Optimizers:

Optimzers or optimization algorithms are helpful in reducing the losses (cost function),improve the speed of learning process and give accurate obtained predication as output.

Types of Optimizers

1. Gradient Descent
2. Adam
3. Stochastic Gradient Descent
4. Adagrad
5. AdaDelta

In the above , only first three are considered and widely used

In all those three, Adam is frequently used than other two in DL.

**Gradient Descent**:

It is basic and most popular optimizer Algorithm , which is mostly used in neural network

It is used to update weight in model and tune parameters used in model to minimize the Loss(Cost Function)

|  |  |
| --- | --- |
| Advantages | Disadvantages |

It is easy to use it stops at minima

It is easy to understand it may take much time if it is working on

large dataset.

It may consume large memory of CPU

Formula for Gradient Descent

**θ=θ−η⋅∇J(θ)**

**where**

where ‘**η’**is the learning rate,’**∇J(θ)’**is the**Gradient** of **Loss function-J(**θ**)**with respect to parameters-‘θ’.

**Stochastic Gradient**

It is a version of Gradient Descent, which updates the model parameters more frequently

It is fast technique, which performs one update at a time.

For example if we have 100 rows ina Dataseet, it will update weight 1000 times in SCG.

|  |  |
| --- | --- |
| Advantages | Disadvantages |
| It is Less time consuming since it has more frequency | Due to frequent updates, parameter updates have high variance, and will keep overshooting due to the frequent fluctuations |
| No need to store the Loss Values (Cost Function). So requires less memory | It reduce the learning rate , in order to get the same convergence |
| it helps us to discover new and possibly better local minima |  |

Formula

**θ=θ−η⋅∇J (θ;x(i);y(i)) ,**

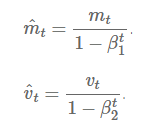
**where {x(i) ,y(i)} are the training examples.**

**Adam**

Adam stands for **Adaptive Moment Estimation.**

Adaptive Moment Estimation (Adam) is another method that computes adaptive learning rates for each parameter.

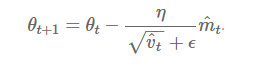
**Formula**



First equation represents Mean(first moment)

Second equation represents Varaince of Gradients(Second Moment).

final formula for the Parameter update is —



The values for β1 is 0.9 , 0.999 for β2, and (10 x exp(-8)) for ϵ.

**Advantages**

* Adam works well in practice and compares favourably to other adaptive learning-method algorithms
* It converges very fast and the learning speed of the Model is quite fast and efficient.
* It rectifies every problem that is faced in other optimization techniques such as vanishing Learning rate , slow convergence or High variance in the parameter updates which leads to fluctuating Loss function
* Adam was designed to combine the advantages of Adagrad, which works well with sparse gradients, and RMSprop, which works well in on-line settings

**Disadvantages**

* The Cost of Computation is very high.