1 What is Gradient Descent

When we break the word Gradient Descent and understand what it actually means, gradient means inclined part of a pathway or a slope, descent means to move towards the bottom of the slope. Imagine yourself to be a mountaineer, and you are trying to get to the bottom of the mountain, you are descending the gradient of a mountain.

2 Understanding Gradient Descent for $f(x) = x^2$

Suppose function $y = f(x)$, where x, y are real numbers.

- This function has minimum at $x = 0$ which we want to determine using gradient descent.
- Derivative of function denoted: $f^{'}(x)$ or as $\frac{dy}{dx}$
- Derivative $f^{'}(x)$ gives the slope of $f(x)$ at point x.
- It specifies how to scale a small change in input to obtain a corresponding change in the output:

  $f(x + \eta) \approx f(x) + \eta f^{'}(x)$ where η is a small change made.
- It tells how you make a small change in input to make a small improvement in y.
- We know that $f(x - \eta sign(f^{'}(x))))$ is less than $f(x)$ for small η. Thus we can reduce $f(x)$ by moving x in small steps with opposite sign of derivative

  where $sign(x) = \{ \begin{array}{cc} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{array}$

This technique is called gradient descent.

Consider a simple math equation $f(x) = x^2$,

x	$f(x) = x^2$
1	1
2	4
3	9
4	16
5	25
6	36
0	0
-1	1
-2	4
-3	6
-4	8
-5	25
-6	36

When above values are plotted we get the following graph

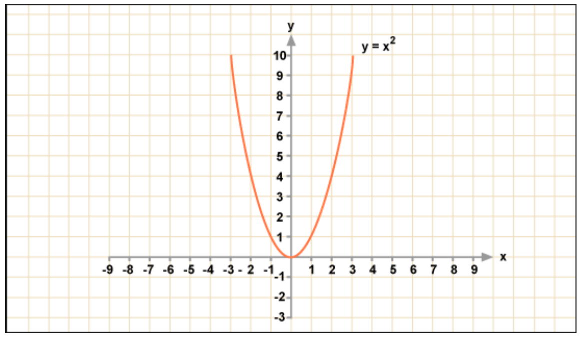

Figure 1

Here, using gradient descent we want to find the minimum of the function $f(x)$.
This function has minimum at $x = 0$ which we want to determine using gradient descent.
We have $f^{'}(x) = 2x$
For gradient descent, we update by $-f^{'}(x)$
If $x(t) > 0$ then $x(t+1) < x(t)$
If $x(t) < 0$ then $f^{'}(x) = 2x$ is negative, thus $x(t+1) > x(t)$

Updation rule: $x^{'} \leftarrow x - \eta \frac{dy}{dx}$

Procedure:
Gradient-Descent(
$x^{'}$ //Initial starting point
f //function to be minimized
δ //Convergence threshold)
1 t $\leftarrow$ 1
2 do
3 $x(t+1) \leftarrow x(t) - \eta \frac{dy}{dx}$
4 $t \leftarrow t + 1$
5 while $||x(t+1) - x(t)|| > \delta$
6 return $(x(t))$

References:

Gradient descent over multi-dimensional parameters:
`https://gluon.mxnet.io/chapter06_optimization/gd-sgd-scratch.html`
An Overview of Gradient Descent Algorithm:
`http://ruder.io/optimizing-gradient-descent/index.html#batchgradientdescent`