

BATCH VS STOCHAISTIC GRADIENT DESCENT

In both gradient descent (GD) and stochastic gradient descent (SGD), you update a set of parameters in an iterative manner to minimize an error function.

While in GD, you have to run through **ALL** the samples in your training set to do a single update for a parameter in a particular iteration, in SGD, on the other hand, you use **ONLY ONE** or SUBSET of training sample from your training set to do the update for a parameter in a particular iteration.

If you use SUBSET, it is called Minibatch Stochastic gradient Descent.

Thus, if the number of training samples are large, in fact very large, then using gradient descent may take too long because in every iteration when you are updating the values of the parameters, you are running through the complete training set.

On the other hand, using SGD will be faster because you use only one training sample and it starts improving itself right away from the first sample.

SGD often converges much faster compared to GD but the error function is not as well minimized as in the case of GD. Often in most cases, the close approximation that you get in SGD for the parameter values are enough because they reach the optimal values and keep oscillating there.



