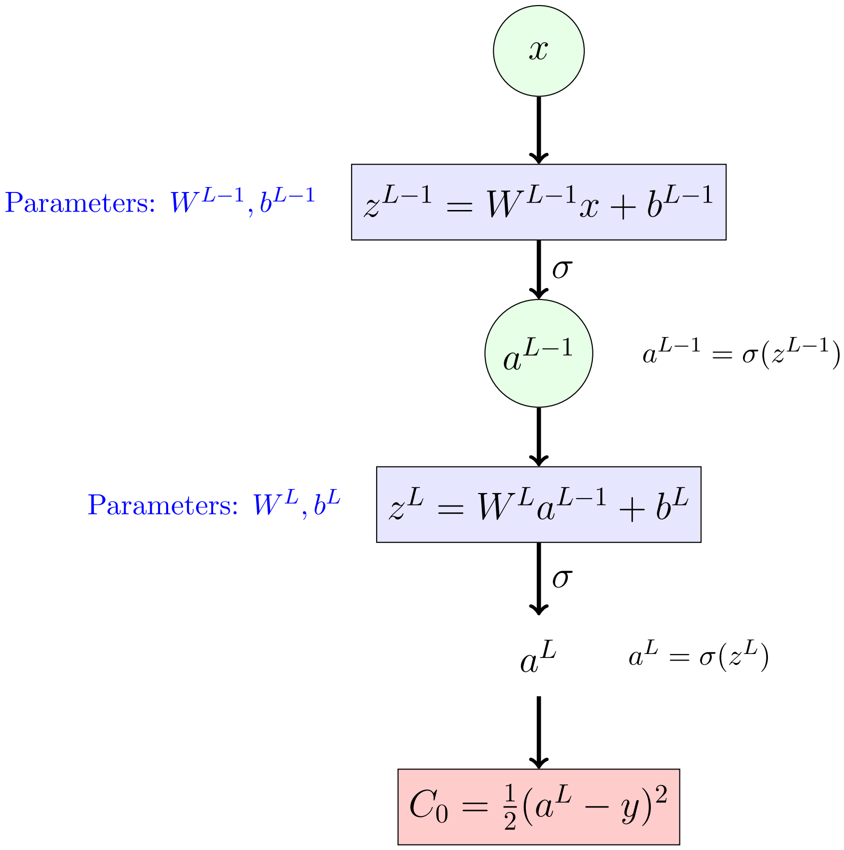
Backpropagation

1 Network Architecture

We consider a simple neural network with two linear layers:



2 Forward Pass Equations

The forward pass through the network is defined by:

$$z^{L-1} = W^{L-1}x + b^{L-1} \tag{1}$$

$$a^{L-1} = \sigma(z^{L-1}) \tag{2}$$

$$z^L = W^L a^{L-1} + b^L \tag{3}$$

$$a^L = \sigma(z^L) \tag{4}$$

3 Cost Function

For a single training example, the cost function is:

$$C_0 = \frac{1}{2}(a^L - y)^2 \tag{5}$$

where y is the target output.

4 Backpropagation: Chain Rule Decomposition

To update the weights W^L, we need to compute $\frac{\partial C_0}{\partial W^L}$. Using the chain rule:

$$\frac{\partial C_0}{\partial W^L} = \frac{\partial z^L}{\partial W^L} \cdot \frac{\partial a^L}{\partial z^L} \cdot \frac{\partial C_0}{\partial a^L} \tag{6}$$

We will compute each term step by step.

5 Derivative 1: $\frac{\partial C_0}{\partial a^L}$

Starting with the cost function:

$$C_0 = \frac{1}{2}(a^L - y)^2 \tag{7}$$

Taking the derivative with respect to a^L:

$$\begin{align}
\frac{\partial C_0}{\partial a^L} &= \frac{\partial}{\partial a^L}\left[\frac{1}{2}(a^L - y)^2\right] \tag{8}\\
&= \frac{1}{2} \cdot 2(a^L - y) \cdot \frac{\partial}{\partial a^L}(a^L - y) \tag{9}\\
&= (a^L - y) \cdot 1 \tag{10}\\
&= a^L - y \tag{11}
\end{align}$$

Result:

$$\boxed{\frac{\partial C_0}{\partial a^L} = a^L - y} \tag{12}$$

6 Derivative 2: $\frac{\partial a^L}{\partial z^L}$ (Sigmoid Derivative)

Since $a^L = \sigma(z^L)$ where $\sigma(z) = \frac{1}{1+e^{-z}}$, we need to find the derivative of the sigmoid function.

6.1 Step-by-Step Sigmoid Derivative

$$\sigma(z) = \frac{1}{1 + e^{-z}} = (1 + e^{-z})^{-1} \tag{13}$$

Using the chain rule:

$$\frac{d\sigma}{dz} = \frac{d}{dz}(1 + e^{-z})^{-1} \tag{14}$$
$$= -1 \cdot (1 + e^{-z})^{-2} \cdot \frac{d}{dz}(1 + e^{-z}) \tag{15}$$
$$= -(1 + e^{-z})^{-2} \cdot (-e^{-z}) \tag{16}$$
$$= \frac{e^{-z}}{(1 + e^{-z})^2} \tag{17}$$

Now we rewrite this in terms of $\sigma(z)$:

$$\frac{d\sigma}{dz} = \frac{e^{-z}}{(1 + e^{-z})^2} \tag{18}$$
$$= \frac{1}{1 + e^{-z}} \cdot \frac{e^{-z}}{1 + e^{-z}} \tag{19}$$
$$= \sigma(z) \cdot \frac{e^{-z}}{1 + e^{-z}} \tag{20}$$

Note that:

$$\frac{e^{-z}}{1 + e^{-z}} = \frac{1 + e^{-z} - 1}{1 + e^{-z}} \tag{21}$$
$$= 1 - \frac{1}{1 + e^{-z}} \tag{22}$$
$$= 1 - \sigma(z) \tag{23}$$

Therefore:

$$\frac{d\sigma(z)}{dz} = \sigma(z)(1 - \sigma(z)) \tag{24}$$

Applying this to our network:

$$\frac{\partial a^L}{\partial z^L} = \frac{d\sigma(z^L)}{dz^L} \tag{25}$$
$$= \sigma(z^L)(1 - \sigma(z^L)) \tag{26}$$
$$= a^L(1 - a^L) \tag{27}$$

Result:

$$\boxed{\frac{\partial a^L}{\partial z^L} = a^L(1 - a^L)} \tag{28}$$

7 Derivative 3: $\frac{\partial z^L}{\partial W^L}$

From the forward pass, we have:

$$z^L = W^L a^{L-1} + b^L \tag{29}$$

Taking the derivative with respect to W^L:

$$\frac{\partial z^L}{\partial W^L} = \frac{\partial}{\partial W^L}(W^L a^{L-1} + b^L) \tag{30}$$
$$= a^{L-1} \cdot \frac{\partial W^L}{\partial W^L} + 0 \tag{31}$$
$$= a^{L-1} \tag{32}$$

Result:

$$\boxed{\frac{\partial z^L}{\partial W^L} = a^{L-1}} \tag{33}$$

8 Complete Gradient: $\frac{\partial C_0}{\partial W^L}$

Combining all three derivatives using the chain rule:

$$\frac{\partial C_0}{\partial W^L} = \frac{\partial z^L}{\partial W^L} \cdot \frac{\partial a^L}{\partial z^L} \cdot \frac{\partial C_0}{\partial a^L} \tag{34}$$
$$= a^{L-1} \cdot a^L(1 - a^L) \cdot (a^L - y) \tag{35}$$

Final Result:

$$\boxed{\frac{\partial C_0}{\partial W^L} = a^{L-1} \cdot a^L(1 - a^L) \cdot (a^L - y)} \tag{36}$$

9 Gradient for Bias: $\frac{\partial C_0}{\partial b^L}$

Similarly, for the bias term:

$$\frac{\partial C_0}{\partial b^L} = \frac{\partial z^L}{\partial b^L} \cdot \frac{\partial a^L}{\partial z^L} \cdot \frac{\partial C_0}{\partial a^L} \tag{37}$$

Since $z^L = W^L a^{L-1} + b^L$:

$$\frac{\partial z^L}{\partial b^L} = 1 \tag{38}$$

Therefore:

$$\frac{\partial C_0}{\partial b^L} = 1 \cdot a^L(1 - a^L) \cdot (a^L - y) \tag{39}$$

Final Result:

$$\boxed{\frac{\partial C_0}{\partial b^L} = a^L(1 - a^L) \cdot (a^L - y)} \tag{40}$$

10 Gradients for Previous Layer: $\frac{\partial C_0}{\partial W^{L-1}}$ and $\frac{\partial C_0}{\partial b^{L-1}}$

10.1 Chain Rule for W^{L-1}

To propagate the error back to the previous layer:

$$\frac{\partial C_0}{\partial W^{L-1}} = \frac{\partial z^{L-1}}{\partial W^{L-1}} \cdot \frac{\partial a^{L-1}}{\partial z^{L-1}} \cdot \frac{\partial z^L}{\partial a^{L-1}} \cdot \frac{\partial a^L}{\partial z^L} \cdot \frac{\partial C_0}{\partial a^L} \tag{41}$$

10.2 Computing Each Term

1. $\frac{\partial z^{L-1}}{\partial W^{L-1}}$**:**

$$z^{L-1} = W^{L-1}x + b^{L-1} \implies \frac{\partial z^{L-1}}{\partial W^{L-1}} = x \tag{42}$$

2. $\frac{\partial a^{L-1}}{\partial z^{L-1}}$**:**

$$\frac{\partial a^{L-1}}{\partial z^{L-1}} = a^{L-1}(1 - a^{L-1}) \tag{43}$$

3. $\frac{\partial z^{L}}{\partial a^{L-1}}$**:**

$$z^L = W^L a^{L-1} + b^L \implies \frac{\partial z^L}{\partial a^{L-1}} = W^L \tag{44}$$

4. We already computed:

$$\frac{\partial a^L}{\partial z^L} = a^L(1 - a^L) \tag{45}$$

$$\frac{\partial C_0}{\partial a^L} = a^L - y \tag{46}$$

10.3 Final Result for $\frac{\partial C_0}{\partial W^{L-1}}$

$$\frac{\partial C_0}{\partial W^{L-1}} = x \cdot a^{L-1}(1 - a^{L-1}) \cdot W^L \cdot a^L(1 - a^L) \cdot (a^L - y) \tag{47}$$

Final Result:

$$\boxed{\frac{\partial C_0}{\partial W^{L-1}} = x \cdot a^{L-1}(1 - a^{L-1}) \cdot W^L \cdot a^L(1 - a^L) \cdot (a^L - y)} \tag{48}$$

10.4 Gradient for b^{L-1}

$$\frac{\partial z^{L-1}}{\partial b^{L-1}} = 1 \tag{49}$$

Therefore:

$$\frac{\partial C_0}{\partial b^{L-1}} = 1 \cdot a^{L-1}(1 - a^{L-1}) \cdot W^L \cdot a^L(1 - a^L) \cdot (a^L - y) \tag{50}$$

Final Result:

$$\boxed{\frac{\partial C_0}{\partial b^{L-1}} = a^{L-1}(1 - a^{L-1}) \cdot W^L \cdot a^L(1 - a^L) \cdot (a^L - y)} \tag{51}$$

11 Summary of All Gradients

$$\frac{\partial C_0}{\partial W^L} = a^{L-1} \cdot a^L(1 - a^L) \cdot (a^L - y) \tag{52}$$

$$\frac{\partial C_0}{\partial b^L} = a^L(1 - a^L) \cdot (a^L - y) \tag{53}$$

$$\frac{\partial C_0}{\partial W^{L-1}} = x \cdot a^{L-1}(1 - a^{L-1}) \cdot W^L \cdot a^L(1 - a^L) \cdot (a^L - y) \tag{54}$$

$$\frac{\partial C_0}{\partial b^{L-1}} = a^{L-1}(1 - a^{L-1}) \cdot W^L \cdot a^L(1 - a^L) \cdot (a^L - y) \tag{55}$$