3)})$$

Weight Gradients:

$$\frac{\partial L}{\partial W_{kl}^{(3)}} = \delta_k^{(3)} a_l^{(2)}$$

Bias Gradients:

$$\frac{\partial L}{\partial b_k^{(3)}} = \delta_k^{(3)}$$

Hidden Layer 2 (L=2)

Error Term:

$$\delta_l^{(2)} = \left(\sum_{k=1}^{2} W_{kl}^{(3)} \delta_k^{(3)} \right) g'(z_l^{(2)})$$

Weight Gradients:

$$\frac{\partial L}{\partial W_{lm}^{(2)}} = \delta_l^{(2)} a_m^{(1)}$$

Bias Gradients:

$$\frac{\partial L}{\partial b_l^{(2)}} = \delta_l^{(2)}$$

Hidden Layer 1 (L=1)

Error Term:

$$\delta_m^{(1)} = \left(\sum_{l=1}^{4} W_{lm}^{(2)} \delta_l^{(2)} \right) g'(z_m^{(1)})$$

Weight Gradients:

$$\frac{\partial L}{\partial W_{mn}^{(1)}} = \delta_m^{(1)} a_n^{(0)}$$

Bias Gradients:

$$\frac{\partial L}{\partial b_m^{(1)}} = \delta_m^{(1)}$$

General Formulas

Error Term for Hidden Layer l

$$\delta_j^{(l)} = \left(\sum_{p=1}^{n_{l+1}} W_{pj}^{(l+1)} \delta_p^{(l+1)} \right) g'(z_j^{(l)})$$

Weight Gradients

$$\frac{\partial L}{\partial W_{jk}^{(l)}} = \delta_j^{(l)} a_k^{(l-1)}$$

Bias Gradients

$$\frac{\partial L}{\partial b_j^{(l)}} = \delta_j^{(l)}$$