Backpropagation Equations

Parth Bansal

June 20, 2025

Backpropagation Derivation for a 3-Layer Neural Network

We consider a fully connected feedforward neural network with the following layer dimensions:

- Input layer: $X \in \mathbb{R}^{n_x \times m}$
- Hidden layer 1: $A^{[1]} \in \mathbb{R}^{n_{h1} \times m}$
- Hidden layer 2: $A^{[2]} \in \mathbb{R}^{n_{h2} \times m}$
- Output layer: $A^{[3]} \in \mathbb{R}^{n_y \times m}$ (softmax output)
- True labels: $Y \in \{0, 1\}^{n_y \times m}$ (one-hot encoded)

Forward Propagation

$$\begin{aligned}
Z^{[1]} &= W^{[1]}X + b^{[1]} \\
A^{[1]} &= \text{ReLU}(Z^{[1]}) \\
Z^{[2]} &= W^{[2]}A^{[1]} + b^{[2]} \\
A^{[2]} &= \text{ReLU}(Z^{[2]}) \\
Z^{[3]} &= W^{[3]}A^{[2]} + b^{[3]} \\
A^{[3]} &= \text{softmax}(Z^{[3]})
\end{aligned}$$

Loss Function

We use the categorical cross-entropy loss: $\mathcal{L} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_y} Y_{j,i}\log A^{[3]}_{j,i}$

Detailed Backpropagation Derivation

Notation and Loss

Let

$$Z^{[3]} = W^{[3]}A^{[2]} + b^{[3]}, \quad A^{[3]} = \mathrm{softmax}(Z^{[3]}), \quad \mathcal{L} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_y} Y_{j,i}\log A^{[3]}_{j,i}.$$

Step 1: $dZ^{[3]} = \dfrac{\partial \mathcal{L}}{\partial Z^{[3]}}$

We first compute

$$\frac{\partial \mathcal{L}}{\partial A^{[3]}_{k,i}} = -\frac{1}{m}\,\frac{Y_{k,i}}{A^{[3]}_{k,i}}.$$

But $A^{[3]}_{k,i} = \dfrac{e^{Z^{[3]}_{k,i}}}{\sum_r e^{Z^{[3]}_{r,i}}}$, so

$$\frac{\partial A^{[3]}_{\ell,i}}{\partial Z^{[3]}_{k,i}} = A^{[3]}_{\ell,i}(\delta_{k\ell} - A^{[3]}_{k,i}).$$

Hence by the chain rule,

$$\frac{\partial \mathcal{L}}{\partial Z^{[3]}_{k,i}} = \sum_{\ell=1}^{n_y} \frac{\partial \mathcal{L}}{\partial A^{[3]}_{\ell,i}}\,\frac{\partial A^{[3]}_{\ell,i}}{\partial Z^{[3]}_{k,i}} = \frac{1}{m}\Big(A^{[3]}_{k,i} - Y_{k,i}\Big).$$

In matrix form:

$$\boxed{dZ^{[3]} = A^{[3]} - Y.}$$

Step 2: Parameter Gradients at Output

Using

$$Z^{[3]} = W^{[3]}A^{[2]} + b^{[3]},$$

we have

$$\frac{\partial \mathcal{L}}{\partial W^{[3]}} = dZ^{[3]}\,A^{[2]^\top}, \qquad \frac{\partial \mathcal{L}}{\partial b^{[3]}} = \sum_{i=1}^{m} dZ^{[3]}_{:,i}.$$

Averaging over m examples gives

$$\boxed{dW^{[3]} = \frac{1}{m}\,dZ^{[3]}A^{[2]^\top}, \quad db^{[3]} = \frac{1}{m}\sum_{i=1}^{m} dZ^{[3]}_{:,i}.}$$

Step 3: Propagate into Hidden Layer 2

We use

$$dA^{[2]} = W^{[3]^\top}\,dZ^{[3]},$$

and since $A^{[2]} = \mathrm{ReLU}(Z^{[2]})$, $'(z) = \mathbf{1}_{z>0}$, so

$$\boxed{dZ^{[2]} = dA^{[2]} \circ \mathbf{1}_{Z^{[2]}>0}.}$$

Then by the same linear-layer rule:

$$\boxed{dW^{[2]} = \tfrac{1}{m}\,dZ^{[2]}A^{[1]^\top}, \quad db^{[2]} = \tfrac{1}{m}\sum_{i=1}^{m} dZ^{[2]}_{:,i}.}$$

Step 4: Propagate into Hidden Layer 1

Similarly,

$$dA^{[1]} = W^{[2]\top} \, dZ^{[2]}, \quad dZ^{[1]} = dA^{[1]} \circ \mathbf{1}_{Z^{[1]}>0},$$

and

$$\boxed{dW^{[1]} = \tfrac{1}{m} \, dZ^{[1]} X^\top, \quad db^{[1]} = \tfrac{1}{m} \sum_{i=1}^{m} dZ^{[1]}_{:,i}.}$$

Compact Summary

For each layer $l = 3, 2, 1$:

$$dZ^{[l]} = \begin{cases} A^{[l]} - Y, & l = 3, \\ \big(W^{[l+1]\top} dZ^{[l+1]}\big) \circ \mathbf{1}_{Z^{[l]}>0}, & l < 3, \end{cases}$$

$$dW^{[l]} = \tfrac{1}{m} \, dZ^{[l]} A^{[l-1]\top}, \quad db^{[l]} = \tfrac{1}{m} \sum_{i=1}^{m} dZ^{[l]}_{:,i}.$$

Notes

- ReLU derivative: $\frac{\partial A}{\partial Z} = \nVdash_{Z>0}$
- Softmax + cross-entropy simplifies: $\frac{\partial \mathcal{L}}{\partial Z^{[3]}} = A^{[3]} - Y$
- Each gradient is averaged over the batch of m examples