Backward Propagation

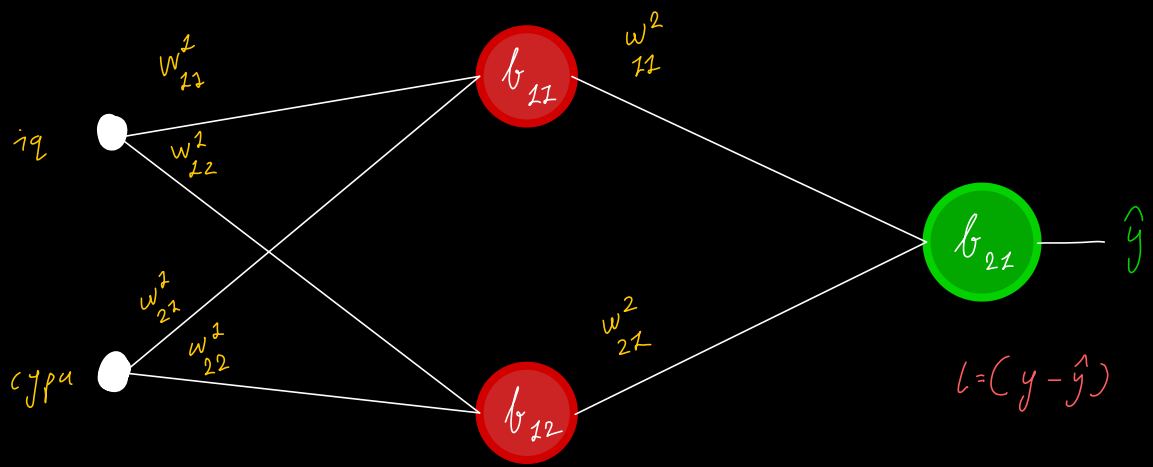
→ Backpropagation "Backward Propagation of errors", is an algorithm for supervised learning of artificial neural network using gradient descent. given an artificial neural network & an error function, the method calculates the gradient of error function with respect to the neural network's weights.

in-simple

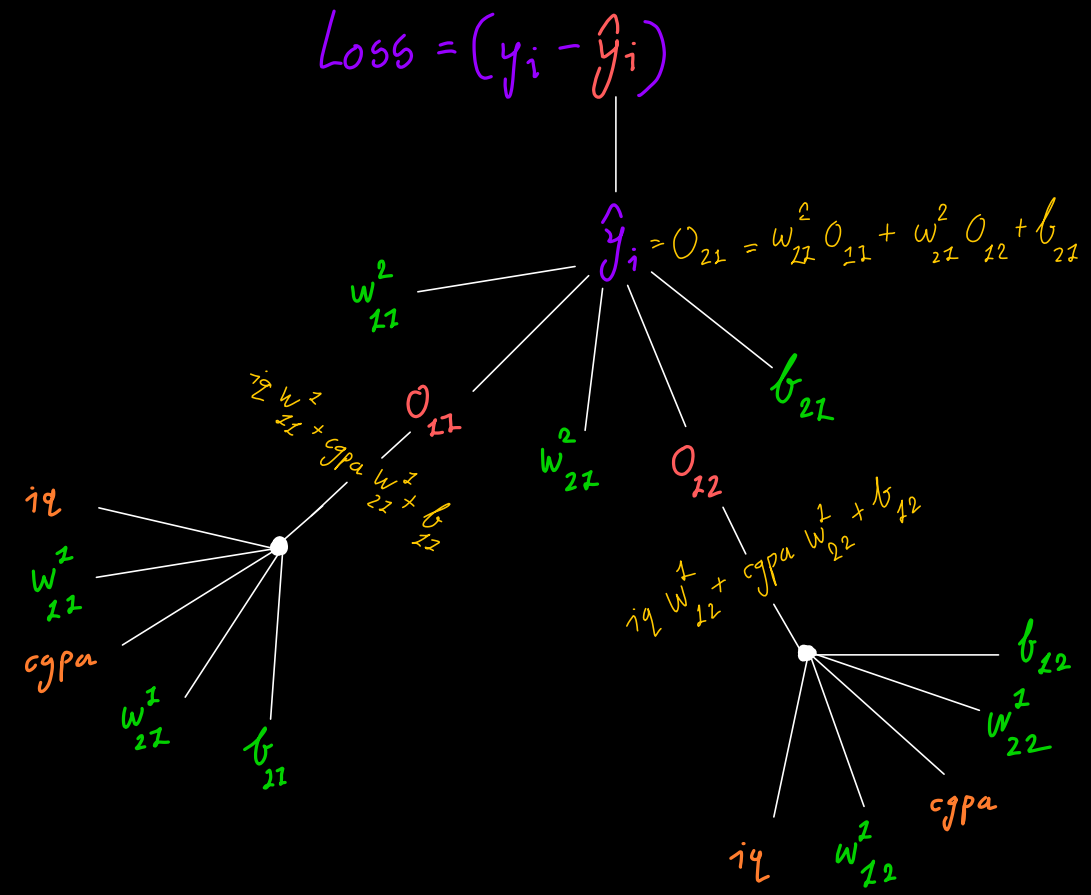
↳ An algorithm to train neural-network.

Steps

1. initialize weights & biases.
2. Select a data row & feed it to network.
3. Make prediction.
4. Select loss function & calculate loss.
5. Adjust weights & biases based on calculated error.



Let's see how all these parameter depends to each other



→ As above dependency tree shows, in order to change loss we need to change $\hat{y}$

→ To change $\hat{y}$ change w^2_{11}, w^2_{21}, O_{11}, O_{12} and b_{21}.

→ Since w^2_{11}, w^2_{21} and b_{21} are indipendent, we can change the value of those weight

→ in order to change O_{11}, we need to update iq, w^1_{11}, cgpa, w^1_{21} and b_{11}

→ in-order to change O_{12}, we need to update iq, w^1_{12}, cgpa, w^1_{22} and b_{12}.

→ Since iq and cgpa are actual i/p data, we can't update it hence remaind parameters w^1_{11}, w^1_{12}, w^1_{21}, w^1_{22}, b_{11} and b_{12} will get update.

formulas for weights & bias updation

for weights

$$W_{new} = W_{old} - \alpha \frac{\partial L}{\partial W_{old}}$$

for bias

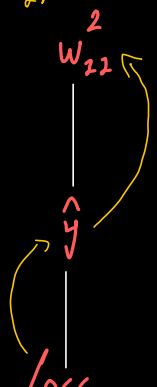
$$b_{new} = b_{old} - \alpha \frac{\partial L}{\partial b_{old}}$$

Total number of derivatives, we need to find....

$$\frac{\partial L}{\partial w^2_{11}} \quad \frac{\partial L}{\partial w^2_{21}} \quad \frac{\partial L}{\partial b_{21}} \quad \Big| \quad \frac{\partial L}{\partial w^2_{12}} \quad \frac{\partial L}{\partial w^2_{22}} \quad \frac{\partial L}{\partial b_{12}} \quad \Big| \quad \frac{\partial L}{\partial w^1_{11}} \quad \frac{\partial L}{\partial w^1_{22}} \quad \frac{\partial L}{\partial b_{12}}$$

derivative $\frac{dy}{dx}$ says that how 'y' is changing with respect to 'x'.

$\frac{\partial L}{\partial w^2_{11}}$: How Loss is changing with respect to w^2_{11}.



$$\frac{\partial L}{\partial w^2_{11}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w^2_{11}}$$

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

$$\frac{\partial L}{\partial \hat{y}} = -2(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial w^2_{11}} = \frac{\partial}{\partial w^2_{11}} (O_{11} w^2_{11} + O_{12} w^2_{22} + b_{21})$$

$$\frac{\partial \hat{y}}{\partial w^2_{11}} = O_{11}$$

$$\boxed{\frac{\partial L}{\partial w^2_{11}} = -2(y - \hat{y}) O_{11}}$$

$$\frac{\partial L}{\partial w_{22}^2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial w_{22}^2}$$

$$\frac{\partial L}{\partial \hat{y}} = \boxed{-2(y-\hat{y})}$$

$$\frac{\partial \hat{y}}{\partial w_{22}^2} = \frac{\partial}{\partial w_{22}^2}\left(o_{21} w_{21}^2 + o_{22} w_{22}^2 + b_{22}\right)$$

$$\frac{\partial \hat{y}}{\partial w_{22}^2} = \boxed{o_{22}}$$

$$\boxed{\frac{\partial L}{\partial w_{22}^2} = -2(y-\hat{y})\, o_{22}}$$

$$\frac{\partial L}{\partial b_{22}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial b_{22}}$$

$$\frac{\partial L}{\partial \hat{y}} = \boxed{-2(y-\hat{y})} \qquad \frac{\partial \hat{y}}{\partial b_{22}} = \frac{\partial}{\partial b_{22}}\left(o_{21} w_{21}^2 + o_{22} w_{22}^2 + b_{22}\right)$$

$$\frac{\partial \hat{y}}{\partial b_{22}} = \boxed{1}$$

$$\boxed{\frac{\partial L}{\partial b_{22}} = -2(y-\hat{y})}$$

Now we will calculate next derivatives.

$$\frac{\partial L}{\partial w^2_{11}} \quad \frac{\partial L}{\partial w^2_{22}} \quad \frac{\partial L}{\partial b_{11}}$$

$$\frac{\partial L}{\partial w^2_{11}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{11}} \frac{\partial O_{11}}{\partial w^2_{11}} \qquad \frac{\partial L}{\partial w^2_{12}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{12}} \frac{\partial O_{12}}{\partial w^2_{12}}$$

$$\frac{\partial L}{\partial w^2_{21}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{11}} \frac{\partial O_{11}}{\partial w^2_{21}} \qquad \frac{\partial L}{\partial w^2_{22}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{12}} \frac{\partial O_{12}}{\partial w^2_{22}}$$

$$\frac{\partial L}{\partial b_{11}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{11}} \frac{\partial O_{11}}{\partial b_{11}} \qquad \frac{\partial L}{\partial b_{12}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{12}} \frac{\partial O_{12}}{\partial b_{12}}$$

Already known

$$\frac{\partial L}{\partial \hat{y}} = -2(y - \hat{y})$$

Not known yet....

$$\frac{\partial \hat{y}}{\partial O_{11}} \quad \frac{\partial \hat{y}}{\partial O_{12}}$$

$$\frac{\partial \hat{y}}{\partial O_{11}} = \frac{\partial}{\partial O_{11}} \left[O_{11} w^2_{11} + O_{12} w^2_{21} + b_{21} \right]$$

$$\frac{\partial \hat{y}}{\partial O_{11}} = \boxed{w^2_{11}}$$

$$\frac{\partial \hat{y}}{\partial O_{12}} = \frac{\partial}{\partial O_{12}} \left[O_{11} w^2_{11} + O_{12} w^2_{21} + b_{21} \right]$$

$$\frac{\partial \hat{y}}{\partial O_{12}} = \boxed{w^2_{21}}$$

$$\frac{\partial O_{11}}{\partial w_{11}^1} = \frac{\partial}{\partial w_{11}^1}\left[iq\, w_{11}^1 + cgpa\, w_{21}^1 + b_{11}\right]$$

$$\frac{\partial O_{11}}{\partial w_{11}^1} = \boxed{iq}\ (x_{i1})$$

$$\frac{\partial O_{11}}{\partial w_{21}^1} = \frac{\partial}{\partial w_{21}^1}\left[iq\, w_{11}^1 + cgpa\, w_{21}^1 + b_{11}\right]$$

$$\frac{\partial O_{11}}{\partial w_{21}^1} = \boxed{cgpa}\ (x_{i2})$$

$$\frac{\partial O_{11}}{\partial b_{11}} = \frac{\partial}{\partial b_{11}}\left[iq\, w_{11}^1 + cgpa\, w_{21}^1 + b_{11}\right]$$

$$\frac{\partial O_{11}}{\partial b_{11}} = \boxed{1}$$

$$\frac{\partial O_{12}}{\partial w_{12}^1} = \frac{\partial}{\partial w_{12}^1}\left[iq\, w_{12}^1 + cgpa\, w_{22}^1 + b_{12}\right]$$

$$\frac{\partial O_{12}}{\partial w_{12}^1} = \boxed{iq}\ (x_{i1})$$

$$\frac{\partial O_{12}}{\partial w_{22}^1} = \frac{\partial}{\partial w_{22}^1}\left[iq\, w_{12}^1 + cgpa\, w_{22}^1 + b_{12}\right]$$

$$\frac{\partial O_{12}}{\partial w_{22}^1} = \boxed{cgpa}\ (x_{i2})$$

$$\frac{\partial O_{12}}{\partial b_{12}} = \frac{\partial}{\partial b_{12}}\left[iq\, w_{12}^1 + cgpa\, w_{22}^1 + b_{12}\right]$$

$$\frac{\partial O_{12}}{\partial b_{12}} = \boxed{1}$$

Put them all together

1. $$\frac{\partial L}{\partial w_{11}^2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{11}} \frac{\partial O_{11}}{\partial w_{11}^2}$$

$$= \boxed{-2(y - \hat{y})\, w_{11}^2\, x_{i1}}$$

2. $$\frac{\partial L}{\partial w_{22}^1} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{22}} \frac{\partial O_{22}}{\partial w_{22}^1}$$

$$-2(y-\hat{y})\, w_{11}^2\, x_{i2}$$

3. $$\frac{\partial L}{\partial b_{22}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{22}} \frac{\partial O_{22}}{\partial b_{22}}$$

$$-2(y-\hat{y})\, w_{11}^2$$

4. $$\frac{\partial L}{\partial w_{22}^1} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{12}} \frac{\partial O_{12}}{\partial w_{12}^2}$$

$$-2(y-\hat{y})\, w_{21}^2\, x_{i2}$$

5. $$\frac{\partial L}{\partial w_{22}^1} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{12}} \frac{\partial O_{12}}{\partial w_{22}^2}$$

$$-2(y-\hat{y})\, w_{21}^2\, x_{i2}$$

$$\frac{\partial L}{\partial b_{12}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{12}} \frac{\partial O_{12}}{\partial b_{12}}$$

$$-2(y-\hat{y})\, w_{21}^2$$

How everything works together

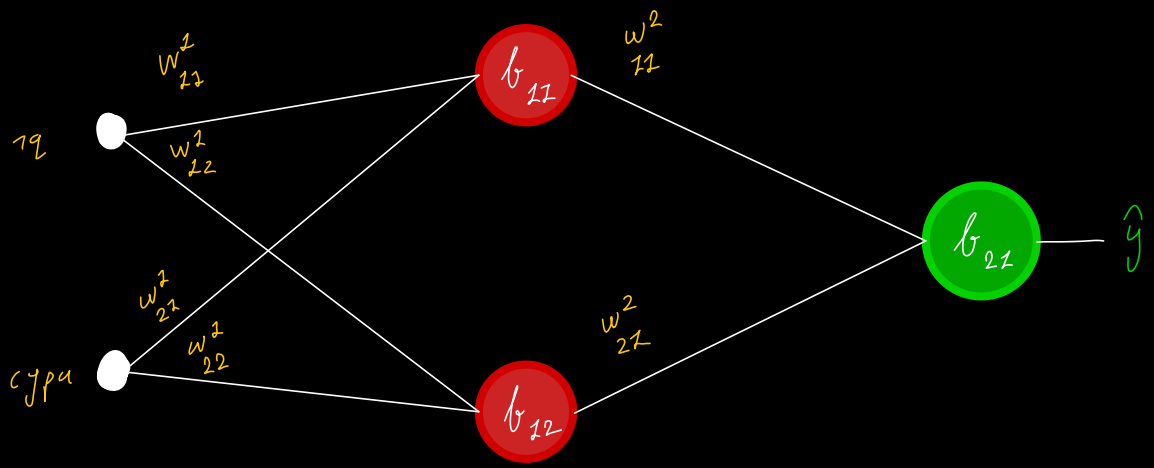
epoch = 10

```
for i in range(epoch):
    for j in range(x.shape[0]):
```

1. Select 1 row randomly
2. forward propagation & prediction
3. Calculate loss
4. Update weights & biases using G.D.

Calculate average loss for each epochs

Let's see the same thing in classification task.



$$Loss = -y \log(\hat{y}) - (1-y)\log(1-\hat{y})$$

$$\hat{y} \quad (O_{11} w^2_{11} + O_{12} w^2_{21} + b_{21})$$

z_{11} w^2_{11} z_{12} w^2_{21} b_{21}

$(x_{i1} w^1_{11} + x_{i2} w^1_{21} + b_{11})$ $(x_{i1} w^1_{12} + x_{i2} w^1_{22} + b_{12})$

x_{i1} w^1_{11} x_{i2} w^1_{21} b_{11} x_{i1} w^1_{12} x_{i2} w^1_{22} b_{12}

$$\frac{\partial L}{\partial w_{21}^2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial w_{21}^2} \qquad \frac{\partial L}{\partial w_{21}^2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial w_{21}^2}$$

$$\frac{\partial L}{\partial b_{21}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial b_{21}}$$

$$\frac{\partial L}{\partial \hat{y}} = \frac{\partial}{\partial \hat{y}} \left[-y \log(\hat{y}) - (1-y) \log(1-\hat{y}) \right]$$

$$= \frac{-y}{\hat{y}} + \frac{(1-y)}{(1-\hat{y})}$$

$$= \frac{-y(1-\hat{y}) + \hat{y}(1-y)}{\hat{y}(1-\hat{y})}$$

$$= \frac{-y + y\hat{y} + \hat{y} - y\hat{y}}{\hat{y}(1-\hat{y})}$$

$$\frac{\partial L}{\partial \hat{y}} = \boxed{\frac{-(y-\hat{y})}{\hat{y}(1-\hat{y})}}$$

$$\frac{\partial \hat{y}}{\partial z} = \frac{\partial}{\partial z} (\sigma(z))$$

$$= \sigma(z)[1 - \sigma(z)]$$

$$\hat{y}$$

$$\frac{\partial \hat{y}}{\partial z} = \boxed{\hat{y}(1-\hat{y})}$$

$$\frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z} = \frac{-(y-\hat{y})}{\hat{y}(1-\hat{y})} \; \hat{y}(1-\hat{y})$$

$$= \boxed{-(y-\hat{y})}$$

$$\frac{\partial z}{\partial w_{11}^2} = \frac{\partial}{\partial w_{11}^2}\left[O_{11}w_{11}^2 + O_{12}w_{21}^2 + b_{21}\right] \qquad \frac{\partial z}{\partial w_{21}^2} = \frac{\partial}{\partial w_{21}^2}\left[O_{11}w_{11}^2 + O_{12}w_{21}^2 + b_{21}\right]$$

$$= \boxed{O_{11}} \qquad = \boxed{O_{12}}$$

$$\frac{\partial z}{\partial b_{21}} = \frac{\partial}{\partial b_{21}}\left[O_{11}w_{11}^2 + O_{12}w_{21}^2 + b_{21}\right]$$

$$= \boxed{1}$$

$$\frac{\partial L}{\partial w_{11}^2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial w_{11}^2} \qquad \frac{\partial L}{\partial w_{21}^2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial w_{21}^2}$$

$$= \boxed{-(y-\hat{y})\, O_{11}} \qquad = \boxed{-(y-\hat{y})\, O_{12}}$$

$$\frac{\partial L}{\partial b_{21}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial b_{21}}$$

$$= \boxed{-(y-\hat{y})}$$

$$\frac{\partial L}{\partial w_{11}^2}, \quad \frac{\partial L}{\partial w_{22}^2}, \quad \frac{\partial L}{\partial b_{21}}$$

$$z_{fin} = O_{11}w_{11}^2 + O_{12}w_{21}^2 + b_{21}$$

$$O_{11} = \sigma(z_{prev})$$

$$\frac{\partial L}{\partial w_{11}^1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_{fin}} \cdot \frac{\partial z_{fin}}{\partial O_{11}} \cdot \frac{\partial O_{11}}{\partial z_{prev}} \frac{\partial O_{11}}{\partial w_{11}^1}$$

$$\frac{\partial z_{fin}}{\partial o_{11}} = \boxed{w_{11}^2} \qquad \frac{\partial o_{11}}{\partial z_{prev}} = \boxed{o_{11}(1-o_{11})}$$

$$\frac{\partial z_{prev}}{\partial w_{11}^1} = \frac{\partial}{\partial w_{11}^1}\left[w_{11}^1 x_{i1} + w_{21}^1 x_{i2} + b_{11}\right]$$

$$= \boxed{x_{i1}}$$

$$\frac{\partial L}{\partial w_{11}^1} = \boxed{-(y-\hat{y})\, w_{11}^2\, o_{11}(1-o_{11})\, x_{i1}}$$

$$\frac{\partial L}{\partial w_{21}^1} = \boxed{-(y-\hat{y})\, w_{11}^2\, o_{11}(1-o_{11})\, x_{i2}}$$

$$\frac{\partial L}{\partial w_{12}^2} \qquad \frac{\partial L}{\partial w_{22}^1} \qquad \frac{\partial L}{\partial b_{12}}$$

$$\frac{\partial L}{\partial w_{12}^2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_{fin}} \frac{\partial z_{fin}}{\partial o_{12}} \frac{\partial o_{12}}{\partial z_{prev}} \frac{\partial z_{prev}}{\partial w_{12}^1}$$

$$\frac{\partial z_{fin}}{\partial o_{12}} \to w_{21}^2 \qquad \frac{\partial o_{12}}{\partial z_{prev}} \to o_{12}(1-o_{12}) \qquad \frac{\partial z_{prev}}{\partial w_{12}^1} \to x_{i1}$$

$$\frac{\partial L}{\partial w_{12}^1} = \boxed{-(y-\hat{y})\, w_{21}^2\, o_{12}(1-o_{12})\, x_{i1}}$$

$$\frac{\partial L}{\partial w_{22}^1} = \boxed{-(y-\hat{y})\, w_{21}^2\, o_{12}(1-o_{12})\, x_{i2}}$$

$$\frac{\partial L}{\partial b_{12}} = \boxed{-(y-\hat{y})\, w_{21}^2\, o_{12}(1-o_{12})}$$

Q. Why this works.....?

$y = f(x)$

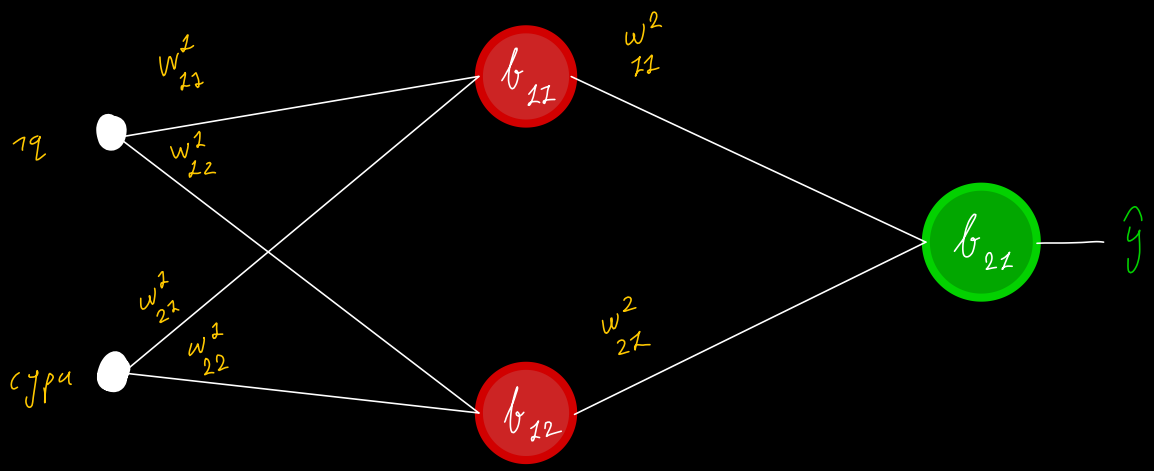
-> y is the function of 'x'.

$$L = (y - \hat{y})$$

-> So, L is the function of whom..?

-> y is coming from data, hence it is constant

-> What is this $\hat{y}$....?



$$\hat{y} = w^2_{11} O_{11} + w^2_{21} O_{12} + b_{21}$$

-> Here w^2_{11} and w^2_{21} are constant, what about O_{11} and O_{12}

$$O_{11} = x_{i1} w^1_{11} + x_{i2} w^1_{21} + b_{11}$$

$$O_{12} = x_{i1} w^1_{12} + x_{i2} w^1_{22} + b_{12}$$

-> Here x_{i1} and x_{i2} are input data-points iq and cgpa.

Hence...

$$\hat{y} = w^2_{11}\left[x_{i1} w^1_{11} + x_{i2} w^1_{21} + b_{11}\right] + w^2_{21}\left[x_{i1} w^1_{12} + x_{i2} w^1_{22} + b_{12}\right] + b_{21}$$

-> That's why L is function of $\hat{y}$.

$L(\hat{y})$

-> We can say that loss function is the function of all trainable parameters. ($w^1_{11}, w^2_{12}, w^1_{21}, w^1_{22}, b_{11}, b_{12}, W^2_{11}, W^2_{21}, b_{21}$)

Concept of gradient

-> A nickname of derivative.

$y = f(x) = x^2 + x$ task : Differentiate 'y' with-respect-to 'x'.

$$\frac{dy}{dx} = \frac{d}{dx} f(x) = \frac{d}{dx}(x^2 + x) = \boxed{2x + 1}$$

Derivative VS Gradient

Derivative : The o/p of a function depends on only 1 parameter

exc

$y = f(x)$

$y \longrightarrow x \quad \frac{d}{dx}$

Gradient : The o/p of a function depends on >1 parameters.

exc

$y = f(x, z) = x^2 + z^2$

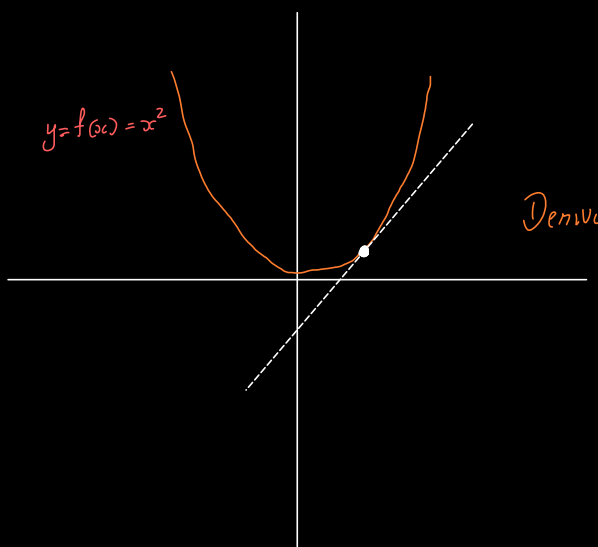
$y \rightarrow (x, y)$

Notation = ∂

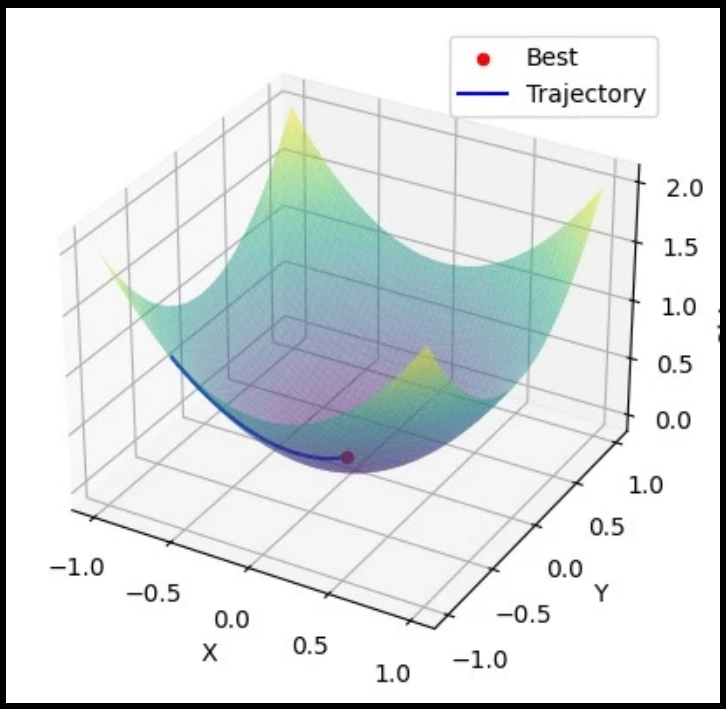
$\frac{\partial y}{\partial x} = 2x \qquad \frac{\partial y}{\partial z} = 2z$

Our loss function depends on 9 parameter of network. behind the scene, we will calculate gradient of each params.

in 2D



Derivative = Calculating slope at specific location.



3D :

$z = f(x, y) = x^2 + y^2$

Calculating slope with-respect-to x, y.

-> Our loss function is a complex 9 dimentional function
-> By calculating derivatives, we're calculating 9 different slopes WRT each dimentions.

Concept of derivative (dy/dx)

-> How y is changing by making changes in x.

$y = x^2 + 2x$

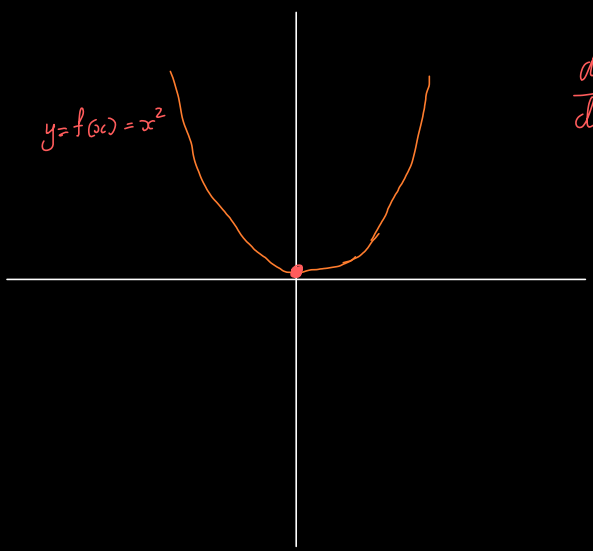
$\frac{dy}{dx} = 2x + 7$

for $x = 5$

$\frac{dy}{dx} = 2(5) + 7$

$= 11$ slope

Concept of minima



2D

$\frac{dy}{dx} = 2x = 0$

At $x = 0$ we are at global minima.

3D

for $z = x^2 + y^2$

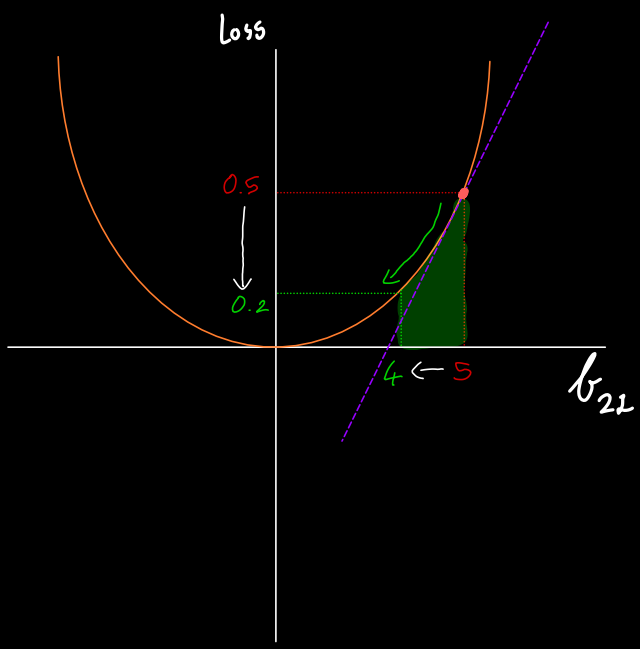
$\frac{\partial z}{\partial x} = 2x = 0$ $\frac{\partial z}{\partial y} = 2y = 0$

At $x = 0$ and $y = 0$, we're at global minima

$b_{21} = b_{21} \boxed{- \frac{\partial L}{\partial b_{21}}}$ Why we're subtracting $\frac{\partial L}{\partial b_{21}}$ .... ?

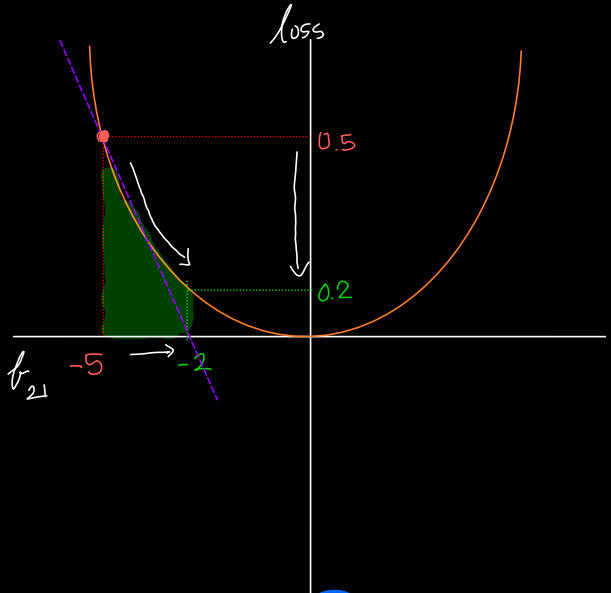
Suppose $b_{21} = 5$ & o/p of $\frac{\partial L}{\partial b_{21}}$ is +ve

Hence o/p of $\frac{\partial L}{\partial b_{21}}$ = +ve means sum loss is increasing as we increase the value of b_{21}. Hence we will subtract it.



→ Similar in opposite, if $\frac{\partial L}{\partial b_{21}}$ is -ve then the value will be added in b_{21}

→ It shows that as b_{21} increases loss is getting negative.

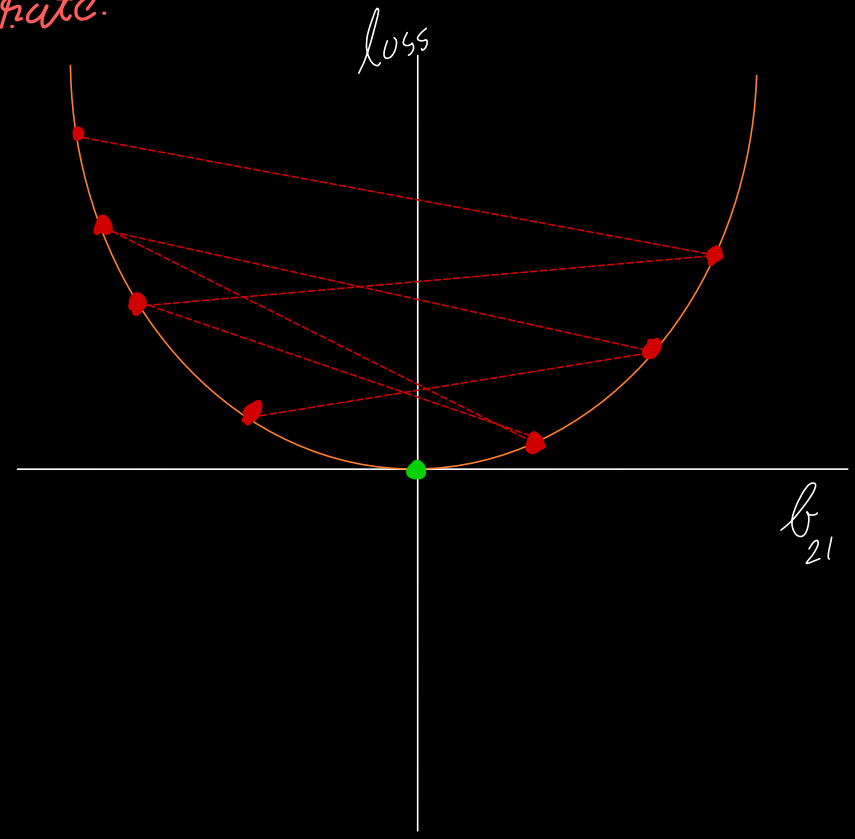


What about learning rate...?

↳ Learning rate makes process smoother & makes process stable.

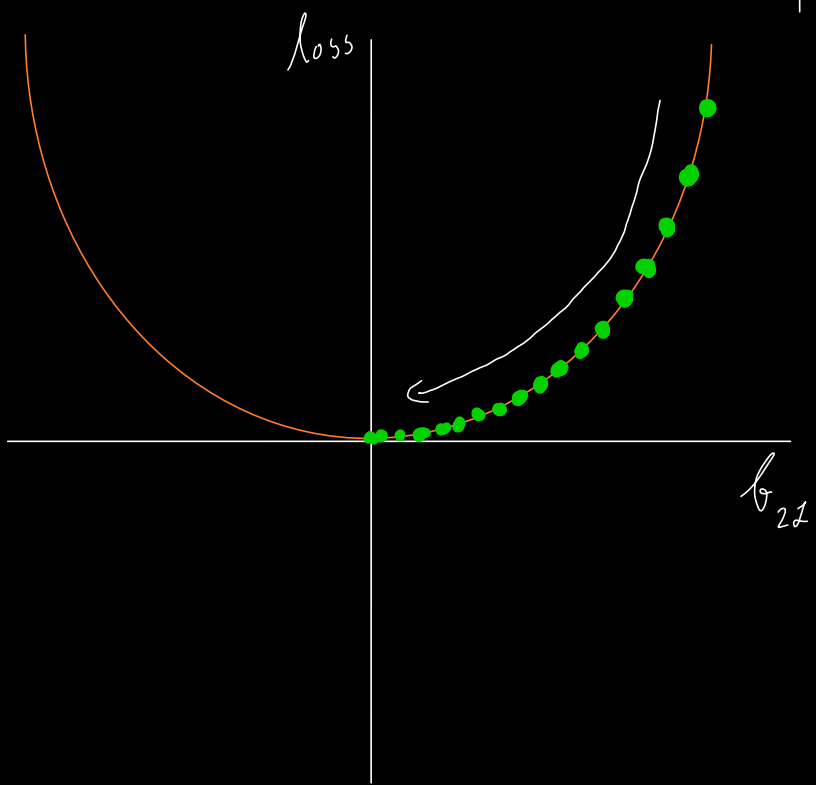
Without learning rate.

$$b_{21} = b_{21} - \frac{\partial L}{\partial b_{21}}$$



with learning rate

$$b_{21} = b_{22} - \alpha \frac{\partial L}{\partial b_{21}}$$



What is convergence...?

$$b_{new} = b_{old} - \alpha \frac{\partial L}{\partial b_{old}} = 0$$

-> $b_{new} = b_{old}$ says that now we've reached at global-minima

-> We should keep updating parameters untill

$$b_{new} \approx b_{old}$$

-> As you reach at minima, stop because further updation won't give any meaningful results.