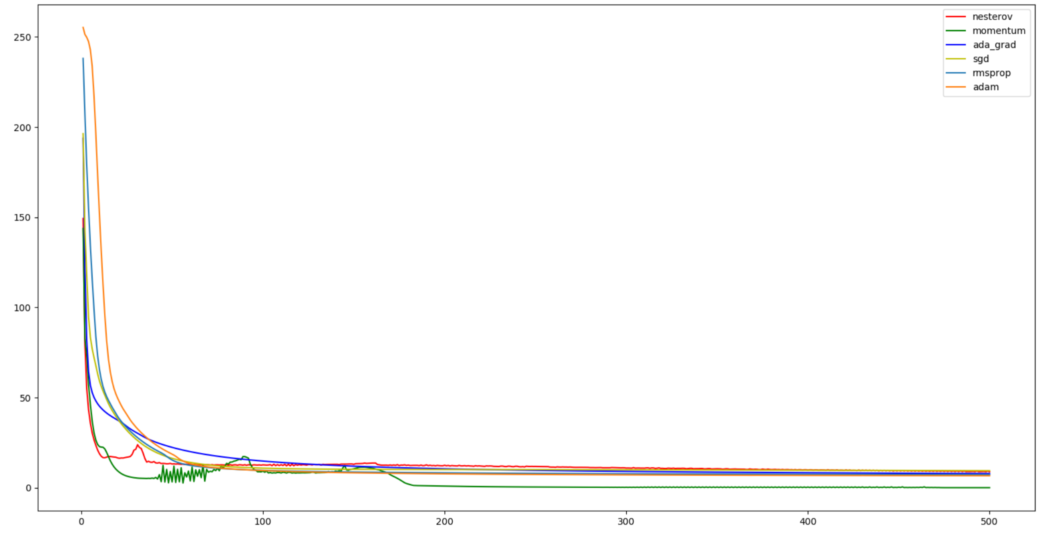
REPORT

EE16BTECH11023
HAVISH

The comparision plots of all the 6 optimizers can be found below. The x-axis is number of epochs and the y-axis is the total loss.



Note : Decay was not applied in the case of SGD since dataset is small.
The following are some of my observations about the working of these optimizers:

- Ada-grad, RMSProp and Adam optimizers can tolerate higher learning rates when compared to momentum, nesterov and SGD. (The loss gets trapped in a valley in the case of momentum, nesterov and SGD).

- Of all the optimizers, it can be observed that SGD has slower convergence rate and Adam has higher convergence rate for a given learning rate.

- The reason for the above might be because in SGD we are naively updating the gradient whereas in other cases, our updates are dependent on the previous step.

- However, all of these observations are made on a very small dataset having 120 training and 30 test samples. So, most of the above comparisions need not be consistent when done on other datasets.