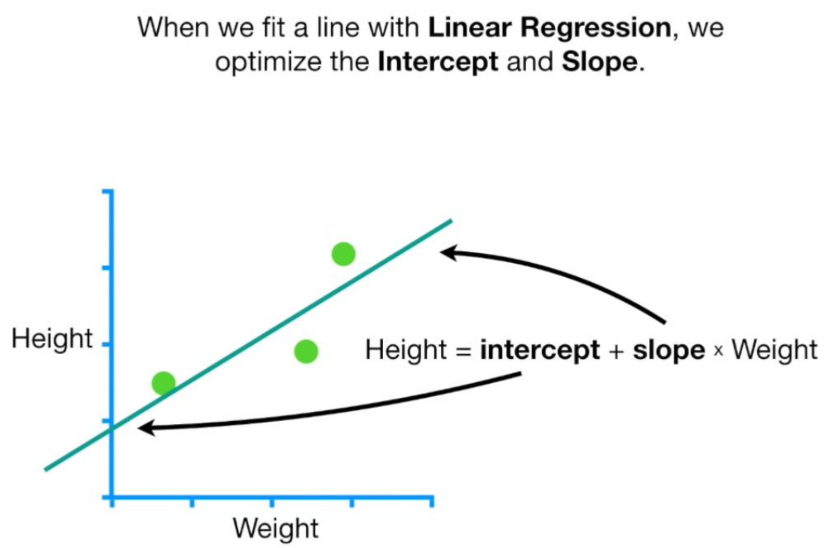
ASAD ASHRAF KAREL

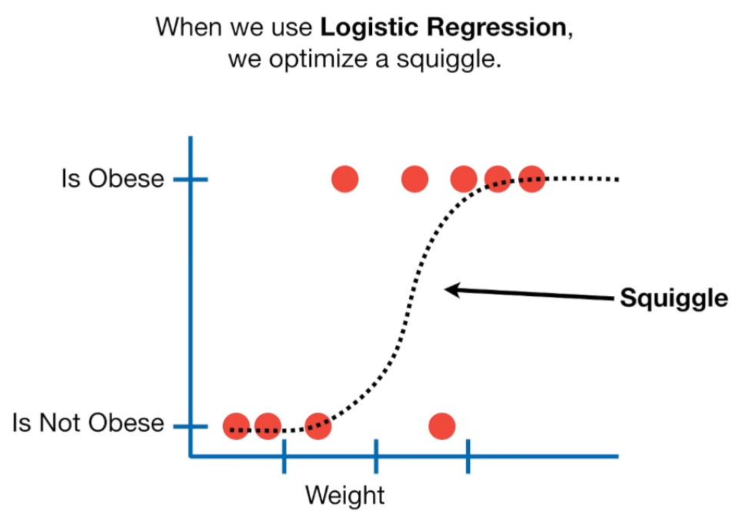
# Gradient Descent

Purpose to find out the best intercept value for best fit line of Regression

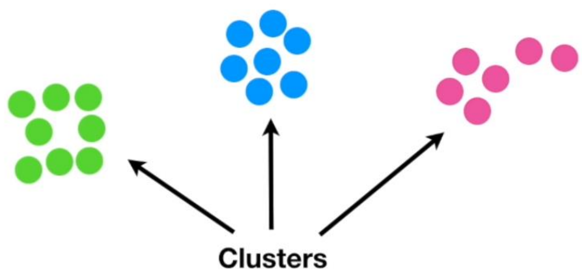
Entire game is around slope and intercept of the line(s)



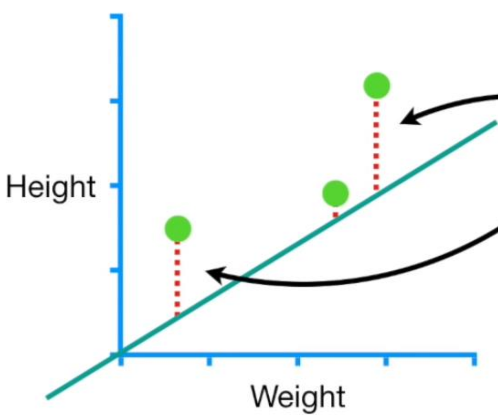

---



And when we use **t-SNE**, we optimize clusters.

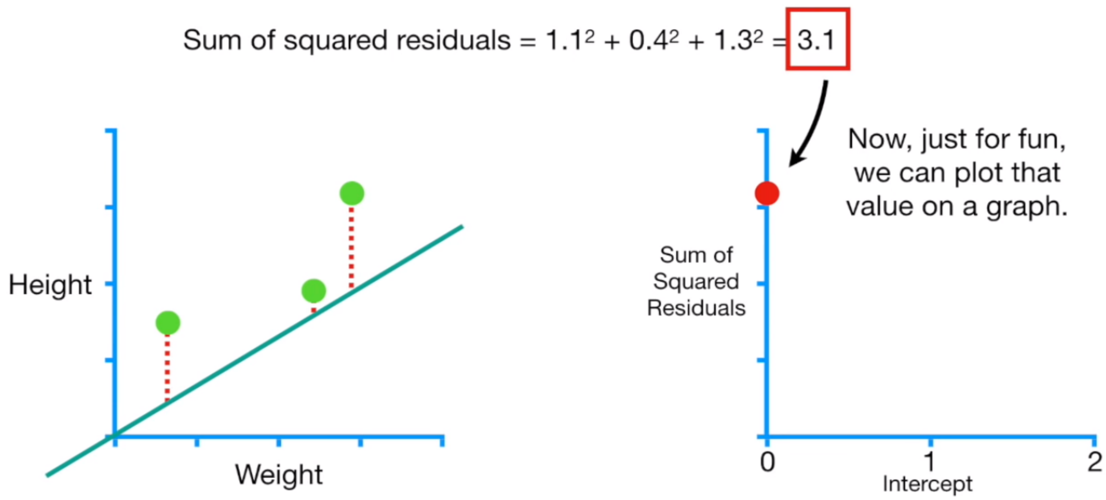


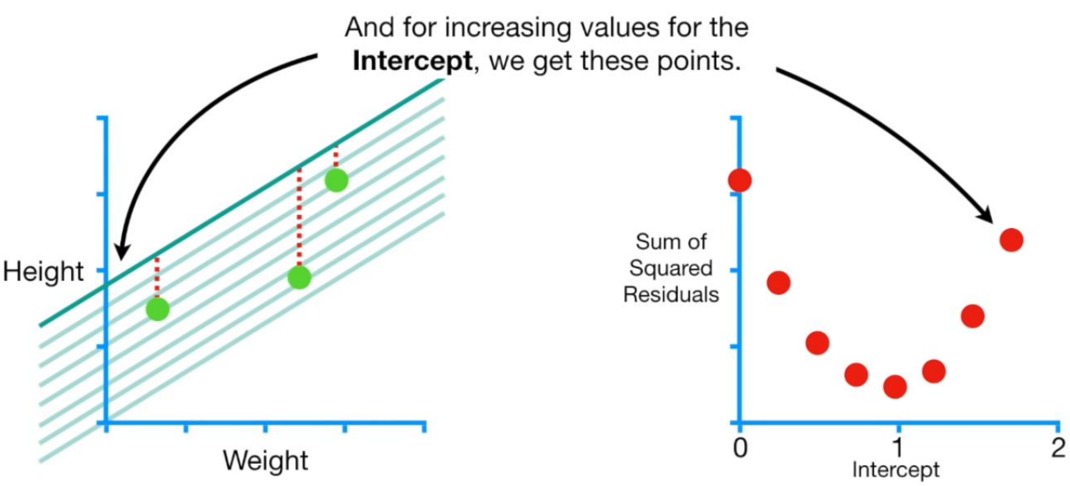

---




In this example, we will evaluate how well this line fits the data with the **Sum of the Squared Residuals**.

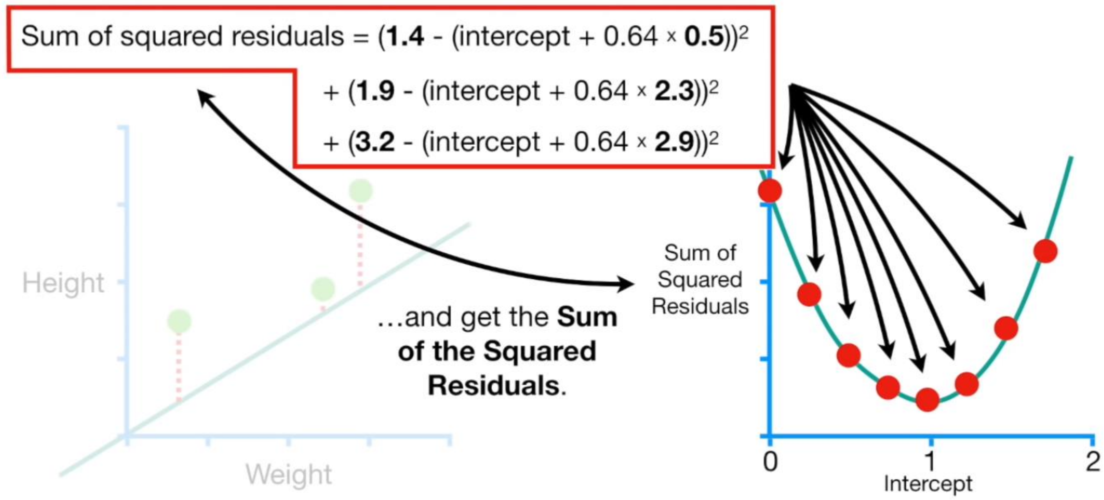
**NOTE:** In Machine Learning lingo, The Sum of the Squared Residuals is a type of **Loss Function**.




Here we have several fitting lines, but we choose the best fit line (least residual value).

Here at intercept 1, we are getting least residual value, which mean line passing through the intercept with the respective slope is best fit line.




$$\frac{d}{d\ intercept}\boxed{(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5})}^2 = 2(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5}))\boxed{\times -1}$$

$$\frac{d}{d\ intercept}\ \mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5})$$

$$\frac{d}{d\ intercept}\ \cancel{\mathbf{1.4}} + (-1)\text{intercept} - \cancel{0.64 \times \mathbf{0.5}} = -1$$

These parts don't contain a term
for the **Intercept**, so they go away.

We use **Chain Rule** for taking derivatives to check the slope of the line.

Negative sign is nothing but minus sign of the part of predicted value (line), and 1 is derivative of respective part is 1.

$$\frac{d}{d\ intercept}(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5}))^2 = 2(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5})) \times -1$$

$$\boxed{= -2(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5}))}$$

…and this…

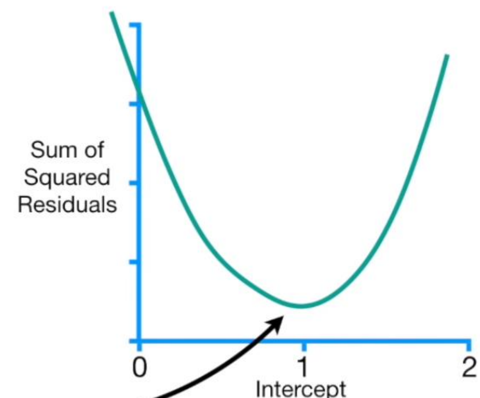
…is the derivative of the first part…

$$\frac{d}{d\ intercept}\text{Sum of squared residuals} = \boxed{\frac{d}{d\ intercept}(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5}))^2}$$

$$+ \frac{d}{d\ intercept}(\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))^2$$

$$+ \frac{d}{d\ intercept}(\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))^2$$

$$\frac{d}{d\ intercept}\text{Sum of squared residuals} = -2(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5}) + -2(\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3})) + -2(\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))$$

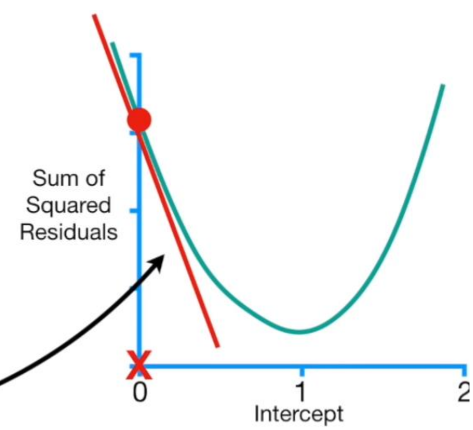



Now that we have the derivative, **Gradient Descent** will use it to find where the Sum of Squared Residuals is lowest.

---

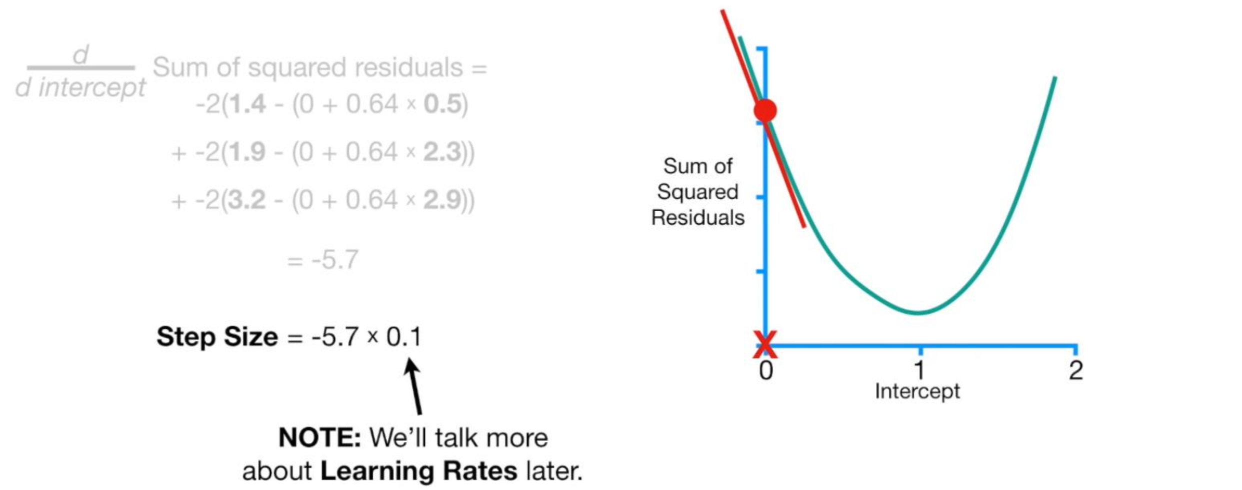
$$\frac{d}{d\ intercept}\text{Sum of squared residuals} = -2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5}) + -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3})) + -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

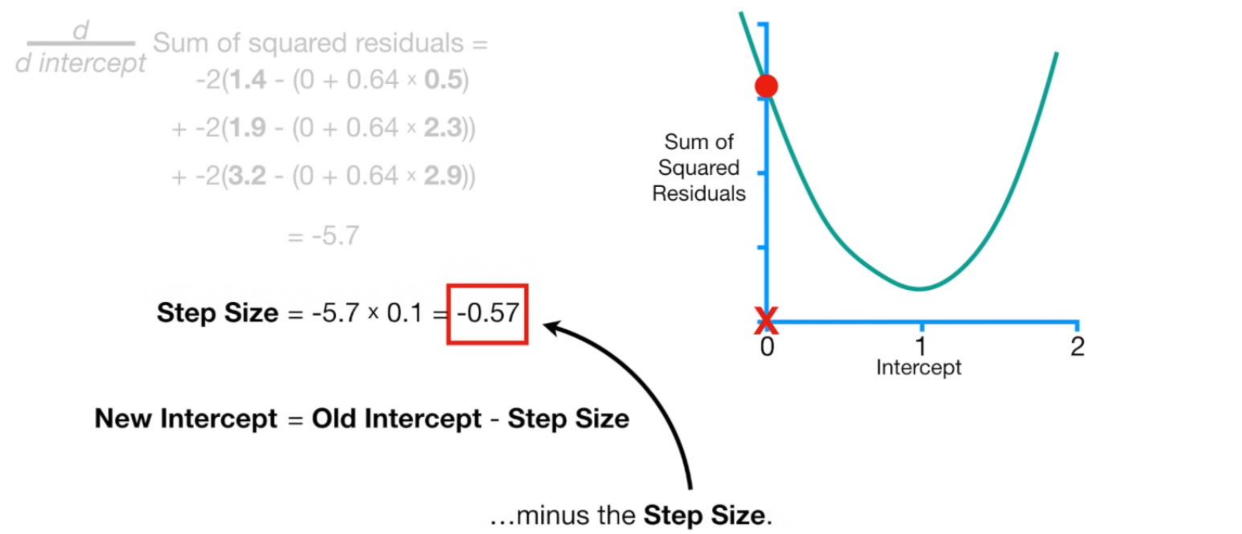
$$= -5.7$$




So when the **Intercept** = **0**, the slope of the curve = **-5.7**.

When slope is far from zero, we should take **big steps** (not much big, but moderate) and when we are close zero then we should take **baby steps**, so that we could get a very close value of the intercept.



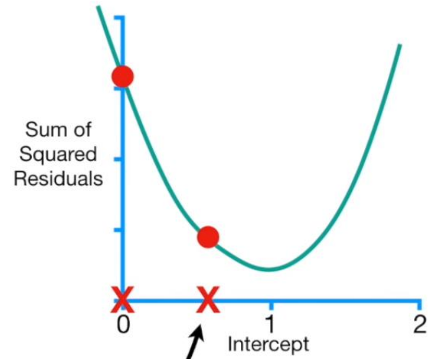

**Step-size** we kept by 0.1. We change it respectively along the calculations.

Negative sign indicates the slope going down, which means line is going towards down.

Our aim is to go close to zero.

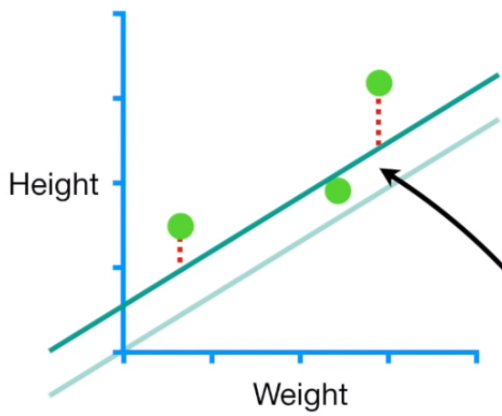
$$\frac{d}{d\ intercept}\text{ Sum of squared residuals} = -2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5}) + -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3})) + -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9})) = -5.7$$

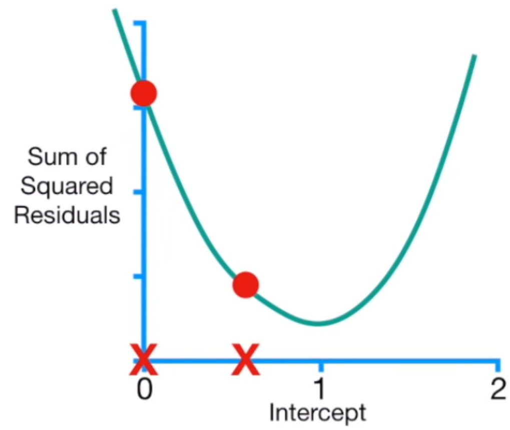
**Step Size** = -5.7 × 0.1 = -0.57




**New Intercept** = 0 - (-0.57) = $\boxed{\mathbf{0.57}}$

…and the the **New Intercept** = **0.57**.



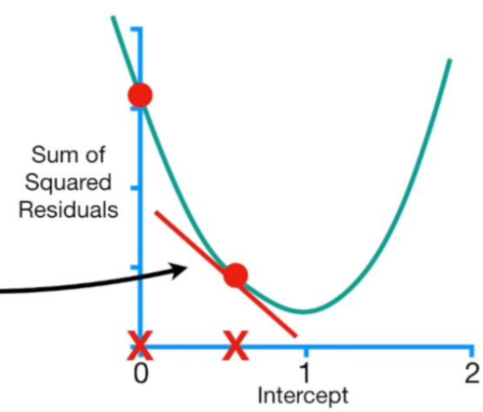




…we can see how much the residuals shrink when the **Intercept** = **0.57**.

---

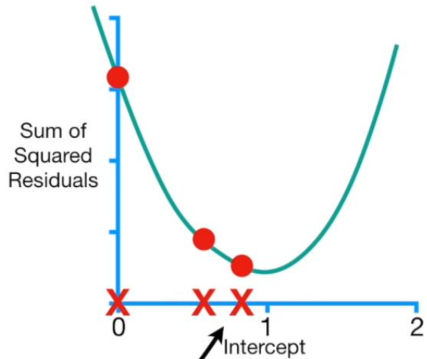
$$\frac{d}{d\ intercept}\text{ Sum of squared residuals} = -2(\mathbf{1.4} - (0.57 + 0.64 \times \mathbf{0.5}) + -2(\mathbf{1.9} - (0.57 + 0.64 \times \mathbf{2.3})) + -2(\mathbf{3.2} - (0.57 + 0.64 \times \mathbf{2.9})) = \boxed{-2.3}$$

…and that tells us the slope of the curve = **-2.3**.



$$\frac{d}{d\ intercept}\text{ Sum of squared residuals} = -2(\mathbf{1.4} - (0.57 + 0.64 \times \mathbf{0.5}) + -2(\mathbf{1.9} - (0.57 + 0.64 \times \mathbf{2.3})) + -2(\mathbf{3.2} - (0.57 + 0.64 \times \mathbf{2.9})) = -2.3$$

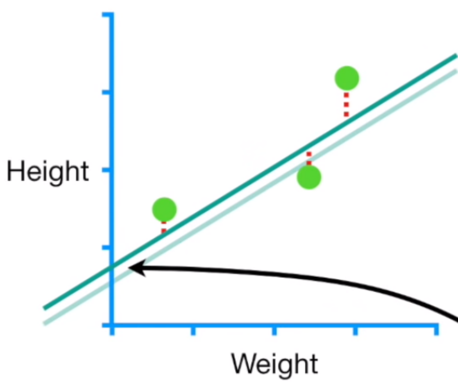
**Step Size** = -2.3 × 0.1 = **-0.23**

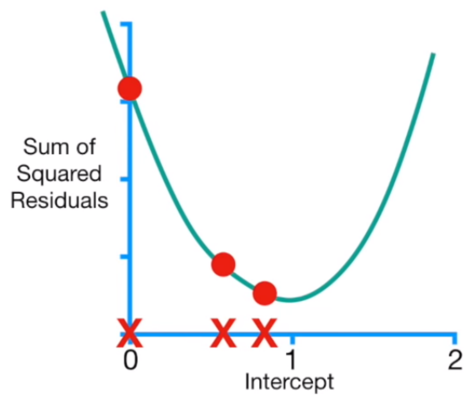



**New Intercept** = 0.57 - (-0.23) = $\boxed{\mathbf{0.8}}$

…and the **New Intercept** = **0.8**

---



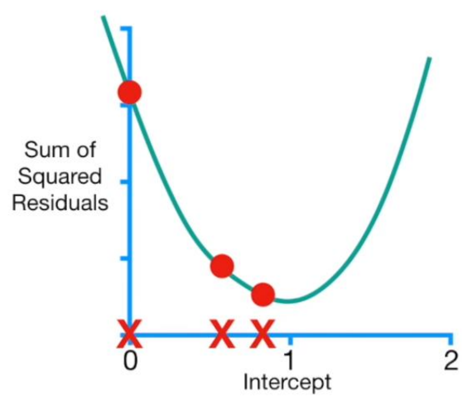



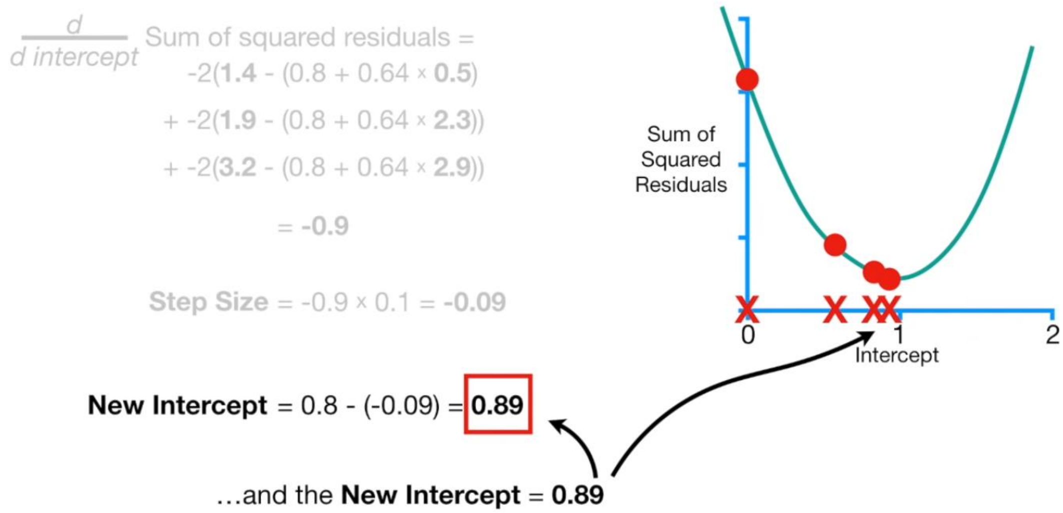

…to when the
**Intercept** = **0.8**

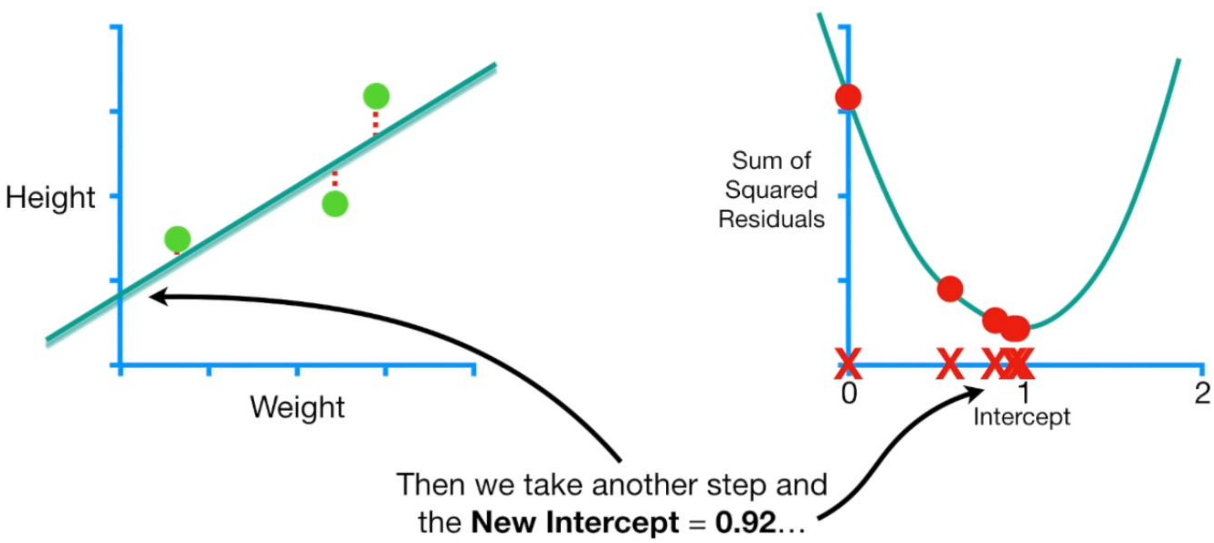
---

$$\frac{d}{d\ intercept}\text{ Sum of squared residuals} = -2(\mathbf{1.4} - (0.8 + 0.64 \times \mathbf{0.5}) + -2(\mathbf{1.9} - (0.8 + 0.64 \times \mathbf{2.3})) + -2(\mathbf{3.2} - (0.8 + 0.64 \times \mathbf{2.9})) = \boxed{\mathbf{-0.9}}$$

…and we get **-0.9**.





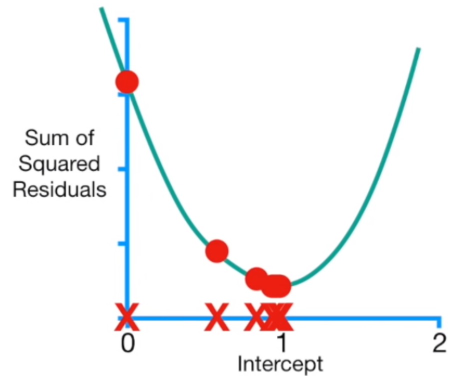

**Big step** and **Baby step** is automatically handle by Gradient Descent, as shown in above figure.

As we approach towards zero residuals, we get a proper slope. That's all the magic done by Gradient Descent.

After 6 steps, the **Gradient Descent** estimate for the **Intercept** is **0.95**.

**NOTE:** The **Least Squares** estimate for the intercept is also **0.95**.

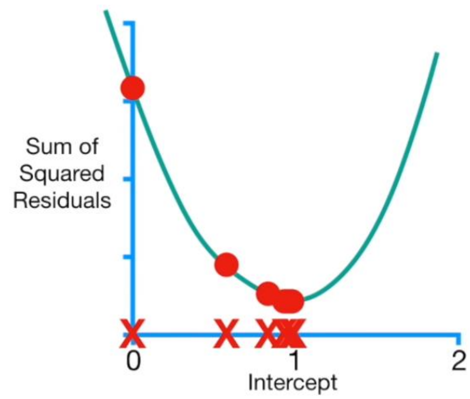
So we know that **Gradient Descent** has done its job, but without comparing its solution to a gold standard, how does **Gradient Descent** know to stop taking steps?




---

**Gradient Descent** stops when the **Step Size** is **Very Close To 0**.

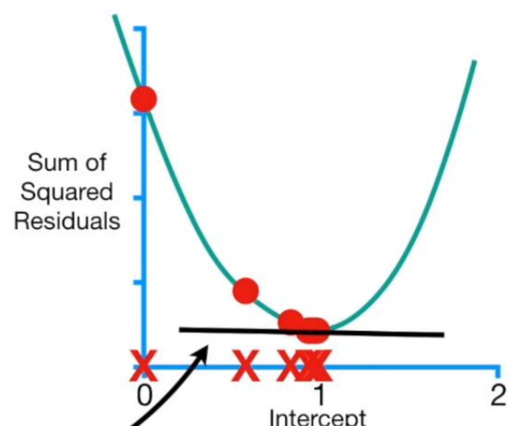
**Step Size** = **Slope** × **Learning Rate**




---

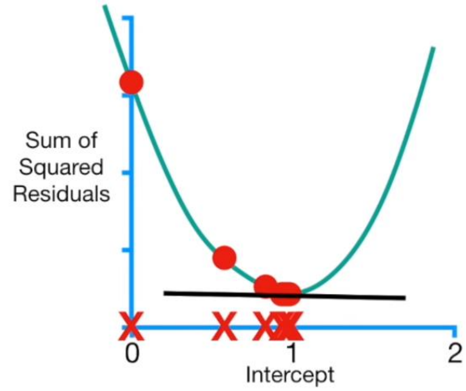
The **Step Size** will be **Very Close to 0** when the **Slope** is very close to **0**.

**Step Size** = $\boxed{\textbf{Slope}}$ × **Learning Rate**



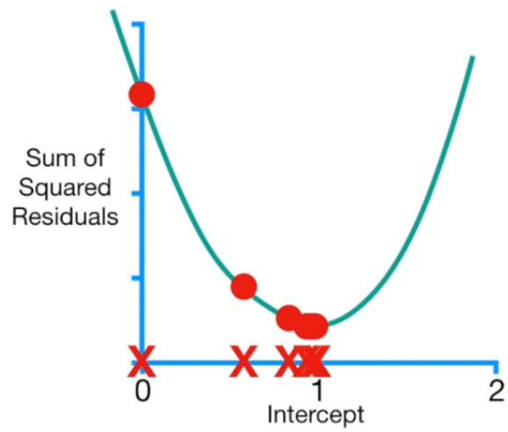
Then we would plug in **0.009** for the **Slope** and **0.1** for the **Learning Rate**..

**Step Size** = **0.009** × **0.1**




---

So, even if the **Step Size** is large, if there have been more than the **Maximum Number of Steps**, **Gradient Descent** will stop.




---

$$\frac{d}{d\ intercept} \text{Sum of squared residuals} = -2(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}) + -2(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3})) + -2(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))$$

**NOTE:** When you have two or more derivatives of the same function, they are called a **Gradient**.

$$\frac{d}{d\ slope} \text{Sum of squared residuals} = -2 \times \mathbf{0.5}(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5})) + -2 \times \mathbf{2.9}(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2 + -2 \times \mathbf{2.3}(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2$$

$$\frac{d}{d\ intercept}\ \text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))$$
$$+ -2(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))$$
$$+ -2(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))$$

We will use this **Gradient** to **descend** to lowest point in the **Loss Function**, which, in this case, is the Sum of the Squared Residuals…

…thus, this is why this algorithm is called **Gradient Descent**!

$$\frac{d}{d\ slope}\ \text{Sum of squared residuals} =$$

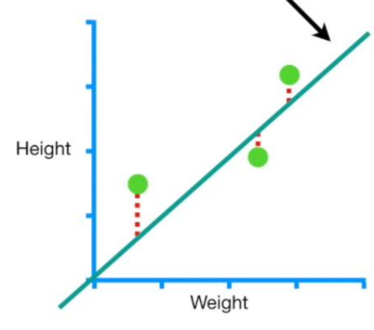
$$-2 \times \mathbf{0.5}(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))$$
$$+ -2 \times \mathbf{2.9}(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$$
$$+ -2 \times \mathbf{2.3}(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2$$

---

$$\frac{d}{d\ intercept}\ \text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))$$
$$+ -2(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))$$
$$+ -2(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))$$

Thus, this line, with **Intercept** = **0** and **Slope** = **1**, is where we will start.




$$\frac{d}{d\ slope}\ \text{Sum of squared residuals} =$$

$$-2 \times \mathbf{0.5}(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))$$
$$+ -2 \times \mathbf{2.9}(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$$
$$+ -2 \times \mathbf{2.3}(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2$$

$$\frac{d}{d\ intercept}\ \text{Sum of squared residuals} = -2(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5}) + -2(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3})) + -2(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9})) = \mathbf{-1.6}$$

**Step Size**$_{\text{Intercept}}$ = -1.6 × 0.01 = **-0.016**

**New Intercept** = 0 - (-0.016) = 0.016

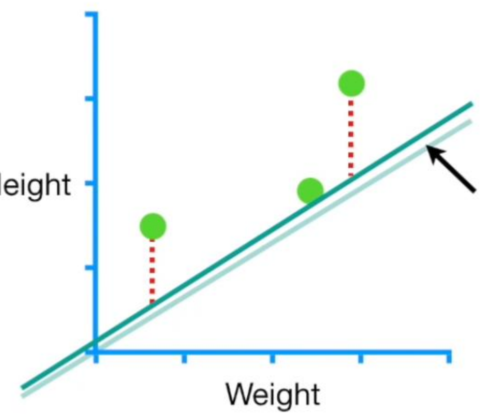
…and we end up with a **New Intercept** and a **New Slope**.

$$\frac{d}{d\ slope}\ \text{Sum of squared residuals} = -2 \times \mathbf{0.5}(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5})) + -2 \times \mathbf{2.9}(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9}))^2 + -2 \times \mathbf{2.3}(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3}))^2 = \mathbf{-0.8}$$

**Step Size**$_{\text{Slope}}$ = -0.8 × 0.01 = **-0.008**

**New Slope** = 1 - (-0.008) = 1.008

---

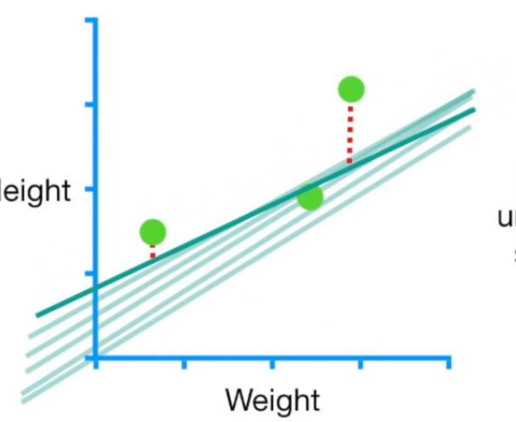



**New Intercept** = 0 - (-0.016) = 0.016

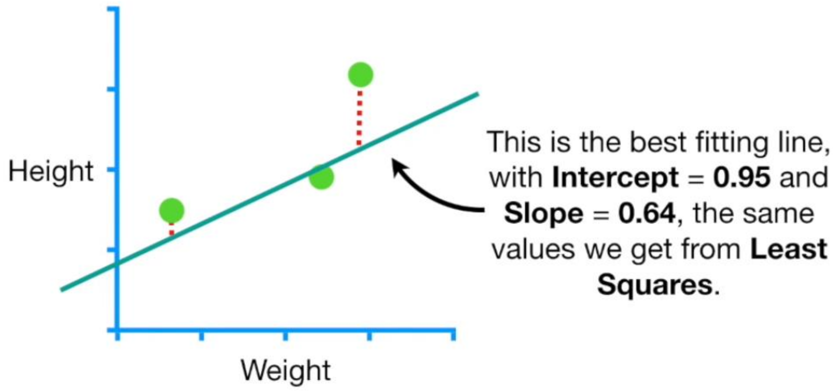
…and this is the new line (with **Slope = 1.008** and **Intercept = 0.016**) after the first step.

**New Slope** = 1 - (-0.008) = 1.008

---




Now we just repeat what we did until all of the **Steps Sizes** are very small or we reach the **Maximum Number of Steps**.




Here the points are very less, but what if the points are in millions of values. Hence we approach to **Stochastic Gradient Descent**. Because, for less points the mathematics is simpler but for huge values it’s not. It is great if we have tones of data and lots of parameters. The mathematics is all same like regular gradient descent, but stochastic descent takes either a single value from the data or from the cluster for each step.