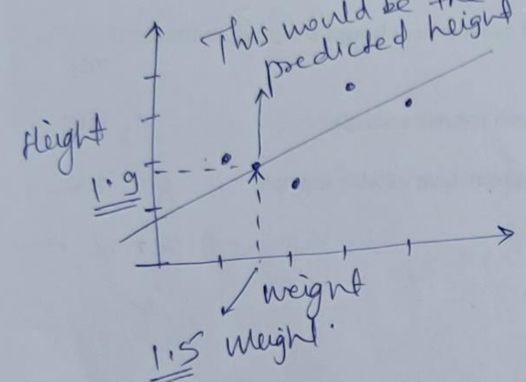
Linear Regression Using Gradient Descent
(An Iterative Approach.)

⇒ When we fit a line with Linear Regression, we optimize the Intercept and slope.



H = Intercept + slop * W.

$H \rightarrow$ height, $W \rightarrow$ weight.

$\hat{y} = a_0 + a_1 x$.

Equation of the line.

⇒ Gradient Descent is used to optimize the Slope & Intercept so that line is the best fit for given data points.

→ Statistics, Machine Learning & Data science

$$\text{Predicted Height} = \underbrace{\text{Intercept} + \text{slop}} * \text{weight}.$$

So. gradient descent can fit a line to data by finding the optimal values..

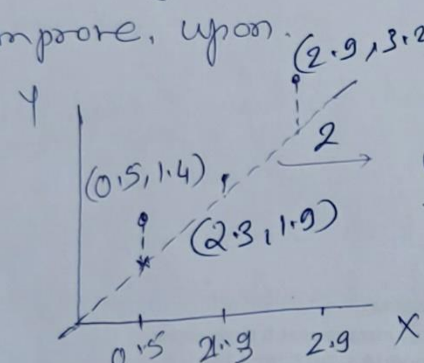
⇒ Start GD to find Intercept; then we used it to solve for Intercept & slope.

→ Currently assume that least square estimate for the slope = 0.64

Predicted Height = Intercept + 0.64 * weight

& we will decide GD to find the optimal value for intercept.

→ Pick a random value for Intercept.
→ This is just initial guess that gives GD something to improve upon.



Initial guess = 0

Equation for this line.

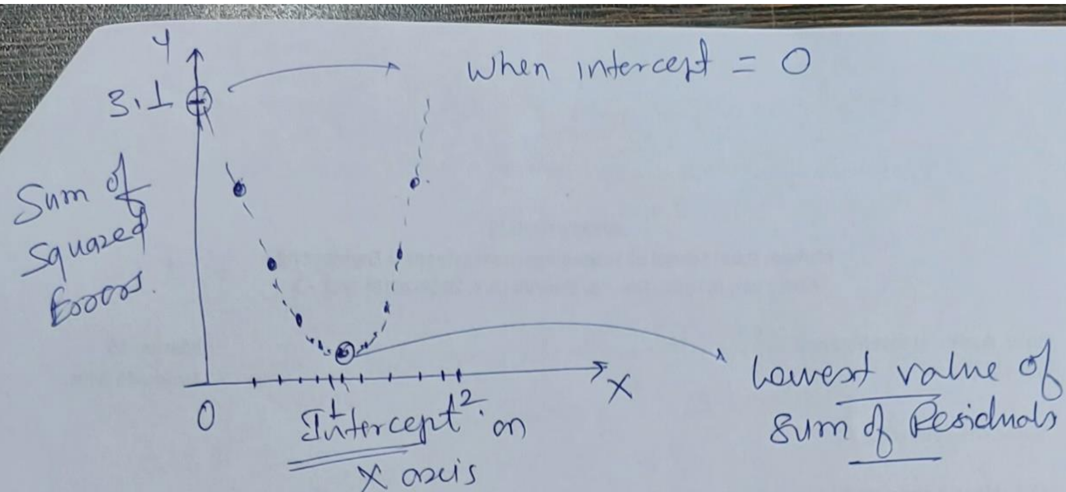
→ we will find out sum of residuals for this line.

In ML - Sum of the squared Residuals is type of loss function - put this value in the eqn.

$$PH = 0 + 0.64 * 0.5 = 0.32$$

Residual = OH − PH.
= 1.4 − 0.32
= 1.1 → Residuals.

Sum of Square Residual $= (1.1)^2 + (0.4)^2 + (1.3)^2$
$= 3.1$



plotting the point on graph for increasing value of residuals.

Is the best ? what happend if best value is in between.

So, it is very slow and tedius method to find out the intercept for the line (by least square error method)

Don't ~~dispaires~~ you think GD will be more efficient method to find intercept (best).

⇒ GD only does a "few calculations" far from the optimal soln. and increases the number of calculations closer to the optimal value.

→ It takes big steps when it is far away and baby steps when it is close.

Lets understand GD to find optimal value of intercept starting from Random value (RV)

$RV = 0 \Rightarrow$

① $RV = 0 \Rightarrow$ when we calculated SSE

$$SSE = \underbrace{(OV - PV)^2}_{\text{First Point}} + \underbrace{(OV - PV)^2}_{\text{Second Point}} + \underbrace{(OV - PV)^2}_{\text{Third Point}}$$

$$= (1.4 - (I + 0.64 * W))^2$$

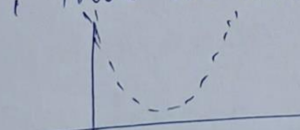
$$\boxed{(y - \bar{y})^2}$$

$$PH = I + \underset{\substack{\uparrow \\ 0.64}}{S} * W$$

$$SSE = (1.4 - (I + 0.64 * 0.5))^2 + (1.9 - (I + 0.64 * 2.3))^2 + (3.2 - (I + 0.64 * 2.9))^2$$

Note: we can put any value for intercept and predict the new height.

Now we can put any values of Intercept and get SSE, Thus we have equation for this curve



$\Rightarrow$ Now we can take derivative of this function and determine the slope at any value for the intercept.

$\Rightarrow$ Lets take derivative of SSE with respect to the Intercept

$$\frac{d}{d(\text{intercept})} SSE = \frac{d}{d(\text{intercept})} (1.4 - (I + 0.64 * 0.5))^2 + \frac{d}{d(\text{intercept})} \text{second Point} + \frac{d}{d(\text{intercept})} \text{Third point}$$

$$\frac{d}{d(\text{intercept})}(1.4 - (I + 0.64 * 0.5))^2 =$$

To take derivative of this, we need to apply (Chain rule) → moving square to the front

$$= 2(1.4 - (I + 0.64 * 0.5)) * (-1)$$

$$= \cancel{2}2(1.4 - I - \underline{0.64 * 0.5}) * (-1)$$

do not content intercept to derivative of Constant is zero

$$= \underbrace{-2(1.4 - (I + 0.64 * 0.5))}_{\text{Derivative of first Point}} +$$

$$-2(1.9 - (I + 0.64 \times 2.3)) +$$

$$-2(3.2 - (I + 0.64 \times 2.9)) \leftarrow \text{eq}^n.1$$

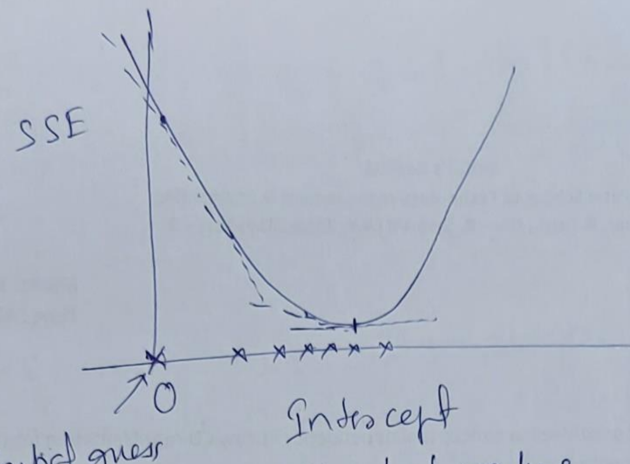
Now we have derivatives, so GD will use to find where SSE is lowest.

Note

If we use LS method to solve for the optimal value for intercept, we would simply find whether the slope of curve $= 0$

— x —

In contrast GD find minimum value by taking steps from initial guess untill it reaches the best value



Until it reaches the best value

This make GD very useful when it is not possible to solve for where the derivative = 0,

It is useful in many situation.

⇒ We started by taking intercept to the random number, in this case, it was 0.

put 0 in the equation of derivative's

$$\frac{d}{d(\text{intercept})} SSE = -2(1.4 - (0 + 0.64 \times 0.5)) + -2(1.9 - (0 + 0.64 \times 2.3)) + -2(3.2 - (0 + 0.64 \times 2.9))$$

$$= -5.7$$

when Intercept = 0 → slope of the curve is -5.7

Intercept =

Note - closer we get to the optimal value for the Intercept, the closer the slop of the curve gets to 0.

This mean when the slope of the curve is close to 0, then we should take baby steps. because we are closer to the optimal value.

When we the slope is far from zero then take bigger steps. because we are far from optimal value of intercept.

=> If we take huge step then increase the SSE. So the size of the step should be related to the slope.

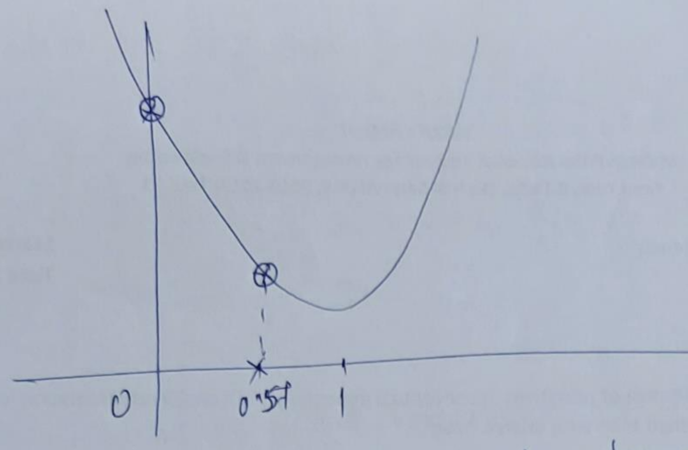
GD determine the step size by multiplying slope by a small number called "The learning Rate"

$$\text{Step size} = -5.7 \times 0.1 = -0.57$$

when intercept was zero step size was ~~-5.7~~ -0.57

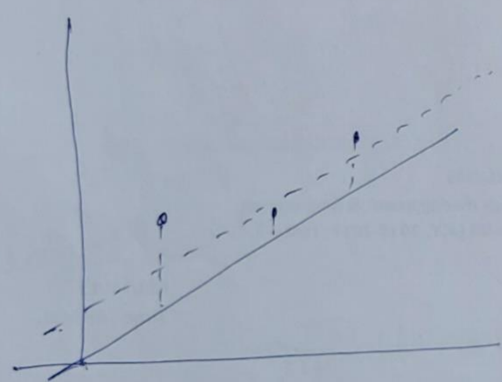
$$\text{New Intercept} = \text{old intercept} - \text{step size}$$
$$= 0 - (-0.57) = 0.57$$

with step size we can calculate new "Intercept"



In one big step we moved to much closer to the optimal value for the intercept.

⇒ Going back to the original data the original line with intercept 0 is moved much closer to the points.



Now put new intercept in derivative eqn.

$$\frac{d}{d\,(\text{intercept})}\,SSE = -2(1.4-(0.57+0.64\times 0.5)) + -2(1.9-(0.57+0.64\times 2.3)) + -2(3.2-(0.57+0.64\times 2.9))$$

$$= \boxed{-2.3}$$ Now this is the slope of the line

Calculate the step size

Step size = slope × Learning Rate

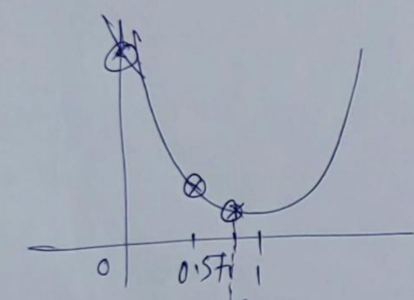
$= -2.3 \times 0.1 = -0.23$

step size $= -0.23$

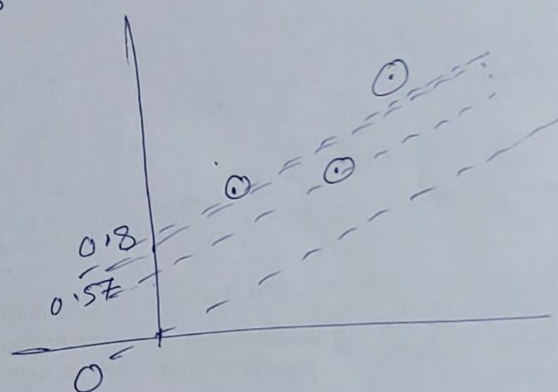
New Intercept = old Intercept − step size

$= 0.57 - (-0.23)$

$= 0.8$



Intercepts



Overall SSE getting smaller with the value of intercept

———x———

⇒ Now calculate derivative again -

$$\frac{d}{d(\text{intercept})} SSE = (-2(1.4-(0.8+0.64\times0.5))+ (-2(1.9-(0.8+0.64\times2.3))+ (-2(3.2-(0.8+0.64\times2.9))$$

$= \boxed{-0.9}$ we get.

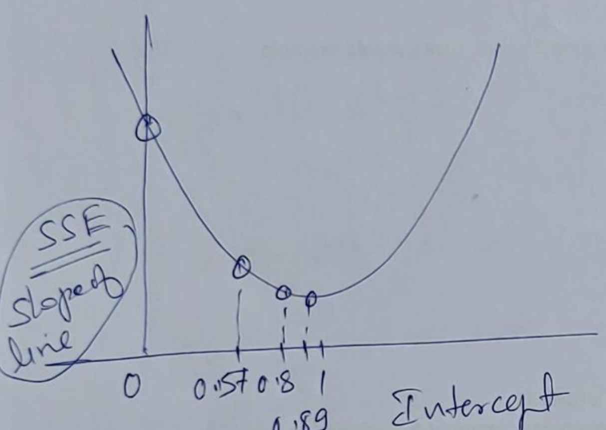
Step Size = slope * LR

= $-0.9 \times 0.1 = -0.09$

Step Size = -0.09

New Intercept = $0.8 - (-0.09)$

= $\boxed{0.89}$



New Intercept = 0.92

New I = 0.94

New I = 0.95

Step gets smaller and smaller when we get closer at the bottom.

After 6 steps GD estimate for the Intercept is 0.95

Note: The least Squares estimate for the intercept is also 0.95

So, GD has done a job, but without comparing the soln to gold standard, How does GD know to stop taking step.?

GD stops when the step size is very close to 0.

⇒ Step size will be very close to 0, when the slope is very close to 0.

In practice minimum step size = 0.001 or smaller.

So the slope = 0.009.

Step size = 0.009 × 0.1

= 0.0009 which is smaller than 0.1 (LoR).

So GD stops.

—x—

⇒ GD also includes a limit on the number of steps it will take before giving up.

In practice, ~~step size~~ = maximum no. of steps = 1,000 or greater

1) → we ~~have~~ ~~is~~ decided to use sum of squared Error SSE as loss function, to evaluate how well a line fits the data.

2) Then we took the derivative of the SSE. In other words we took derivative of loss function.

3) Picked Random value for the intercept = 0

4) Calculate derivative when intercept was. 0
(slope)

5) put that slop. into the step size calculations

$$\text{step size} = \text{slop} \times LR$$

6) Calculated New Intercept

$$NI = OI - \text{step size}$$

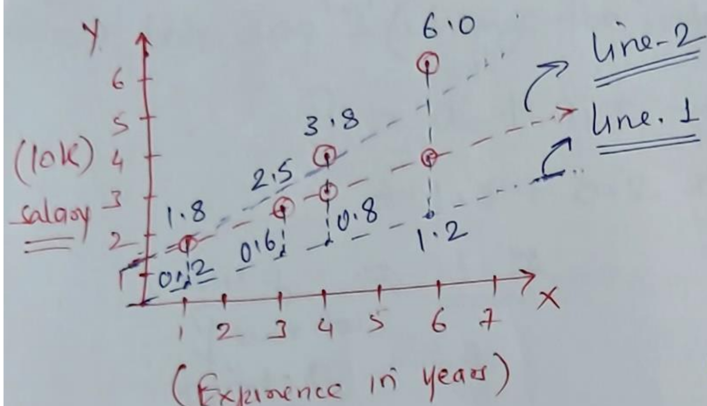
7) we put new intercept into the derivative and repeated everything until step size was close to 0.

LR using Gradient Descent Method (Example)

Data

(1, 18), (3, 25), (4, 38), (6, 60)

(1, 1.8) (3, 2.5), (4, 3.8), (6, 6.0)



Now we have to set LR model for this → straight line

eqn. $y = mx + c$

we can rewrite it as

$y = a_0 + a_1 x$.

a_0 ↓ Intercept, a_1 ↓ slope.

lets assume that $a_0 = 0$ & $a_1 = 0.2$

what will be the equation of line ?

$\bar{y} = 0 + 0.2x$ — line - 1

Need to calculate the "errors" with respect to this line.

Formula for error = Observed value − Estimated value

Error w.r.to point.1 $= (y - \bar{y}) = (1.8 - 0.2)^2$
$= 2.56$.

Sum of squared Error (Loss function) $= (1.8 - 0.2)^2 + (2.5 - 0.6)^2 + (3.8 - 0.8)^2 + (6.0 - 1.2)^2$

$= 2.56 + 3.61 + 9 + 23.04$

$= 38.21$

Value of loss function with respect to line - 1

SSE $= \boxed{38.21} \Rightarrow$ when $a_0 = 0$ & $a_1 = 0.2$

$\Rightarrow$ lets say 'I' (change the intercept) as - 1.5

$\bar{y} = a_0 + a_1 x \longrightarrow$ line - 2

$\bar{y} = 1.5 + 0.2x$. — eqⁿ of the line.

Value of $\bar{y}$ w.r.t. point. 1 $= 1.7$

Sum of Squared Error (Loss Function) $= (y_1 - \bar{y}_1)^2 + (y_2 - \bar{y}_2)^2 + (y_3 - \bar{y}_3)^2 + (y_4 - \bar{y}_4)^2$

$= \cancel{(1.7)}$

$= (1.8 - 1.7)^2 + (2.5 - 2.1)^2 + (3.8 - 2.3)^2 + (6 - 2.9)^2$

$= 0.01 + 0.16 + 2.25 + 9.61$

SSE $= \boxed{12.03} \Rightarrow$ when $a_0 = 1.5$ & $a_1 = 0.2$

Now we will change the slope from 0.2 to 0.6

$\bar{y} = 1.5 + 0.6x$ — Line - 3

$(y_1 - \bar{y}_1)^2 + \ldots$

<u>loss function</u> $= (1.8 - 2.1)^2 + (2.5 - 3.3)^2 + (3.8 - 3.9) + (6 - 5.1)^2$

$= (0.09) + (0.64) + (0.01) + (0.81)$

SSE with respect to all the points $= \boxed{1.55} \Rightarrow$ When $a_0 = 1.5$ & $a_1 = 0.6$

That means we require ~~both~~ (Intercept & Slope)

How do we get this line directly by <u>GD method</u>

What will be the eqn General eq^n of <u>Loss Function</u>

$\boxed{\bar{y} = I + Sx}$

Intercept, slope

<u>SSE</u> $= (y_1 - \bar{y}_1)^2 + \ldots (y_n - \bar{y}_n)^2$

$= (y_1 - (I + Sx))^2 + (y_2 - (I + Sx))^2 + (y_3 - (I + Sx))^2 + (y_4 - (I + Sx))^2$

Now we need to take derivative of loss function with r. to. I & S. seperately.

<u>SSE</u> $= (1.8 - (I + S(1)))^2 + (2.5 - (I + S \times 3))^2 + (3.8 - (I + S \times \cancel{4.8}))^2 + (6 - (I + S \times 6))^2$

Step 1

Derivative w.r.t. to Intercept when Slope is Constant.

$$\frac{d(SSE)}{d(\text{Intercept})} = -2(1.8-(I+S\times 1)) + -2(2.5-(I+S\times 3)) + -2(3.8-(I+S\times 4)) + -2(6-(I+S\times 6)).$$

$$\frac{d}{d(\text{Slope})} = -2\times(1.8-(I+S)) + -2\times 3(2.5-(I+S\times 3)) + -2\times 4(3.8-(I(+S\times 4)) + -2\times 6(6-(I+S\times 6))$$

step 2

Start with random values of I & S.

$I = 0$ & $S = 0.2$

$$\frac{d\ SSE}{d\left(\substack{I=0\\ S=0.2}\right)} = ?$$

we will get Intercept (I) (please calculate)

$$\frac{d\ SSE}{d\left(\substack{S=0.2\\ I=0}\right)} = ?$$

we will get slope (S). (please calculate)

step 3 Need to calculate step size for Intercept & slope

Step size (I) = value obtained in step 2 × LR

Step size (S) = —"—

LR is random

$\boxed{0.01}$

$\boxed{\text{Need to Calculate}}$.

LR → Learning rate that define the step size.

Step 4

Calculate New Intercept

New (I) = old (I) − step size (I)

New (I) = 0 −

New (S) = old (S) − step size (S)

Step 5 put new values of I & S in eqn of derivative of I & S again and repeat the steps till we reach.

$\boxed{SSE = 0}$ or some random number of iterations.

Use of Loss Function to calculate LR;

→ We used SSE as loss function (Example).

Data

X	Y
0.5	1.4
2.3	1.9
2.9	3.2

$$SSE = (1.4-(I+S*0.5))^2 + (1.9-(I+S*2.3))^2 + (3.2-(I+S*2.9))^2 \quad \text{--- (2)}$$

We need to find out value of the Intercept & slope, such that, SSE would be minimum.

⇒ Also, we need to take derivative with respect to slope.

$$\frac{d}{d(\text{Intercept})} SSE = (1.4-(I+S*0.5))^2$$

$$= 2(1.4-(I+S\times0.5))\times(-1)$$

Since we are taking derivative with respect to Intercept; slope would be considered as constant and derivative of Constant is 0.

⇒ So we get (-1) just like before

$$\frac{d}{d(\text{Intercept})} SSE = -2(1.4-(I+S\times0.5)) + (-2(1.9-(I+S\times2.3))) + (-2(3.2-(I+S\times2.9))).$$

$$\frac{d}{d(\text{slope})} SSE = \frac{d}{d(\text{slope})}(1.4-(I+S\times 0.5))^2$$

$$+ \frac{d}{d(\text{slope})}(1.9-(I+S\times 2.3))^2$$

$$+ \quad -\text{''}-$$

$$\cancel{= 2(1.4-(I+S\times 0.5))\times 6(0.5)}$$

$$\frac{d(SSE)}{d(\text{slope})} = -2\times 0.5(1.4-(I+S\times 0.5))$$
$$+ -2\times 2.3(1.9-(I+S\times 2.3))$$
$$+ -2\times 2.9(1.9-(I+S\times 2.9))$$

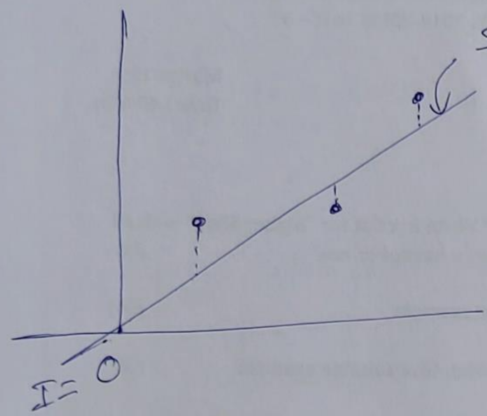
<u>Note</u> : when you have two or more derivatives of the same function, they are called as <u>Gradient</u>

We use this Gradient to Descend to lowest point in the <u>loss function</u>, (SSE).

Thats why it is called as <u>Gradient Descent</u>

Now we start with Random value of I & S

I = 0 & S = 1



Now put the values of I = 0 & S = 1 in Both the equations.

we get

$$\frac{d}{d(I)} SSE = -1.6 \quad \text{(Intercept)}$$

we get

$$\frac{d}{d(S)} SSE = -0.8 \quad \text{(slope)}$$

$\Rightarrow$ Step Size(I) = $-1.6 \times LR$

$\Rightarrow$ step size (S) = $-0.8 \times LR$ $\qquad$ $LR = 0.01$

$\Rightarrow$ Large LR will not work in this case.

$$SS(I) = -1.6 \times 0.01 = -0.016$$

$$SS(S) = -0.8 \times 0.01 = -0.008$$

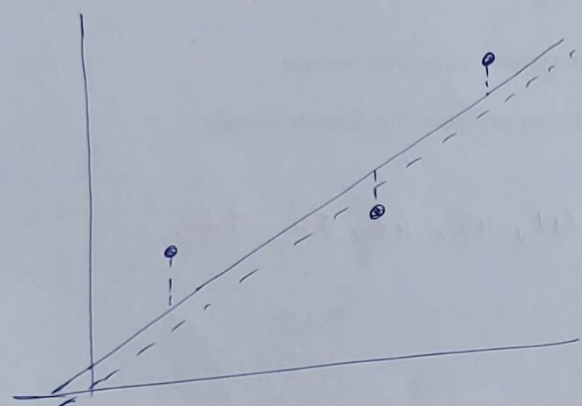
GD is very sensitive to the LR.

In practical Situation LR can be Large Initially and getting smaller at each step.

Now, we calculate .

$$\text{New Intercept} = \text{old } I - \text{step size}$$
$$= 0 - (-0.016) = 0.016$$

$$\text{New slope} = S - \text{step size}$$
$$= 1 - (-0.008)$$
$$= 1.008$$



Now we have to repeat the same process untill all the steps sizes are very small or we reach the maximum number of steps

we have the find out best fitting line

$$\left.\begin{aligned} I &= \underline{\underline{0.95}} \\ S &= \underline{\underline{0.64}} \end{aligned}\right\}$$ Same value we get from least squares.

Now, we have understood how GD will optimize two parameters the slope and Intercept