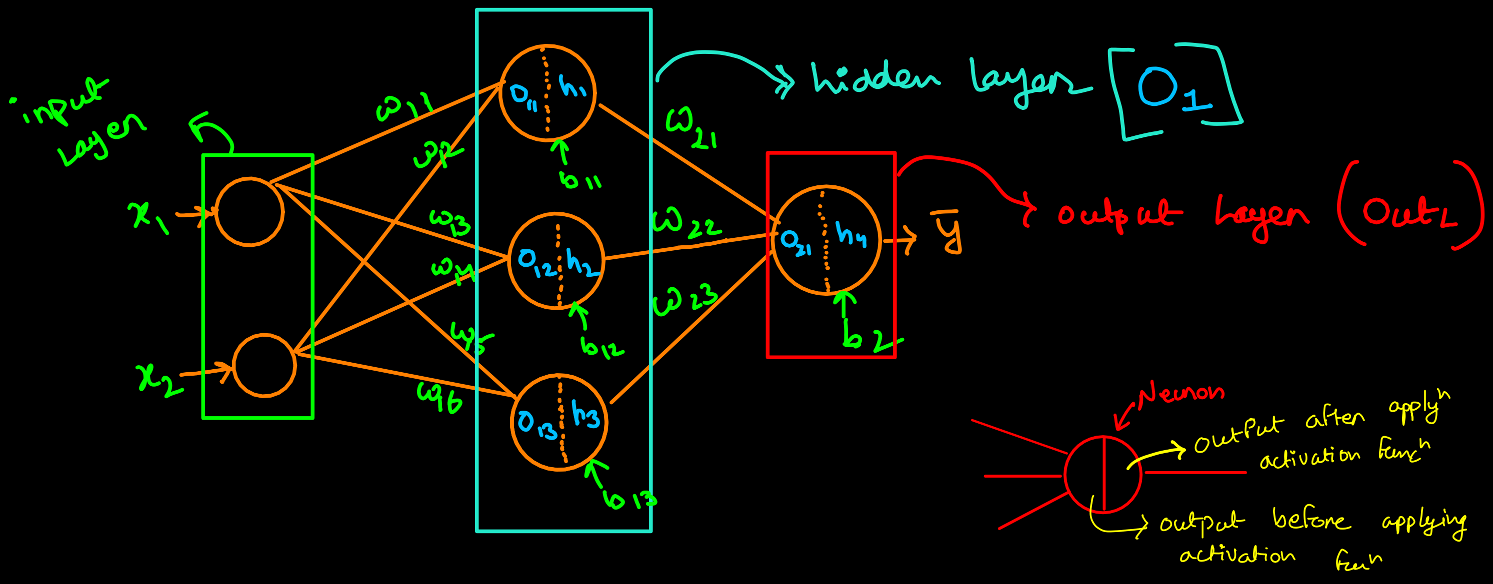
Neural Nets from Scratch using LA



To add non-linearity in the model we will be using sigmoid funcn

$$sig(x) = \frac{1}{1+e^{-x}}$$

$$O_{11} = x_1\omega_{11} + x_2\omega_{12} + b_{11}, \quad h_1 = \frac{1}{1+e^{-O_{11}}}$$

$$O_{12} = x_1\omega_{13} + x_2\omega_{14} + b_{12}, \quad h_2 = \frac{1}{1+e^{-O_{12}}}$$

$$O_{13} = x_1\omega_{15} + x_2\omega_{16} + b_{13}, \quad h_3 = \frac{1}{1+e^{-O_{13}}}$$

$$O_{21} = h_1\omega_{21} + h_2\omega_{22} + h_3\omega_{23} + b_2, \quad h_4 = \frac{1}{1+e^{-O_{21}}}$$

finally $\hat{y} = O_4$

Now, Lets do this using matrix algebra

In this, we are considering that Weight matrix and bias matrix will be different.

so,

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_{2\times 1}$$

input vect

Weight and bias matrices are denoted by W_1 & B_1 respectively so,

$$W_1 = \begin{bmatrix} \omega_{11} & \omega_{13} & \omega_{15} \\ \omega_{12} & \omega_{14} & \omega_{16} \end{bmatrix}_{2\times 3} \quad \& \quad B_1 = \begin{bmatrix} b_{11} \\ b_{12} \\ b_{13} \end{bmatrix}_{3\times 1}$$

The intermediate output of hidden layer is given by H_1

$$H_1 = W_1^T X + B_1 = \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{13} & \omega_{14} \\ \omega_{15} & \omega_{16} \end{bmatrix}_{3\times 2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_{2\times 1} + \begin{bmatrix} b_{11} \\ b_{12} \\ b_{13} \end{bmatrix}_{3\times 1}$$

$$\begin{bmatrix} x_1\omega_{11} + x_2\omega_{12} + b_{11} \\ x_1\omega_{13} + x_2\omega_{14} + b_{12} \\ x_1\omega_{15} + x_2\omega_{16} + b_{13} \end{bmatrix}_{3\times 1} \begin{matrix} \rightarrow O_{11} \\ \rightarrow O_{12} \\ \rightarrow O_{13} \end{matrix}$$

The final output of hidden layer is received by passing H_1 to sigmoid function and is denoted by O_1

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$$

$$O_1 = \sigma(H_1) = \begin{bmatrix} \sigma(x_1\omega_{11} + x_2\omega_{12} + b_{11}) \\ \sigma(x_1\omega_{13} + x_2\omega_{14} + b_{12}) \\ \sigma(x_1\omega_{15} + x_2\omega_{16} + b_{13}) \end{bmatrix}_{3\times 1} \begin{matrix} \rightarrow h_1 \\ \rightarrow h_2 \\ \rightarrow h_3 \end{matrix}$$

$$\begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}_{3\times 1}$$

Now, we are left with calculation of last layer.

Input of the next layer (last layer) is the output of the previous layer (hidden layer).
So, input of last layer is O_1.

Weight and bias matrices of last layer is denoted by W_2 and B_2
So

$$W_2 = \begin{bmatrix} w_{21} \\ w_{22} \\ w_{23} \end{bmatrix}_{3\times 1} \quad \text{and} \quad B_2 = \begin{bmatrix} b_2 \end{bmatrix}_{1\times 1}$$

this is 1×1 because, output layer have single neuron

$O_{21} \rightarrow$ intermediate output of last layer.
$h_4 \rightarrow$ Final output of last layer (i.e $\hat{y}$)
So,

$$O_{21} = W_2^T O_1 + B_2 = \underbrace{\begin{bmatrix} w_{21} & w_{22} & w_{23} \end{bmatrix}_{1\times 3}}_{W_2^T} \underbrace{\begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}_{3\times 1}}_{O_1} + \underbrace{\begin{bmatrix} b_2 \end{bmatrix}_{1\times 1}}_{B_2}$$

$$O_{21} = \begin{bmatrix} h_1 w_{11} + h_2 w_{22} + h_3 w_{23} + b_2 \end{bmatrix}_{1\times 1}$$

$$h_4 = \hat{y} = \sigma(O_{21}) = \sigma(h_1 w_{21} + h_2 w_{22} + h_3 w_{23} + b_2)$$

Writing the complete forward propagation of the network.

1. $O_1 = \sigma(\underset{3\times 2}{W_1^T} \underset{2\times 1}{X} + \underset{2\times 1}{B_1})_{3\times 1}$

2. $\underset{h_4}{O_2} = \sigma(\underset{1\times 3}{W_2^T} \underset{3\times 1}{O_1} + \underset{1\times 1}{B_2})$

Now, we will see how the weights and biases are updated using back-propagation algorithm.

In general, the simple way to update the weights are:-

$$\left\{\begin{array}{l} W = W - \alpha . \nabla_W L \\ B = B - \alpha . \nabla_B L \end{array}\right\}$$

$\alpha \rightarrow$ Learning rate

$\nabla_W L \rightarrow$ gradient of loss w.r.t to W

$\nabla_B L \rightarrow$ gradient of loss w.r.t to B

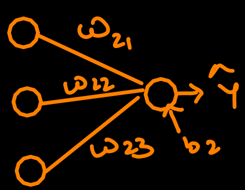
Before finding the gradient, we need to know the loss function.

Lets consider the simplest Loss function $\rightarrow$ mean sq. Error

$$MSE = \frac{1}{2}(\underbrace{Y}_{\text{actual outcome}} - \underbrace{\hat{Y}}_{\text{predicted outcome}})^2$$

this is the loss function that we are going to use.

Now, here the things get complicated, lets understand by taking individual weight & after this we will see how to do the same with matrix algebra.



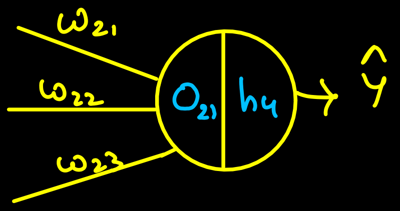
this is the same Network that we have seen above. We are going to update w_{21}, w_{22}, w_{23} & b_2.

$$\left.\begin{array}{l} w_{21} = w_{21} - \alpha . \nabla_{w_{21}} L \\ w_{22} = w_{22} - \alpha . \nabla_{w_{22}} L \\ w_{23} = w_{23} - \alpha . \nabla_{w_{23}} L \\ b_2 = b_2 - \alpha . \nabla_b L \end{array}\right\}$$

In this, $\nabla_{w_{ij}} L$ is nothing but partial derivative of L (Loss) w.r.t to w_{ij}

So, we need to find $\nabla_{w_{21}} L$, $\nabla_{w_{22}} L$, $\nabla_{w_{23}} L$ & $\nabla_b L$

<u>Note</u>



this was the structure of last layer of our network.

$$L = \frac{1}{2}(y-\hat{y})^2 = \frac{1}{2}\left(y - \sigma(h_1 w_{21} + h_2 w_{22} + h_3 w_{23} + b_2)\right)^2$$

Now,

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial L}{\partial h_4} \times \frac{\partial h_4}{\partial O_{21}} \times \frac{\partial O_{21}}{\partial w_{21}}$$ } using chain rule.

$$\rightarrow \frac{\partial L}{\partial h_4} = \frac{\partial}{\partial h_4} \frac{1}{2}\Big(y - \underbrace{\sigma(h_1 w_{21} + h_2 w_{22} + h_3 w_{23} + b_2)}_{h_4 / \hat{y}}\Big)^2$$

$$\boxed{\frac{\partial L}{\partial h_4} = -(y-\hat{y})}$$

$$\rightarrow \frac{\partial h_4}{\partial O_{21}} = \frac{\partial}{\partial O_{21}} \sigma(\underbrace{h_1 w_{21} + h_2 w_{22} + h_3 w_{23} + b_2}_{O_{21}})$$

$$= \frac{\partial}{\partial O_{21}}\left(\frac{1}{1+e^{-O_{21}}}\right) = \boxed{h_4(1-h_4)}$$

$$\rightarrow \frac{\partial O_{21}}{\partial w_{21}} = \frac{\partial}{\partial w_{21}}(h_1 w_{21} + h_2 w_{22} + h_3 w_{23} + b_2)$$

$$= \boxed{h_1}$$

$$\therefore \quad \boxed{\frac{\partial L}{\partial w_{21}} = -(y-\hat{y}) \cdot h_4(1-h_4) \cdot h_1}$$

In the same way, we can find $\nabla_{w_{ij}} L$ for other weights

$$\frac{\partial L}{\partial w_{22}} = -(y-\hat{y}) \cdot h_4(1-h_4) \cdot h_2$$

$$\frac{\partial L}{\partial w_{23}} = -(y-\hat{y}) \cdot h_4(1-h_4) \cdot h_3$$

$$\frac{\partial L}{\partial b_2} = -(y-\hat{y}) \cdot h_4 \,(1-h_4) \cdot 1$$

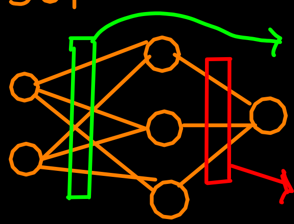
Now, once we get the $\nabla_{\omega_{ij}} L$ & $\nabla_b L$, we can use the below methods to update the weights.

$$\boxed{\omega = \omega - \alpha \cdot \nabla_\omega L} \quad \& \quad \boxed{b = b - \alpha \cdot \nabla_b L}$$

Till now, we have seen how to update the weights of last layer.

Now we are going to see how to update the weights of the hidden layer

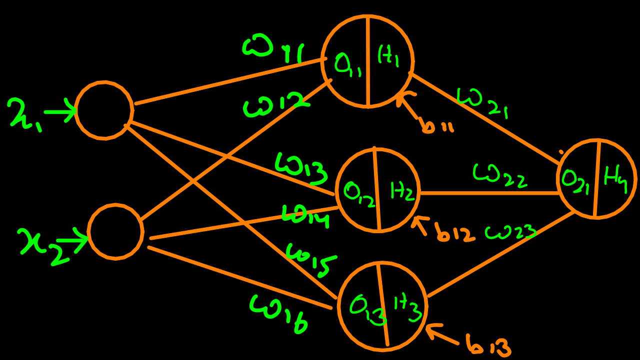
So,



we are going to update these weights

already seen the update rule of these weights.

So,



Lets see how to update ω_{11}.

So,

$$\boxed{\omega_{11} = \omega_{11} - \alpha \cdot \nabla_{\omega_{11}} L}$$

Now,

$$\boxed{\frac{\partial L}{\partial \omega_{11}} = \frac{\partial L}{\partial H_4} \cdot \frac{\partial H_4}{\partial O_{21}} \cdot \frac{\partial O_{21}}{\partial H_1} \cdot \frac{\partial H_1}{\partial O_{11}} \cdot \frac{\partial O_{11}}{\partial \omega_{11}}}$$

- $\frac{\partial L}{\partial H_4} = \frac{\partial}{\partial H_4} \frac{1}{2}(y-H_4)^2 = -(y-H_4)$

→ $\frac{\partial H_4}{\partial O_{21}} = \frac{\partial}{\partial O_{21}} \left(\frac{1}{1+e^{-O_{21}}}\right) = h_4(1-h_4)$

→ $\frac{\partial O_{21}}{\partial H_1} = -(H_1\omega_{21} + H_2\omega_{22} + H_3\omega_{33} + b_2) = \omega_{21}$

$\frac{\partial H_1}{\partial O_{11}} = \frac{\partial}{\partial O_{11}} \left(\frac{1}{1+e^{-O_{11}}}\right) = h_1 \cdot (1-h_1)$

$\frac{\partial O_{11}}{\partial \omega_{11}} = \frac{\partial}{\partial \omega_{11}} (x_1\omega_{11} + h_2\omega_{12} + b_1) = x_1$

So,

$$\boxed{\frac{\partial L}{\partial \omega_{11}} = -(y-H_4) \cdot H_4(1-H_4) \cdot \omega_{21} \cdot H_1(1-H_1) \cdot x_1}$$

we need to find the $\nabla \omega_{ij}$ for all other weights & biases so,

① $$\frac{\partial L}{\partial \omega_{12}} = -(y - H_4) \cdot H_2(1 - H_2) \cdot \omega_{21} \cdot H_1(1 - H_1) \cdot X_2$$

② $$\frac{\partial L}{\partial \omega_{13}} = -(y - H_4) \cdot H_2(1 - H_2) \cdot \omega_{22} \cdot H_2(1 - H_2) \cdot X_1$$

③ $$\frac{\partial L}{\partial \omega_{14}} = -(y - H_4) \cdot H_2(1 - H_2) \cdot \omega_{22} \cdot H_2(1 - H_2) \cdot X_2$$

④ $$\frac{\partial L}{\partial \omega_{15}} = -(y - H_4) \cdot H_2(1 - H_2) \cdot \omega_{23} \cdot H_3(1 - H_3) \cdot X_1$$

⑤ $$\frac{\partial L}{\partial \omega_{16}} = -(y - H_4) \cdot H_2(1 - H_2) \cdot \omega_{23} \cdot H_3(1 - H_3) \cdot X_2$$

⑥ $$\frac{\partial L}{\partial b_{11}} = -(y - H_4) \cdot H_2(1 - H_2) \cdot \omega_{21} \cdot H_1(1 - H_1)$$

⑦ $$\frac{\partial L}{\partial b_{12}} = -(y - H_4) \cdot O_{21}(1 - O_{21}) \cdot \omega_{22} \cdot H_2(1 - H_2)$$

⑧ $$\frac{\partial L}{\partial b_{13}} = -(y - H_4) \cdot O_{21}(1 - O_{21}) \cdot \omega_{23} \cdot H_3(1 - H_3)$$

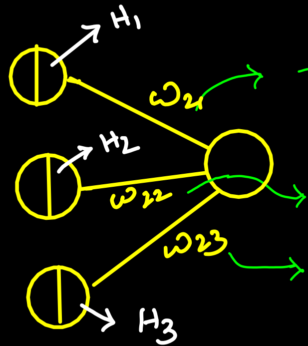
As, we got the $\nabla_{\omega_{ij}} L$ & $\nabla_{b_{ij}} L$ for all the weights and biases, now we just need to update the parameters using

↳ $$\boxed{\omega = \omega - \alpha \cdot \nabla_\omega L \quad \& \quad b = b - \alpha \cdot \nabla_b L}$$

In general, Neural Nets architecture are very complex & finding gradient for each weight will be tidious, so we are going to use LA to do the same task of updating weights.

Lets first update W_2 & B_2 using matrix algebra.

$$W_2 = \begin{bmatrix} w_{21} \\ w_{22} \\ w_{23} \end{bmatrix} \quad \& \quad B_2 = \begin{bmatrix} b_2 \end{bmatrix}$$



$\rightarrow -(y - H_4) \cdot H_4(1 - H_4) \cdot H_1$

$\rightarrow -(y - H_4) \cdot H_4(1 - H_4) \cdot H_2$

$\rightarrow -(y - H_4) \cdot H_4(1 - H_4) \cdot H_3$

} gradient of respective weights.

We can see that all the weight's gradient have a common term

$\hookrightarrow -(y - H_4) \cdot H_4(1 - H_4) \rightarrow$ we will use λ to denote this

Now, lets find out its dimension.

$\hookrightarrow$ we have single output neuron

$\hookrightarrow y \rightarrow 1 \times 1$, $H_4 / \hat{y} \rightarrow 1 \times 1$

$H_4 \rightarrow$ is also a scalar value $\rightarrow \therefore H_4 \rightarrow 1 \times 1$

So, $-(y_{1\times1} - H_{4\,1\times1}) \cdot H_{4\,1\times1}(1 - H_{4\,1\times1}) \rightarrow \lambda_{1\times1}$

Now, output of previous layer can be stored in a matrix

$$\hookrightarrow O_1 = \begin{bmatrix} H_1 \\ H_2 \\ H_3 \end{bmatrix}_{3\times1}$$

Now,

$$\nabla_{w_2} L = [\lambda]_{1\times1} \cdot O_1^T = [\lambda]_{1\times1} \cdot \begin{bmatrix} H_1 & H_2 & H_3 \end{bmatrix}_{1\times3} \rightarrow \begin{bmatrix} \lambda H_1 & \lambda H_2 & \lambda H_3 \end{bmatrix}_{1\times3}$$

$\hookrightarrow$ dot product

Finally,

$$W_2 = W_2 - \alpha \cdot (\nabla_{w_2} L)^T \rightarrow \begin{bmatrix} w_{21} \\ w_{22} \\ w_{23} \end{bmatrix} = \begin{bmatrix} w_{21} \\ w_{22} \\ w_{23} \end{bmatrix} - \alpha \cdot \begin{bmatrix} \lambda H_1 \\ \lambda H_2 \\ \lambda H_3 \end{bmatrix}$$

$$B_2 = B_2 - \alpha \cdot [\lambda]$$

Now. Lets update W_1 & B_1 using matrix algebra. This part is bit complex

$$\frac{\partial L}{\partial \omega_{11}} = -(y - H_4) \cdot H_4(1 - H_4) \cdot \omega_{21} \cdot H_1(1 - H_1) \cdot x_1$$

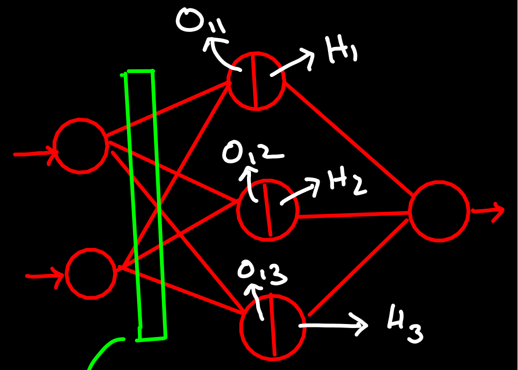
$$\frac{\partial L}{\partial \omega_{12}} = -(y - H_4) \cdot H_2(1 - H_2) \cdot \omega_{21} \cdot H_1(1 - H_1) \cdot x_2$$

$$\frac{\partial L}{\partial \omega_{13}} = -(y - H_4) \cdot H_2(1 - H_2) \cdot \omega_{22} \cdot H_2(1 - H_2) \cdot x_1$$

$$\frac{\partial L}{\partial \omega_{14}} = -(y - H_4) \cdot H_2(1 - H_2) \cdot \omega_{22} \cdot H_2(1 - H_2) \cdot x_2$$

$$\frac{\partial L}{\partial \omega_{15}} = -(y - H_4) \cdot H_2(1 - H_2) \cdot \omega_{23} \cdot H_3(1 - H_3) \cdot x_1$$

$$\frac{\partial L}{\partial \omega_{16}} = \underbrace{-(y - H_4) \cdot H_2(1 - H_2)}_{\lambda_{1\times1}} \cdot \omega_{23} \cdot H_3(1 - H_3) \cdot x_2$$



we are going to update these weights

We have,

$$W_1 = \begin{bmatrix} \omega_{11} & \omega_{13} & \omega_{15} \\ \omega_{12} & \omega_{14} & \omega_{16} \end{bmatrix}_{2\times3}, \quad W_2 = \begin{bmatrix} \omega_{21} \\ \omega_{22} \\ \omega_{23} \end{bmatrix}_{3\times1}, \quad P = \begin{bmatrix} H_1 \\ H_2 \\ H_3 \end{bmatrix}_{3\times1}, \quad X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_{2\times1}$$

Now

① $M_{1\times3} = [\lambda]_{1\times1} \cdot W_2^T \Rightarrow [\lambda]_{1\times1} \cdot [\omega_{21} \quad \omega_{22} \quad \omega_{23}]_{1\times3}$

$$\rightarrow [\lambda\omega_{21} \quad \lambda\omega_{22} \quad \lambda\omega_{23}]_{1\times3} \xrightarrow{\text{transpose}} \begin{bmatrix} \lambda\omega_{21} \\ \lambda\omega_{22} \\ \lambda\omega_{23} \end{bmatrix}_{3\times1}$$

Hadamard product

② $$M^T \odot P = \begin{bmatrix} \lambda\omega_{21} \\ \lambda\omega_{22} \\ \lambda\omega_{23} \end{bmatrix} \odot \begin{bmatrix} H_1 \\ H_2 \\ H_3 \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} \lambda \cdot \omega_{21} \cdot H_1 \\ \lambda \cdot \omega_{22} \cdot H_2 \\ \lambda \cdot \omega_{23} \cdot H_3 \end{bmatrix}_{3\times1}$$

③ $$Z = M^T \odot P \odot (1 - P)$$

$$Z = \begin{bmatrix} \lambda \cdot \omega_{21} \cdot H_1 \\ \lambda \cdot \omega_{22} \cdot H_2 \\ \lambda \cdot \omega_{23} \cdot H_3 \end{bmatrix}_{3\times1} \odot \begin{bmatrix} 1-H_1 \\ 1-H_2 \\ 1-H_3 \end{bmatrix}_{3\times1} = \begin{bmatrix} \boxed{\lambda \cdot \omega_{21} \cdot H_1 \cdot (1-H_1)} \\ \boxed{\lambda \cdot \omega_{22} \cdot H_2 \cdot (1-H_2)} \\ \boxed{\lambda \cdot \omega_{23} \cdot H_3 \cdot (1-H_3)} \end{bmatrix}_{3\times1} \begin{array}{l} \rightarrow Z_{11} \\ \rightarrow Z_{12} \\ \rightarrow Z_{13} \end{array}$$

(4) $\nabla_{\omega_1} L = Z_{3\times1} \cdot X^T_{1\times2}$

$$\begin{bmatrix} Z_{11} \\ Z_{12} \\ Z_{13} \end{bmatrix}_{3\times1} \cdot \begin{bmatrix} x_1 & x_2 \end{bmatrix}_{1\times2} = \begin{bmatrix} Z_{11}x_1 & Z_{11}x_2 \\ Z_{12}x_1 & Z_{12}x_2 \\ Z_{13}x_1 & Z_{13}x_2 \end{bmatrix}_{3\times2}$$

$\therefore$ Dimension of $\nabla_{\omega_1} L$ is 3×2

Finally,

$$\boxed{\omega_1 = \omega_1 - \alpha \cdot (\nabla_{\omega_1} L)^T}$$

$$\begin{bmatrix} \omega_{11} & \omega_{13} & \omega_{15} \\ \omega_{12} & \omega_{14} & \omega_{16} \end{bmatrix} = \begin{bmatrix} \omega_{11} & \omega_{13} & \omega_{15} \\ \omega_{12} & \omega_{14} & \omega_{16} \end{bmatrix} - \alpha \cdot \begin{bmatrix} Z_{11}x_1 & Z_{12}x_1 & Z_{13}x_1 \\ Z_{11}x_2 & Z_{12}x_2 & Z_{13}x_2 \end{bmatrix}$$

and

$$\boxed{B_{1\,(3\times1)} = B_{1\,(3\times1)} - \alpha \cdot Z_{3\times1}}$$

$$\begin{bmatrix} b_{11} \\ b_{12} \\ b_{13} \end{bmatrix}_{3\times1} = \begin{bmatrix} b_{11} \\ b_{12} \\ b_{13} \end{bmatrix}_{3\times1} - \alpha \cdot \begin{bmatrix} Z_{11} \\ Z_{12} \\ Z_{13} \end{bmatrix}_{3\times1}$$

Note, $\nabla_{\omega_1} L$ can also be written in the following maner for understanding purpose

Dot product

Dot product

$$\underset{3\times2}{\nabla_{\omega_1} L} = \left[(\lambda_{1\times1} \cdot \omega^T_{2\,(1\times3)})_{1\times3} \odot P_{3\times1} \odot (1-P)_{3\times1} \right]_{3\times1} \cdot X^T_{1\times2}$$

Hadamard product