AdaGrad

AdaGrad is a well-known optimization method that is used in ML and DL. Duchi, Hazan, and Singer proposed it in 2011 as a way of adjusting the learning rate during training.

Sparse data refers to datasets with many features with zero values. It can cause problems in different fields, especially in machine learning.

AdaGrad's concept is to modify the learning rate for every parameter in a model depending on the parameter's previous gradients.

Specifically, it calculates the learning rate as the sum of the squares of the gradients over time, one for each parameter. This reduces the learning rate for parameters with big gradients while raising the learning rate for parameters with modest gradients.

$$v_t^w = v_{t-1}^w + (\nabla w_t)^2$$

$$\left[\nabla w_t = \frac{\partial L}{\partial w}\right]$$

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{v_t^w + \epsilon}} * \nabla w_t$$

$$v_t^b = v_{t-1}^b + (\nabla b_t)^2$$

$$b_{t+1} = b_t - \frac{\eta}{\sqrt{v_t^b + \epsilon}} * \nabla b_t$$

$(\nabla w_t)^2$ = past goadient's sum

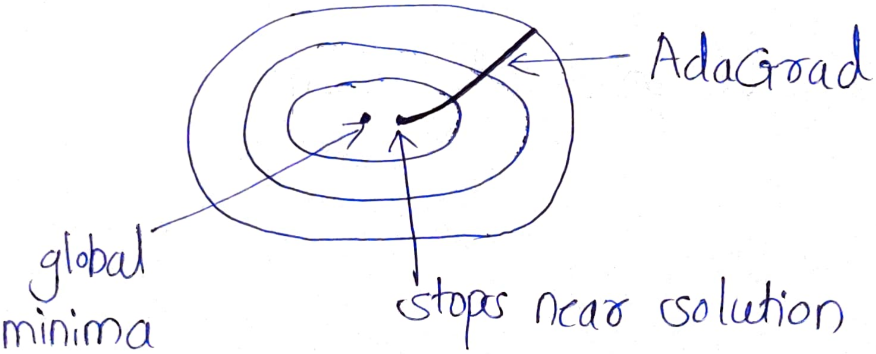
This implies that for parameters with big gradients, the learning rate is lowered, while for parameters with small gradients, the learning rate is raised.

The Adagrad algorithm is particularly useful for dealing with sparse data, where some of the input features have low frequency or are missing. In these cases, Adagrad is able to adaptively adjust the learning rate of each parameter, which allows for better handling of the sparse data.

- Due to sparse data, graph of parameters to loss becomes elongated bowl.
- If data is not sparse, the shape is circular.

Disadvantage:

1) As the number of epochs increases, learning rate decreases due to which updates are very small near to the solution.


AdaGrad
global minima
stops near solution