Optimization Algorithms

Gradient descent :

It is Optimization algorithm used in machine learning to tweak the parameters iteratively to minimize the given cost function(differentiable function) to its local minimum.

The idea is to take repeated steps in the opposite direction of the gradient (or approximate gradient) of the function at the current point, because this is the direction of steepest descent.

Steps :

1. Start with some random values parameters.
2. Keep changing the parameters until we end up at a minimum value of cost function.

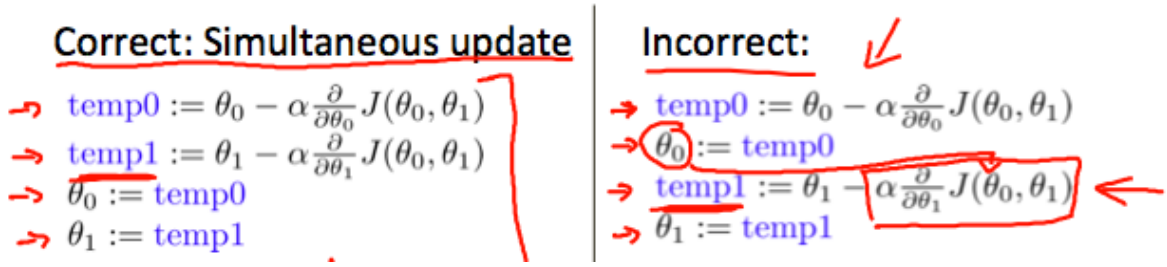
Algorithm :

Repeat until converges {

$$\Theta\text{j} = \Theta\text{j} - \alpha\frac{\partial \text{J}}{\partial\theta} J(\theta_0, \theta_1, \dots\dots, \theta_n)$$

}

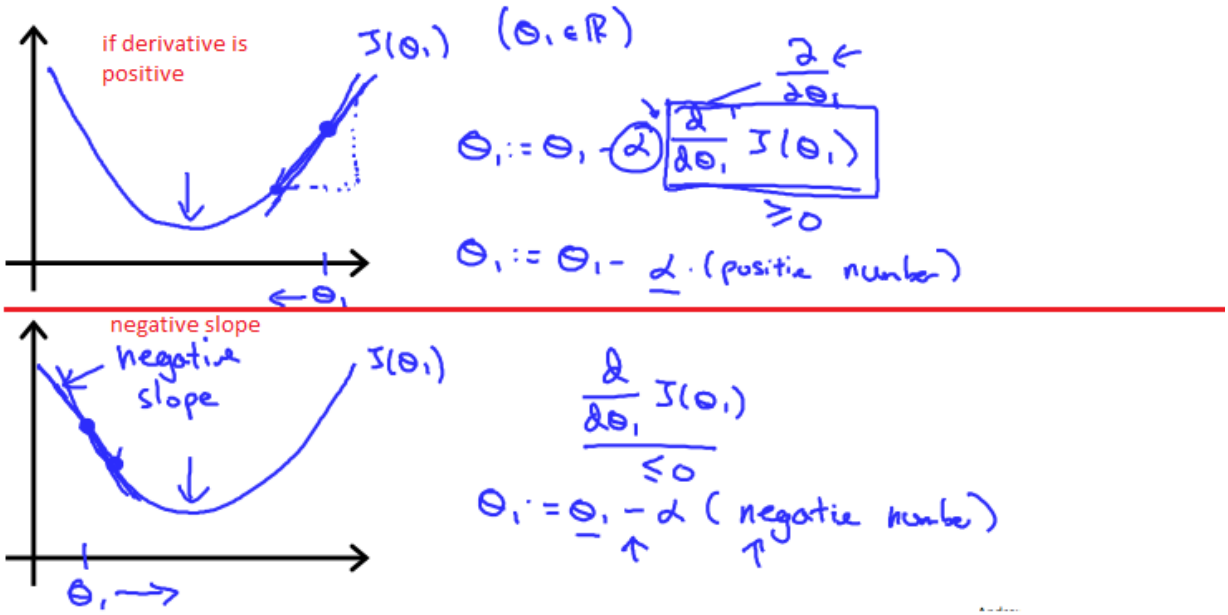
- Here " ***α*** " is the learning rate, it decides how big the step taken by algorithm to reach local minimum.
- $\frac{\partial \text{J}}{\partial\theta} J(\theta_0, \theta_1, \dots\dots, \theta_n)$ is the derivative term.
- We must simultaneously update the parameter values. Suppose we have 2 parameters θ_0, θ_1, then we must update the parameters as follows :



- In the incorrect one the updated **θ_0** will be used in updating the **θ_1** which is an error, so we follow simultaneous update in gradient descent algorithm.

https://coursera.org/share/4e98bc5d188ae89053e81c8101dca65a

To make it simpler consider only θ_1, now plot a graph between θ_1 and cost function J.



- If derivative term is positive then the next θ_1 value will be move towards left side to reach local minimum.
- If derivative term is negative then the next θ_1 value will be move towards right side to reach local minimum.
- If derivative term is zero then it reached the local minimum.

On the other side we also need to adjust the value of our learning Parameter to ensure that the gradient descent algorithm converges at a resonable time.

Failure to converge or too much time to obtain minimum value imply that our step size is wrong.

- If our learning parameter value is too small, then it takes more time to reach minimum.
- On the other hand if its value is large it will fail to converge.

Why there is no need to update the learning parameter along with the parameter ?

Because as we approch to the local minimum the diffentiation value will be decreased and the total product value will also decreases significantly, that is why we can approach local minimum even with fixed value of α.

Note: For the specific choice of cost function $J(\theta_0, \theta_1)$ used in linear regression, there are no local optima (other than the global optimum).

Normal Equation :

Instead of running iterative algorithm, we can solve theta values analytically, i.e. we get weight parameters all in one go.

The equation is $\theta = (\mathrm{X}^T X)^{-1}(\mathrm{X}^T y)$

In the "Normal Equation" method, we will minimize J by explicitly taking its derivatives with respect to the θj ’s and setting them to zero.

Derivation:

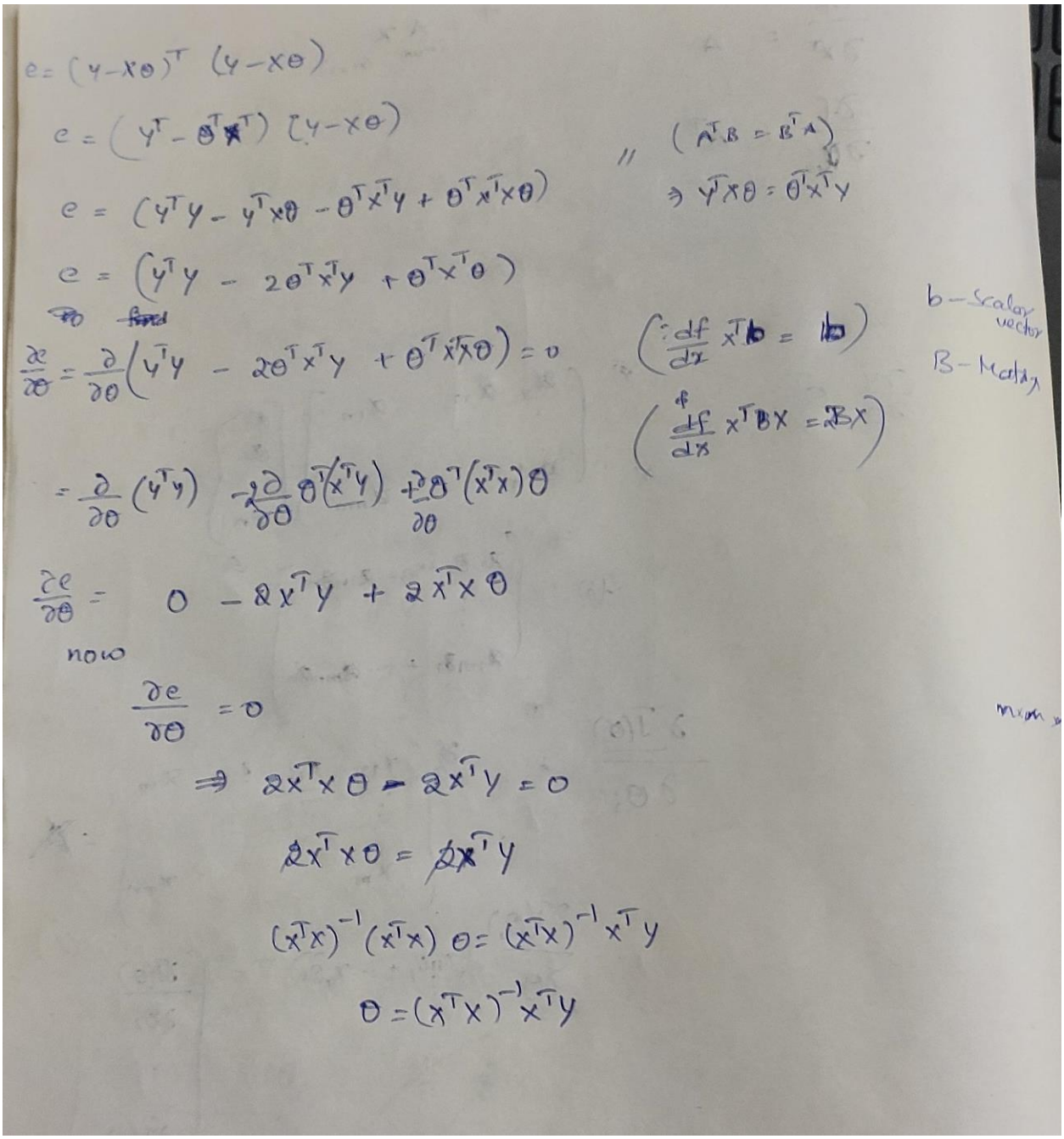
Suppose we have “m” training examples (x_i, y_i) where i= 1..m, and we have ‘n’ features for every training example i.e. $x^j = [x_i^1, x^2_{i,\ldots\ldots\ldots,} x_i^n]$, we add $x_i^0 = 1$ (n X1) vector, for easy matrix manipulations, like it acts as coefficients for our bias theta0), then our feature vector is of size(n+1,1).

Now combining all these feature vectors to form a matrix representation, i.e.

X = $\begin{bmatrix} x_1^0 & \cdots & x_1^n \\ \vdots & \ddots & \vdots \\ x_m^0 & \cdots & x_m^n \end{bmatrix}$ a (n+1 , m) matrix and our corresponding weight matrix $\theta = \begin{matrix} \theta^0 \\ \vdots \\ \theta^n \end{matrix}$ an n+1 dimensional vector. Each row of X represents one data sample.

Now we need to find for what value of theta does X θ = y,

So we inorder to achieve that



References :

1. http://mlwiki.org/index.php/Normal_Equation
2. https://www.youtube.com/watch?v=uoejt0FCWWA
3. http://www.rmi.ge/~kade/LecturesT.Kadeishvili/MathEconomics/Term3/Week3QuadraticLEC.pdf
4. http://www.gatsby.ucl.ac.uk/teaching/courses/sntn/sntn-2017/resources/Matrix_derivatives_cribsheet.pdf