Neural Network Backpropagation with Generic Activation Functions

Architecture

- **Input Layer:** 4 neurons (indexed 1,2,3,4)
- **Hidden Layer 1:** 2 neurons (indexed 1,2)
- **Hidden Layer 2:** 4 neurons (indexed 1,2,3,4)
- **Hidden Layer 3:** 2 neurons (indexed 1,2)
- **Output Layer:** 2 neurons (indexed 1,2)

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
- t_j: Target value for output neuron j

Forward Pass

Input Layer

$$a_1^{(0)} = x_1, \quad a_2^{(0)} = x_2, \quad a_3^{(0)} = x_3, \quad a_4^{(0)} = x_4$$

Hidden Layer 1

$$z_1^{(1)} = W_{11}^{(1)} a_1^{(0)} + W_{12}^{(1)} a_2^{(0)} + W_{13}^{(1)} a_3^{(0)} + W_{14}^{(1)} a_4^{(0)} + b_1^{(1)}$$
$$z_2^{(1)} = W_{21}^{(1)} a_1^{(0)} + W_{22}^{(1)} a_2^{(0)} + W_{23}^{(1)} a_3^{(0)} + W_{24}^{(1)} a_4^{(0)} + b_2^{(1)}$$

$$a_1^{(1)} = g(z_1^{(1)}), \quad a_2^{(1)} = g(z_2^{(1)})$$

Hidden Layer 2

$$z_1^{(2)} = W_{11}^{(2)} a_1^{(1)} + W_{12}^{(2)} a_2^{(1)} + b_1^{(2)}$$
$$z_2^{(2)} = W_{21}^{(2)} a_1^{(1)} + W_{22}^{(2)} a_2^{(1)} + b_2^{(2)}$$
$$z_3^{(2)} = W_{31}^{(2)} a_1^{(1)} + W_{32}^{(2)} a_2^{(1)} + b_3^{(2)}$$
$$z_4^{(2)} = W_{41}^{(2)} a_1^{(1)} + W_{42}^{(2)} a_2^{(1)} + b_4^{(2)}$$

$$a_1^{(2)} = g(z_1^{(2)}), \quad a_2^{(2)} = g(z_2^{(2)}), \quad a_3^{(2)} = g(z_3^{(2)}), \quad a_4^{(2)} = g(z_4^{(2)})$$

Hidden Layer 3

$$z_1^{(3)} = W_{11}^{(3)} a_1^{(2)} + W_{12}^{(3)} a_2^{(2)} + W_{13}^{(3)} a_3^{(2)} + W_{14}^{(3)} a_4^{(2)} + b_1^{(3)}$$
$$z_2^{(3)} = W_{21}^{(3)} a_1^{(2)} + W_{22}^{(3)} a_2^{(2)} + W_{23}^{(3)} a_3^{(2)} + W_{24}^{(3)} a_4^{(2)} + b_2^{(3)}$$

$$a_1^{(3)} = g(z_1^{(3)}), \quad a_2^{(3)} = g(z_2^{(3)})$$

Output Layer

$$z_1^{(4)} = W_{11}^{(4)} a_1^{(3)} + W_{12}^{(4)} a_2^{(3)} + b_1^{(4)}$$
$$z_2^{(4)} = W_{21}^{(4)} a_1^{(3)} + W_{22}^{(4)} a_2^{(3)} + b_2^{(4)}$$

$$a_1^{(4)} = \frac{e^{z_1^{(4)}}}{e^{z_1^{(4)}} + e^{z_2^{(4)}}}, \quad a_2^{(4)} = \frac{e^{z_2^{(4)}}}{e^{z_1^{(4)}} + e^{z_2^{(4)}}}$$

Loss Function

$$L = -t_1 \log(a_1^{(4)}) - t_2 \log(a_2^{(4)})$$

Backpropagation

Output Layer (L=4)

Error Terms:

$$\delta_1^{(4)} = a_1^{(4)} - t_1$$
$$\delta_2^{(4)} = a_2^{(4)} - t_2$$

Weight Gradients:

$$\frac{\partial L}{\partial W_{11}^{(4)}} = \delta_1^{(4)} a_1^{(3)}, \quad \frac{\partial L}{\partial W_{12}^{(4)}} = \delta_1^{(4)} a_2^{(3)}$$
$$\frac{\partial L}{\partial W_{21}^{(4)}} = \delta_2^{(4)} a_1^{(3)}, \quad \frac{\partial L}{\partial W_{22}^{(4)}} = \delta_2^{(4)} a_2^{(3)}$$

Bias Gradients:

$$\frac{\partial L}{\partial b_1^{(4)}} = \delta_1^{(4)}, \quad \frac{\partial L}{\partial b_2^{(4)}} = \delta_2^{(4)}$$

Hidden Layer 3 (L=3)

Error Terms:

$$\delta_1^{(3)} = (W_{11}^{(4)}\delta_1^{(4)} + W_{21}^{(4)}\delta_2^{(4)})g'(z_1^{(3)})$$
$$\delta_2^{(3)} = (W_{12}^{(4)}\delta_1^{(4)} + W_{22}^{(4)}\delta_2^{(4)})g'(z_2^{(3)})$$

Weight Gradients:

$$\frac{\partial L}{\partial W_{11}^{(3)}} = \delta_1^{(3)}a_1^{(2)}, \quad \frac{\partial L}{\partial W_{12}^{(3)}} = \delta_1^{(3)}a_2^{(2)}$$
$$\frac{\partial L}{\partial W_{13}^{(3)}} = \delta_1^{(3)}a_3^{(2)}, \quad \frac{\partial L}{\partial W_{14}^{(3)}} = \delta_1^{(3)}a_4^{(2)}$$
$$\frac{\partial L}{\partial W_{21}^{(3)}} = \delta_2^{(3)}a_1^{(2)}, \quad \frac{\partial L}{\partial W_{22}^{(3)}} = \delta_2^{(3)}a_2^{(2)}$$
$$\frac{\partial L}{\partial W_{23}^{(3)}} = \delta_2^{(3)}a_3^{(2)}, \quad \frac{\partial L}{\partial W_{24}^{(3)}} = \delta_2^{(3)}a_4^{(2)}$$

Bias Gradients:

$$\frac{\partial L}{\partial b_1^{(3)}} = \delta_1^{(3)}, \quad \frac{\partial L}{\partial b_2^{(3)}} = \delta_2^{(3)}$$

Hidden Layer 2 (L=2)

Error Terms:

$$\delta_1^{(2)} = (W_{11}^{(3)}\delta_1^{(3)} + W_{21}^{(3)}\delta_2^{(3)})g'(z_1^{(2)})$$
$$\delta_2^{(2)} = (W_{12}^{(3)}\delta_1^{(3)} + W_{22}^{(3)}\delta_2^{(3)})g'(z_2^{(2)})$$
$$\delta_3^{(2)} = (W_{13}^{(3)}\delta_1^{(3)} + W_{23}^{(3)}\delta_2^{(3)})g'(z_3^{(2)})$$
$$\delta_4^{(2)} = (W_{14}^{(3)}\delta_1^{(3)} + W_{24}^{(3)}\delta_2^{(3)})g'(z_4^{(2)})$$

Weight Gradients:

$$\frac{\partial L}{\partial W_{11}^{(2)}} = \delta_1^{(2)}a_1^{(1)}, \quad \frac{\partial L}{\partial W_{12}^{(2)}} = \delta_1^{(2)}a_2^{(1)}$$
$$\frac{\partial L}{\partial W_{21}^{(2)}} = \delta_2^{(2)}a_1^{(1)}, \quad \frac{\partial L}{\partial W_{22}^{(2)}} = \delta_2^{(2)}a_2^{(1)}$$
$$\frac{\partial L}{\partial W_{31}^{(2)}} = \delta_3^{(2)}a_1^{(1)}, \quad \frac{\partial L}{\partial W_{32}^{(2)}} = \delta_3^{(2)}a_2^{(1)}$$
$$\frac{\partial L}{\partial W_{41}^{(2)}} = \delta_4^{(2)}a_1^{(1)}, \quad \frac{\partial L}{\partial W_{42}^{(2)}} = \delta_4^{(2)}a_2^{(1)}$$

Bias Gradients:

$$\frac{\partial L}{\partial b_1^{(2)}} = \delta_1^{(2)}, \quad \frac{\partial L}{\partial b_2^{(2)}} = \delta_2^{(2)}$$
$$\frac{\partial L}{\partial b_3^{(2)}} = \delta_3^{(2)}, \quad \frac{\partial L}{\partial b_4^{(2)}} = \delta_4^{(2)}$$

Hidden Layer 1 (L=1)

Error Terms:

$$\delta_1^{(1)} = (W_{11}^{(2)}\delta_1^{(2)} + W_{21}^{(2)}\delta_2^{(2)} + W_{31}^{(2)}\delta_3^{(2)} + W_{41}^{(2)}\delta_4^{(2)})g'(z_1^{(1)})$$

$$\delta_2^{(1)} = (W_{12}^{(2)}\delta_1^{(2)} + W_{22}^{(2)}\delta_2^{(2)} + W_{32}^{(2)}\delta_3^{(2)} + W_{42}^{(2)}\delta_4^{(2)})g'(z_2^{(1)})$$

Weight Gradients:

$$\frac{\partial L}{\partial W_{11}^{(1)}} = \delta_1^{(1)}a_1^{(0)}, \quad \frac{\partial L}{\partial W_{12}^{(1)}} = \delta_1^{(1)}a_2^{(0)}$$

$$\frac{\partial L}{\partial W_{13}^{(1)}} = \delta_1^{(1)}a_3^{(0)}, \quad \frac{\partial L}{\partial W_{14}^{(1)}} = \delta_1^{(1)}a_4^{(0)}$$
$$\frac{\partial L}{\partial W_{21}^{(1)}} = \delta_2^{(1)}a_1^{(0)}, \quad \frac{\partial L}{\partial W_{22}^{(1)}} = \delta_2^{(1)}a_2^{(0)}$$
$$\frac{\partial L}{\partial W_{23}^{(1)}} = \delta_2^{(1)}a_3^{(0)}, \quad \frac{\partial L}{\partial W_{24}^{(1)}} = \delta_2^{(1)}a_4^{(0)}$$

Bias Gradients:

$$\frac{\partial L}{\partial b_1^{(1)}} = \delta_1^{(1)}, \quad \frac{\partial L}{\partial b_2^{(1)}} = \delta_2^{(1)}$$

Summary

- **Total Parameters:** 42 weights + 10 biases = 52 parameters
- **Weight Distribution:** Layer 1: 8 weights, Layer 2: 8 weights, Layer 3: 8 weights, Layer 4: 4 weights
- **Bias Distribution:** 2 + 4 + 2 + 2 = 10 biases
- **Key Insight:** Error terms ( ) equal bias gradients for all layers