1 Why Do We Ignore the Bias Term in δ_H but Not in δ_O?

In backpropagation, we handle bias terms differently for the hidden and output layers. The key difference lies in whether the bias neuron has incoming weights that need updates.

1.1 Understanding δ_O (Output Layer Error)

The error at the output layer is computed as:

$$\delta_O = -(y - \hat{y}) \cdot \sigma'(O), \tag{1}$$

where:

- $y - \hat{y}$ represents the error between the actual and predicted outputs.
- $\sigma'(O)$ is the derivative of the activation function at the output layer.

Since the bias neuron in the output layer has an associated weight in $W^{(2)}$, it directly affects the final prediction $\hat{y}$. Thus, we **do not ignore** the bias term in δ_O.

1.2 Understanding δ_H (Hidden Layer Error)

The error at the hidden layer is computed as:

$$\delta_H = \left(\delta_O W^{(2)T}\right) \odot \sigma'(H). \tag{2}$$

However, we ignore the bias term in δ_H because:

- The bias neuron in the hidden layer does not have incoming connections; it is always set to 1.
- Backpropagation only adjusts weights for neurons with incoming connections.

Thus, we exclude the bias term when computing δ_H:

$$\delta_H = \left(\delta_O W^{(2)T}\right)_{[:,:-1]} \odot \sigma'(H)_{[:,:-1]}. \tag{3}$$

Term	Does Bias Have Incoming Weights?	Should We Compute Error for It?
δ_H (Hidden Layer)	No	Ignore bias term
δ_O (Output Layer)	Yes	Do not ignore bias term