Backpropagation

$$\nabla L = (\nabla_w L, \nabla_b L)$$

Start computation <u>backwards</u> (last layer)

For NN on p. 1

$$y = \varphi(s_1^{(2)}) \text{ where } s_1^{(2)} = w_{01}^{(2)} x_0^{(1)} + w_{11}^{(2)} x_1^{(1)} + w_{21}^{(2)} x_2^{(1)}$$

$s_j^{(k)}$: signal input to neuron j in layer k

Loss depends on y, which depends on $w_{ij}^{(2)}$ via $s_1^{(2)}$ and $y = \varphi(s_1^{(2)})$. Applying the chain rule:

$$\frac{\partial L}{\partial b_1^{(2)}} = \frac{\partial L}{\partial w_{01}^{(2)}} = \frac{\partial L}{\partial s_1^{(2)}} \frac{\partial s_1^{(2)}}{\partial w_{01}^{(2)}} = \delta_1^{(2)} x_0^{(1)} = -\delta_1^{(2)} \qquad (1)$$

$$\frac{\partial L}{\partial w_{11}^{(2)}} = \frac{\partial L}{\partial s_1^{(2)}} \frac{\partial s_1^{(2)}}{\partial w_{11}^{(2)}} = \delta_1^{(2)} x_1^{(1)} \qquad (2)$$

$$\frac{\partial L}{\partial w_{21}^{(2)}} = \frac{\partial L}{\partial s_1^{(2)}} \frac{\partial s_1^{(2)}}{\partial w_{21}^{(2)}} = \delta_1^{(2)} x_2^{(1)} \qquad (3)$$

where $\delta_1^{(2)} = \dfrac{\partial L}{\partial s_1^{(2)}}$ (4)

For next layer, $k=1$, L depends on y which depends on $w_{ij}^{(1)}$ via $S_j^{(1)}$, $j=1,2$, where $y = \varphi\left(\sum_{i=1}^{2} w_{i1}^{(2)} \varphi(S_i^{(1)}) - b_1^{(2)}\right)$.

Applying the chain rule:

For $j=1$

$$\frac{\partial L}{\partial b_1^{(1)}} = \frac{\partial L}{\partial w_{01}^{(1)}} = \frac{\partial L}{\partial S_1^{(1)}} \frac{\partial S_1^{(1)}}{\partial w_{01}^{(1)}} = \delta_1^{(1)} x_0^{(0)} = -\delta_1^{(1)} \tag{5}$$

$$\frac{\partial L}{\partial w_{11}^{(1)}} = \frac{\partial L}{\partial S_1^{(1)}} \frac{\partial S_1^{(1)}}{\partial w_{11}^{(1)}} = \delta_1^{(1)} x_1^{(0)} \tag{6}$$

$$\frac{\partial L}{\partial w_{21}^{(1)}} = \frac{\partial L}{\partial S_1^{(1)}} \frac{\partial S_1^{(1)}}{\partial w_{21}^{(1)}} = \delta_1^{(1)} x_2^{(0)} \tag{7}$$

For $j=2$

$$\frac{\partial L}{\partial b_2^{(1)}} = \frac{\partial L}{\partial w_{02}^{(1)}} = \frac{\partial L}{\partial S_2^{(1)}} \frac{\partial S_2^{(1)}}{\partial w_{02}^{(1)}} = \delta_2^{(1)} x_0^{(0)} = -\delta_2^{(1)} \tag{8}$$

$$\frac{\partial L}{\partial w_{12}^{(1)}} = \frac{\partial L}{\partial S_2^{(1)}} \frac{\partial S_2^{(1)}}{\partial w_{12}^{(1)}} = \delta_2^{(1)} x_1^{(0)} \tag{9}$$

$$\frac{\partial L}{\partial w_{22}^{(1)}} = \frac{\partial L}{\partial S_2^{(1)}} \frac{\partial S_2^{(1)}}{\partial w_{22}^{(1)}} = \delta_2^{(1)} x_2^{(0)} \tag{10}$$

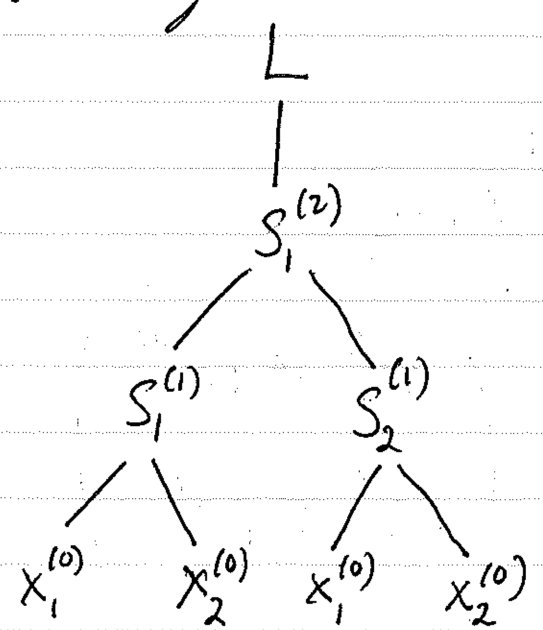
where $S_j^{(1)} = \sum_{i=1}^{2} w_{ij}^{(1)} x_i^{(0)} - b_j^{(1)}$ (11) and $\delta_j^{(1)} = \frac{\partial L}{\partial S_j^{(1)}}$, $j=1,2$ (12)

Can we generalize these results? <u>Yes!</u>

$$\boxed{\frac{\partial L}{\partial b_j^{(k)}} = -\delta_j^{(k)} \quad (13) \qquad \frac{\partial L}{\partial w_{ij}^{(k)}} = \delta_j^{(k)} x_i^{(k-1)} \quad (14)}$$

How do we calculate the $\delta_j^{(k)}$?

Dependency Tree for Chain Rule



$$\delta_1^{(1)} = \frac{\partial L}{\partial S_1^{(1)}} = \frac{\partial L}{\partial S_1^{(2)}} \frac{\partial S_1^{(2)}}{\partial S_1^{(1)}} = \delta_1^{(2)} \frac{\partial S_1^{(2)}}{\partial S_1^{(1)}} \quad (15)$$

Using $s_1^{(2)} = -w_{01}^{(2)} + w_{11}^{(2)} \varphi(s_1^{(1)}) + w_{21}^{(2)} \varphi(s_2^{(1)})$ (16)

$$\frac{\partial s_1^{(2)}}{\partial s_1^{(1)}} = w_{11}^{(2)} \frac{\partial \varphi(s_1^{(1)})}{\partial s_1^{(1)}} = w_{11}^{(2)} \varphi'(s_1^{(1)}) \quad (17)$$

$$\therefore \quad \delta_1^{(1)} = \delta_1^{(2)} w_{11}^{(2)} \varphi'(s_1^{(1)}) \quad (18)$$

Similarly

$$\delta_2^{(1)} = \frac{\partial L}{\partial s_2^{(1)}} = \frac{\partial L}{\partial s_1^{(2)}} \frac{\partial s_1^{(2)}}{\partial s_2^{(1)}} = \delta_1^{(2)} \frac{\partial s_1^{(2)}}{\partial s_2^{(1)}} \quad (19)$$

Using (16),

$$\delta_2^{(1)} = \delta_1^{(2)} w_{21}^{(2)} \varphi'(s_2^{(1)}) \quad (20)$$

$$\boxed{\begin{pmatrix} \delta_1^{(1)} \\ \delta_2^{(1)} \end{pmatrix} = \delta_1^{(2)} \begin{pmatrix} w_{11}^{(2)} \varphi'(s_1^{(1)}) \\ w_{21}^{(2)} \varphi'(s_2^{(1)}) \end{pmatrix} \quad (21)}$$

δ's in layer 1 depend on δ in layer 2!

How do we calculate $\delta_1^{(2)}$?

L is a function of y,

and $y = \varphi(s_1^{(2)})$!

Using (4) and the chain rule

$$\boxed{\delta_1^{(2)} = \frac{\partial L}{\partial s_1^{(2)}} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial s_1^{(2)}} = \frac{\partial L}{\partial y}\varphi'(s_1^{(2)}) \quad (22)}$$

To minimize L, apply gradient descent:

- Initialize weights and biases:

  $w_{ij}^{(k)}(0)$, $b_j^{(k)}(0)$

- Implement following recursion for iteration $n+1$

$$b_j^{(k)}(n+1) = b_j^{(k)}(n) - \eta \frac{\partial L}{\partial b_j^{(k)}}(w_{ij}^{(k)}(n), b_j^{(k)}(n)) \quad (23)$$

$$w_{ij}^{(k)}(n+1) = w_{ij}^{(k)}(n) - \eta \frac{\partial L}{\partial w_{ij}^{(k)}}(w_{ij}^{(k)}(n), b_j^{(k)}(n)) \quad (24)$$

Using (13) and (14)

$$\boxed{\begin{aligned} b_j^{(k)}(n+1) &= b_j^{(k)}(n) + \eta\,\delta_j^{(k)} && (25)\\ w_{ij}^{(k)}(n+1) &= w_{ij}^{(k)}(n) - \eta\,\delta_j^{(k)} x_i^{(k-1)}(n) && (26)\end{aligned}}$$

What about $\varphi'(s_j^{(k)})$?

Depends on activation function $\varphi(\cdot)$.

Assume $\varphi(x) = \sigma(x) = \dfrac{1}{1+e^{-x}}$ (27)

$$\varphi'(x) = \sigma'(x) = -\frac{-e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2} \quad (28)$$

$$= \frac{e^{-x}}{(1+e^{-x})}\,\frac{1}{1+e^{-x}} = \frac{e^{-x}}{1+e^{-x}}\,\sigma(x) \quad (29)$$

and $\dfrac{e^{-x}}{1+e^{-x}} = \dfrac{e^{-x}}{1+e^{-x}} + \dfrac{1}{1+e^{-x}} - \dfrac{1}{1+e^{-x}}$

$$= 1 - \sigma(x)$$

$$\therefore\ \sigma'(x) = \sigma(x)(1-\sigma(x))$$