

# 1 Gradient Descent

If we keep decreasing the  $\epsilon$  in our Finite Difference approach we effectively get the Derivative of the Cost Function.

$$C'(w) = \lim_{\epsilon \rightarrow 0} \frac{C(w + \epsilon) - C(w)}{\epsilon} \quad (1)$$

Let's compute the derivatives of all our models. Throughout the entire paper  $n$  means the amount of samples in the training set.

## 1.1 Linear Model



$$y = x \cdot w \quad (2)$$

### 1.1.1 Cost

$$C(w) = \frac{1}{n} \sum_{i=1}^n (x_i w - y_i)^2 \quad (3)$$

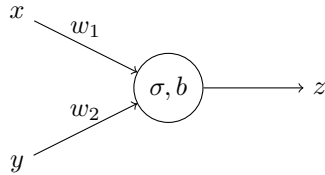
$$C'(w) = \left( \frac{1}{n} \sum_{i=1}^n (x_i w - y_i)^2 \right)' = \quad (4)$$

$$= \frac{1}{n} \left( \sum_{i=1}^n (x_i w - y_i)^2 \right)' \quad (5)$$

$$= \frac{1}{n} \sum_{i=1}^n ((x_i w - y_i)^2)' \quad (6)$$

$$= \frac{1}{n} \sum_{i=1}^n 2(x_i w - y_i)x_i \quad (7)$$

## 1.2 One Neuron Model with 2 inputs



$$z = \sigma(xw_1 + yw_2 + b) \quad (8)$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (9)$$

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)) \quad (10)$$

### 1.2.1 Cost

$$a_i = \sigma(x_iw_1 + y_iw_2 + b) \quad (11)$$

$$\partial_{w_1} a_i = \partial_{w_1} (\sigma(x_iw_1 + y_iw_2 + b)) = \quad (12)$$

$$= a_i(1 - a_i)\partial_{w_1} (x_iw_1 + y_iw_2 + b) = \quad (13)$$

$$= a_i(1 - a_i)x_i \quad (14)$$

$$\partial_{w_2} a_i = a_i(1 - a_i)y_i \quad (15)$$

$$\partial_b a_i = a_i(1 - a_i) \quad (16)$$

$$C = \frac{1}{n} \sum_{i=1}^n (a_i - z_i)^2 \quad (17)$$

$$\partial_{w_1} C = \frac{1}{n} \sum_{i=1}^n \partial_{w_1} ((a_i - z_i)^2) = \quad (18)$$

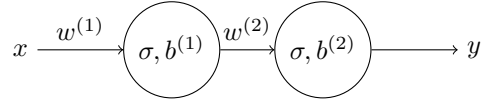
$$= \frac{1}{n} \sum_{i=1}^n 2(a_i - z_i)\partial_{w_1} a_i = \quad (19)$$

$$= \frac{1}{n} \sum_{i=1}^n 2(a_i - z_i)a_i(1 - a_i)x_i \quad (20)$$

$$\partial_{w_2} C = \frac{1}{n} \sum_{i=1}^n 2(a_i - z_i)a_i(1 - a_i)y_i \quad (21)$$

$$\partial_b C = \frac{1}{n} \sum_{i=1}^n 2(a_i - z_i)a_i(1 - a_i) \quad (22)$$

### 1.3 Two Neurons Model with 1 input



$$a^{(1)} = \sigma(xw^{(1)} + b^{(1)}) \quad (23)$$

$$y = \sigma(a^{(1)}w^{(2)} + b^{(2)}) \quad (24)$$

The superscript in parenthesis denotes the current layer. For example  $a_i^{(l)}$  denotes the activation from the  $l$ -th layer on  $i$ -th sample.

### 1.3.1 Feed-Forward

$$a_i^{(1)} = \sigma(x_i w^{(1)} + b^{(1)}) \quad (25)$$

$$\partial_{w^{(1)}} a_i^{(1)} = a_i^{(1)}(1 - a_i^{(1)})x_i \quad (26)$$

$$\partial_{b^{(1)}} a_i^{(1)} = a_i^{(1)}(1 - a_i^{(1)}) \quad (27)$$

$$a_i^{(2)} = \sigma(a_i^{(1)} w^{(2)} + b^{(2)}) \quad (28)$$

$$\partial_{w^{(2)}} a_i^{(2)} = a_i^{(2)}(1 - a_i^{(2)})a_i^{(1)} \quad (29)$$

$$\partial_{b^{(2)}} a_i^{(2)} = a_i^{(2)}(1 - a_i^{(2)}) \quad (30)$$

$$\partial_{a_i^{(1)}} a_i^{(2)} = a_i^{(2)}(1 - a_i^{(2)})w^{(2)} \quad (31)$$

### 1.3.2 Back-Propagation

$$C^{(2)} = \frac{1}{n} \sum_{i=1}^n (a_i^{(2)} - y_i)^2 \quad (32)$$

$$\partial_{w^{(2)}} C^{(2)} = \frac{1}{n} \sum_{i=1}^n \partial_{w^{(2)}} ((a_i^{(2)} - y_i)^2) = \quad (33)$$

$$= \frac{1}{n} \sum_{i=1}^n 2(a_i^{(2)} - y_i) \partial_{w^{(2)}} a_i^{(2)} = \quad (34)$$

$$= \frac{1}{n} \sum_{i=1}^n 2(a_i^{(2)} - y_i) a_i^{(2)} (1 - a_i^{(2)}) a_i^{(1)} \quad (35)$$

$$\partial_{b^{(2)}} C^{(2)} = \frac{1}{n} \sum_{i=1}^n 2(a_i^{(2)} - y_i) a_i^{(2)} (1 - a_i^{(2)}) \quad (36)$$

$$\partial_{a_i^{(1)}} C^{(2)} = \frac{1}{n} \sum_{i=1}^n 2(a_i^{(2)} - y_i) a_i^{(2)} (1 - a_i^{(2)}) w^{(2)} \quad (37)$$

$$e_i = a_i^{(1)} - \partial_{a_i^{(1)}} C^{(2)} \quad (38)$$

$$C^{(1)} = \frac{1}{n} \sum_{i=1}^n (a_i^{(1)} - e_i)^2 \quad (39)$$

$$\partial_{w^{(1)}} C^{(1)} = \partial_{w^{(1)}} \left( \frac{1}{n} \sum_{i=1}^n (a_i^{(1)} - e_i)^2 \right) = \quad (40)$$

$$= \frac{1}{n} \sum_{i=1}^n \partial_{w^{(1)}} ((a_i^{(1)} - e_i)^2) = \quad (41)$$

$$= \frac{1}{n} \sum_{i=1}^n 2(a_i^{(1)} - e_i) \partial_{w^{(1)}} a_i^{(1)} = \quad (42)$$

$$= \frac{1}{n} \sum_{i=1}^n 2(\partial_{a_i^{(1)}} C^{(2)}) a_i^{(1)} (1 - a_i^{(1)}) x_i \quad (43)$$

$$\partial_{b^{(1)}} C^{(1)} = \frac{1}{n} \sum_{i=1}^n 2(\partial_{a_i^{(1)}} C^{(2)}) a_i^{(1)} (1 - a_i^{(1)}) \quad (44)$$

## 1.4 Arbitrary Neurons Model with 1 input

Let's assume that we have  $m$  layers.

### 1.4.1 Feed-Forward

Let's assume that  $a_i^{(0)}$  is  $x_i$ .

$$a_i^{(l)} = \sigma(a_i^{(l-1)} w^{(l)} + b^{(l)}) \quad (45)$$

$$\partial_{w^{(l)}} a_i^{(l)} = a_i^{(l)} (1 - a_i^{(l)}) a_i^{(l-1)} \quad (46)$$

$$\partial_{b^{(l)}} a_i^{(l)} = a_i^{(l)} (1 - a_i^{(l)}) \quad (47)$$

$$\partial_{a_i^{(l-1)}} a_i^{(l)} = a_i^{(l)} (1 - a_i^{(l)}) w^{(l)} \quad (48)$$

#### 1.4.2 Back-Propagation

Let's denote  $a_i^{(m)} - y_i$  as  $\partial_{a_i^{(m)}} C^{(m+1)}$ .

$$C^{(l)} = \frac{1}{n} \sum_{i=1}^n (\partial_{a_i^{(l)}} C^{(l+1)})^2 \quad (49)$$

$$\partial_{w^{(l)}} C^{(l)} = \frac{1}{n} \sum_{i=1}^n 2(\partial_{a_i^{(l)}} C^{(l+1)}) a_i^{(l)} (1 - a_i^{(l)}) a_i^{(l-1)} = \quad (50)$$

$$\partial_{b^{(l)}} C^{(l)} = \frac{1}{n} \sum_{i=1}^n 2(\partial_{a_i^{(l)}} C^{(l+1)}) a_i^{(l)} (1 - a_i^{(l)}) \quad (51)$$

$$\partial_{a_i^{(l-1)}} C^{(l)} = \frac{1}{n} \sum_{i=1}^n 2(\partial_{a_i^{(l)}} C^{(l+1)}) a_i^{(l)} (1 - a_i^{(l)}) w^{(l)} \quad (52)$$