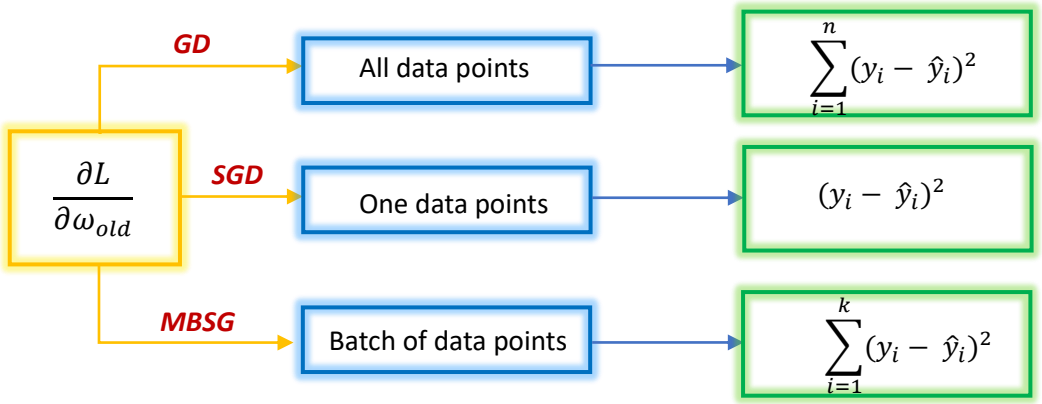
Different Types Of Gradient Decent Techniques Working Nature

1. Key points in weight initialization
2. Weight initialization techniques

1. How All GD Techniques Works:

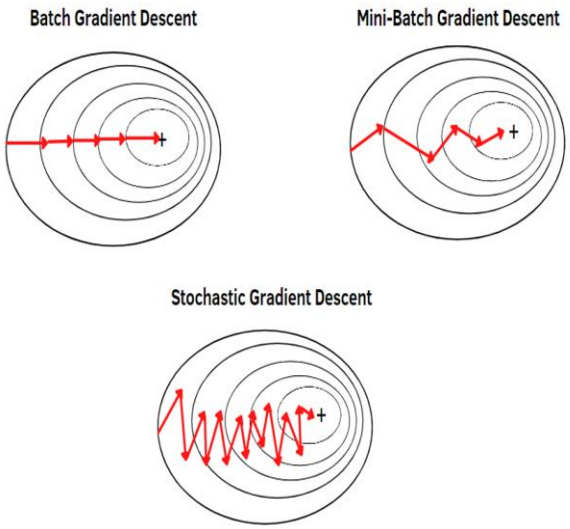
- In *Gradient Descent* all data points will consider in $\frac{\partial L}{\partial \omega_{old}}$ calculation.
- In *Stochastic Gradient Descent* (SGD) only one data point will consider in $\frac{\partial L}{\partial \omega_{old}}$ calculation. In SGD, ANN takes only one data point for forward propagation and calculate $\frac{\partial L}{\partial \omega_{old}}$, Then update the weights in back propagation. This process repeats for all data points in data set.
- Mini batch stochastic gradient descent considers a batch of data points (k) from the dataset to calculate the $\frac{\partial L}{\partial \omega_{old}}$. It will do the forward propagation with a batch of data points k and calculate $\frac{\partial L}{\partial \omega_{old}}$, then update the weights in back propagation. This process rebates for all batches of the data set.



- The most popular technique is MBSGD. Because it is more accurate than GD and SGD.

1.1. Global Minima Path Loss Function:

- The GED path is straightforward to get the global minimum in loss function.
- But SGD and MBSGD follows a jig-jog path to get the global minimum. Due to this jig-jog path it has some noise. Momentum SGD overcome this noise problem.



2. Momentum SGD: