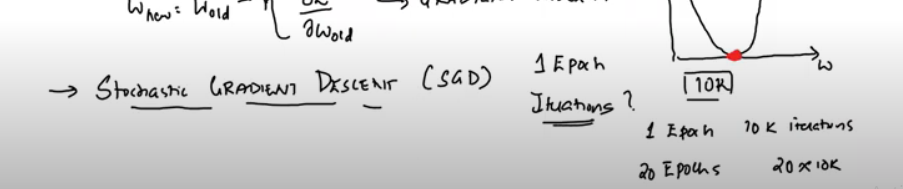


In case of 1million huge records weight intialization and updation occurs for both forward propogation and backward propogation this takes computationally more time to covergence(global minima) hence we go for SGD(Stochastic Gradient Descent)

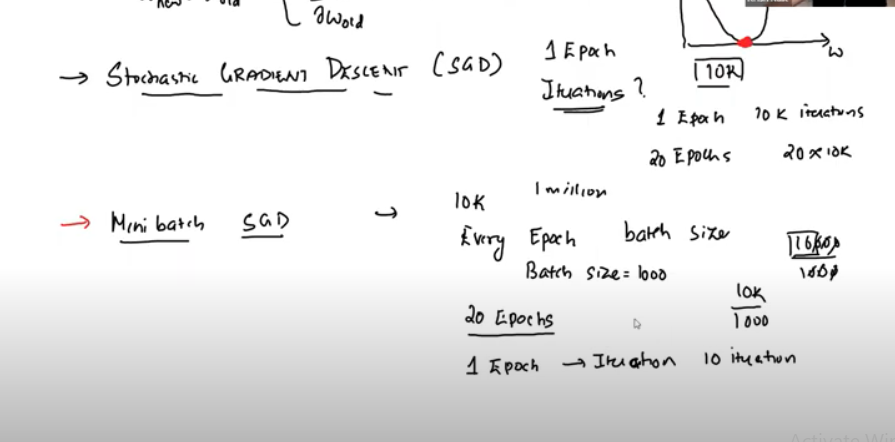


In case of stochastic gradient descent also more time to get to global minima

In stochastic gradient descent only one record is taken and do forward and back ward propogations

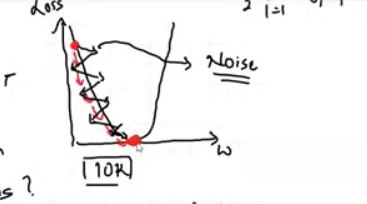
Then it takes more time same as gradient descent but difference is that we don’t require more RAM

And it is not computationally expensive like gradient descent

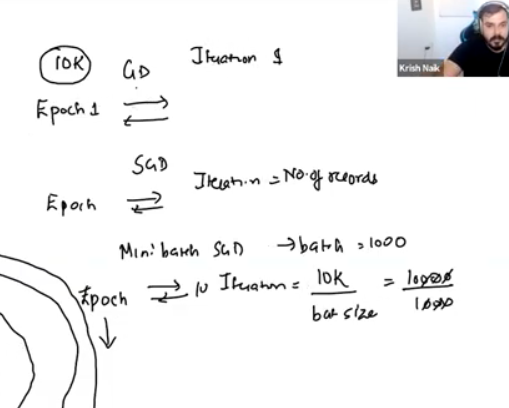
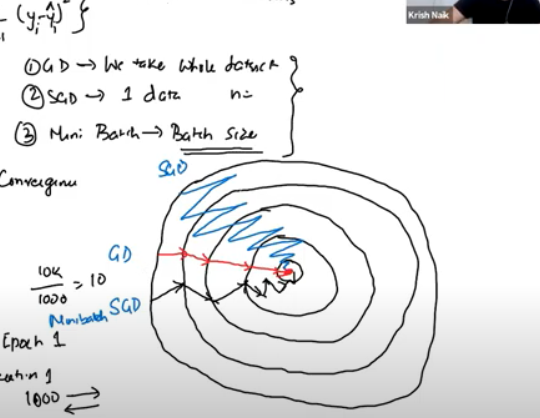
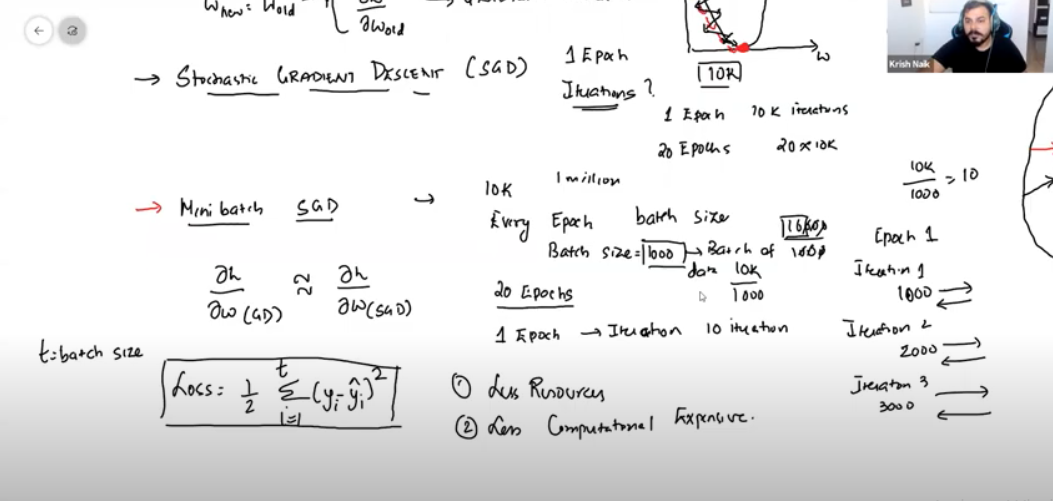
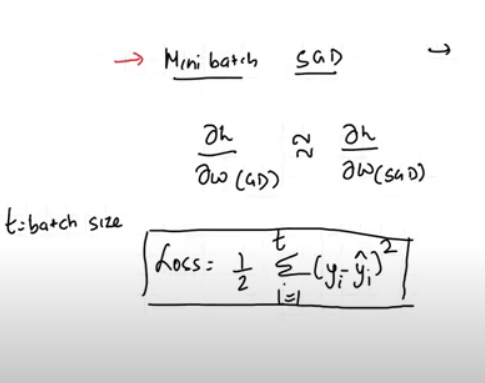
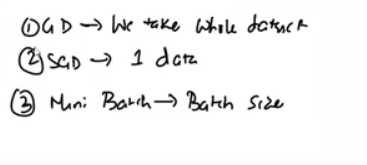


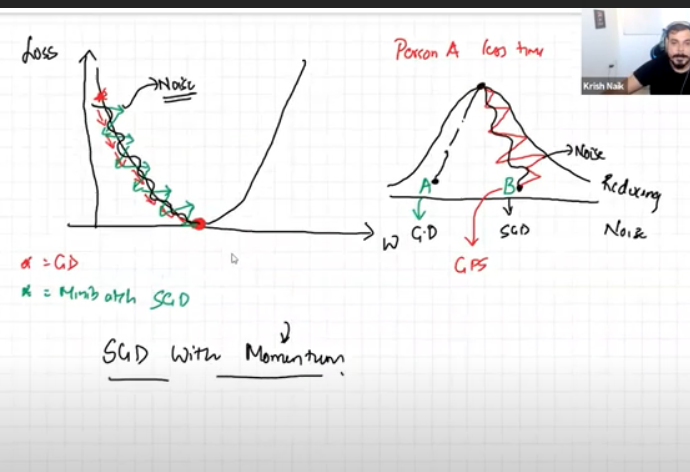
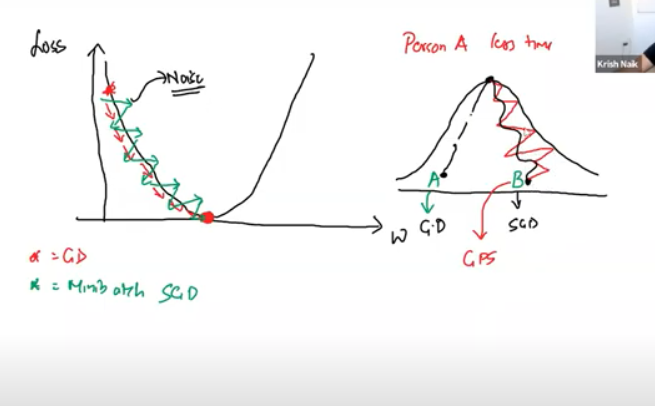
For mini batch SGD we maintain batch size so that for example if batch size=1000 then for 10k records

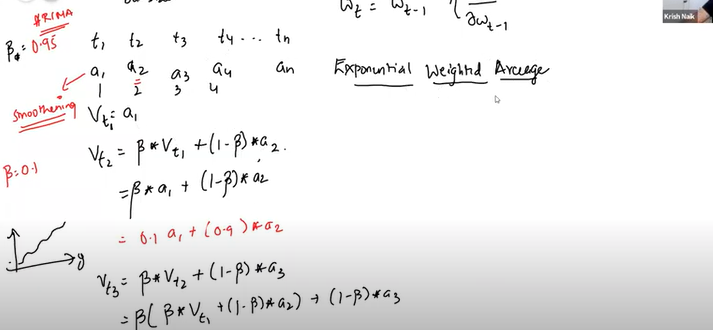
10k/batch size = 10 iterations this make the covergence time fast compare to SGD AND Gradient Descent



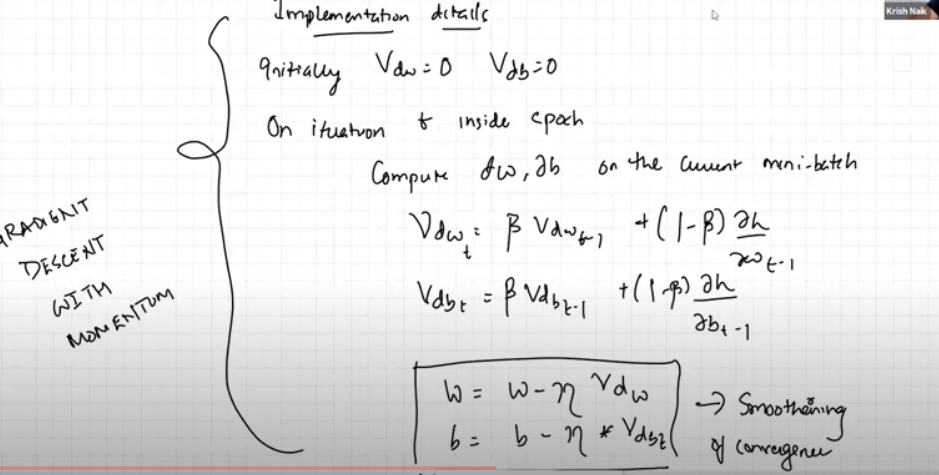
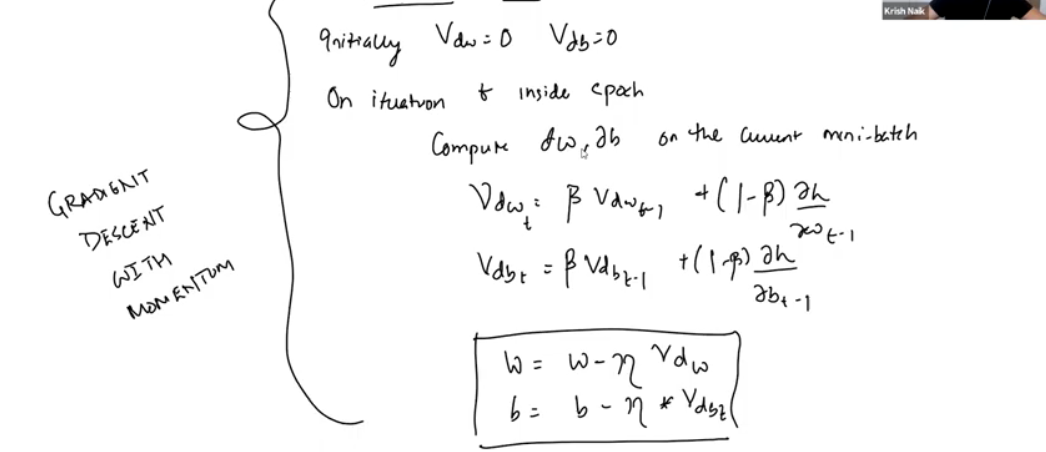
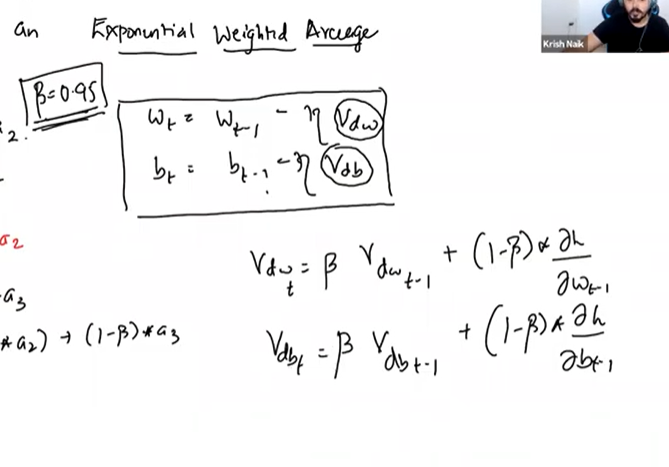
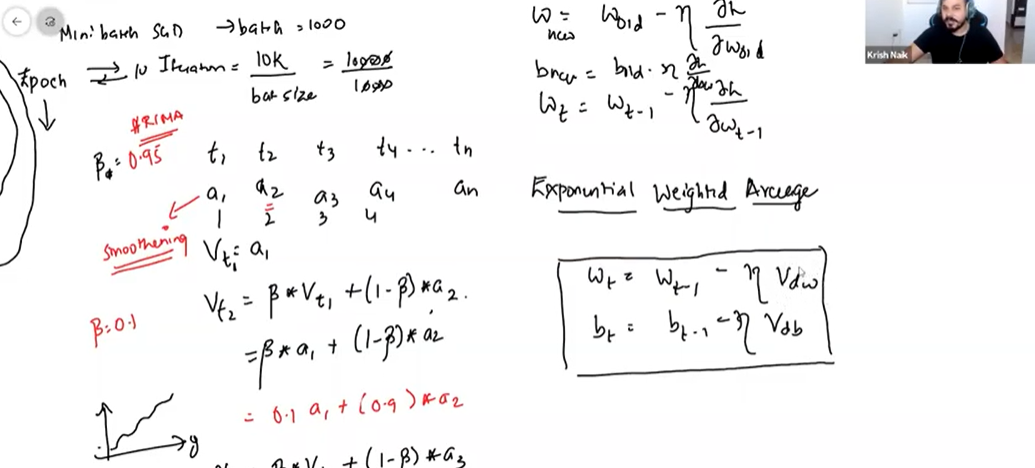
In case of mini batch SGD weight updation undergo as zig zag this is called Noise

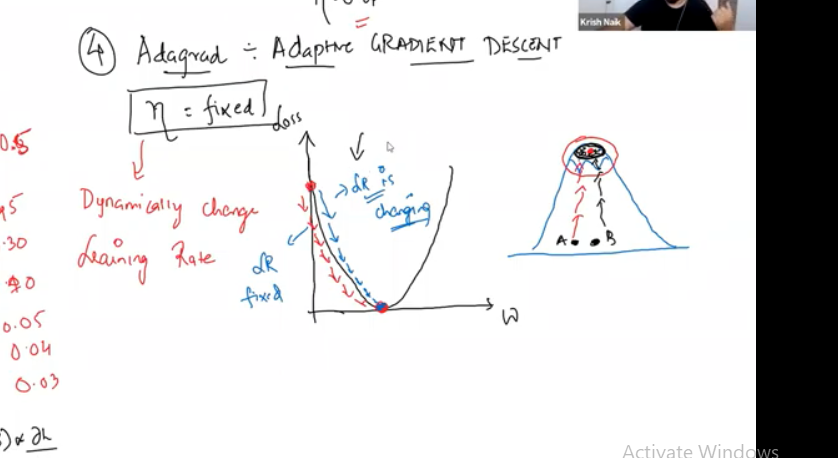


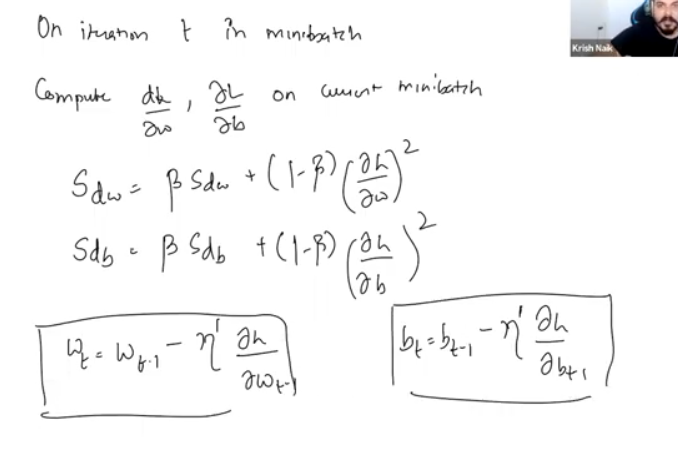
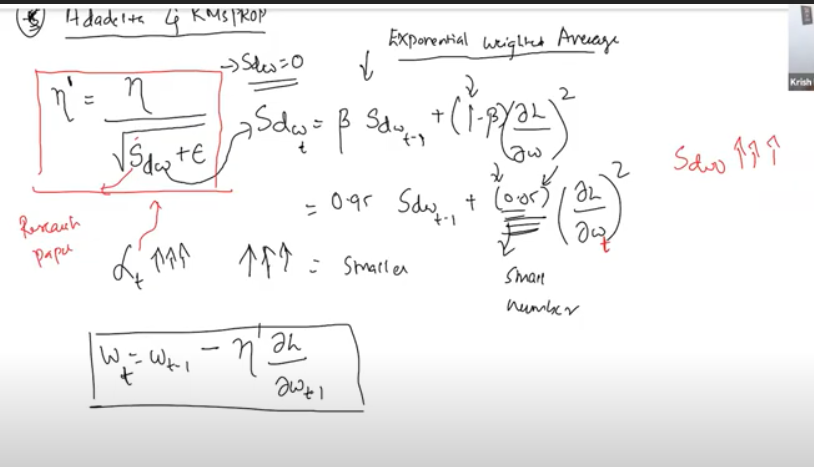
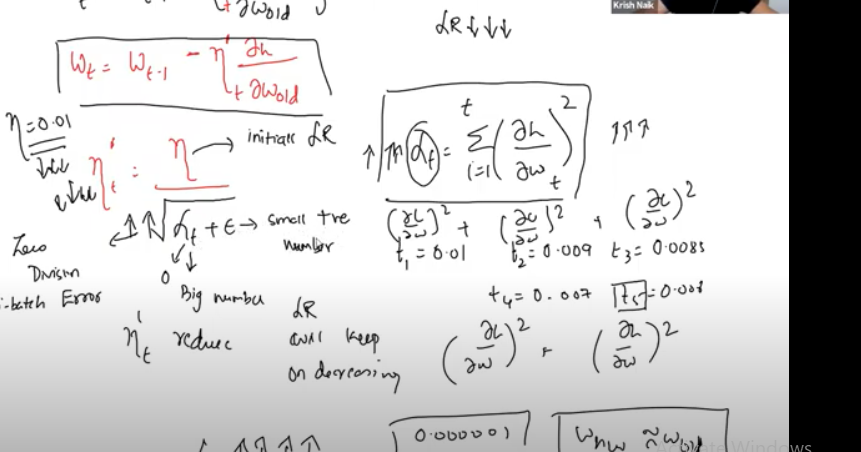
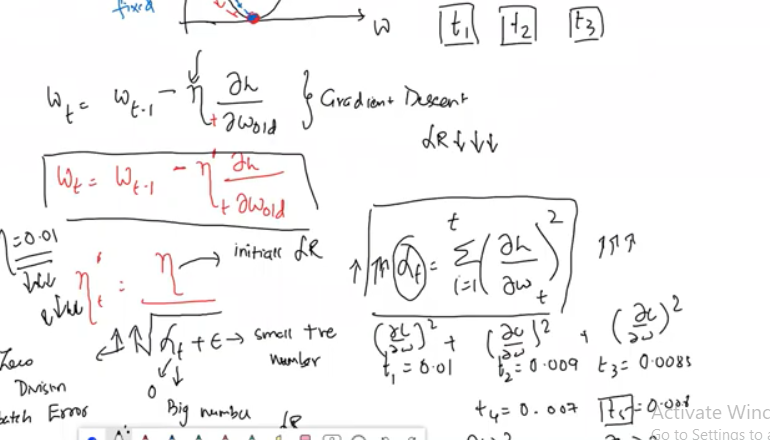


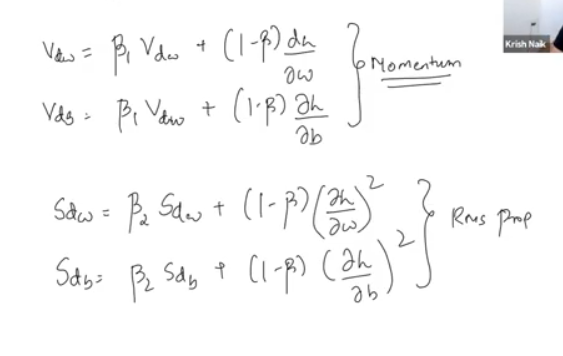
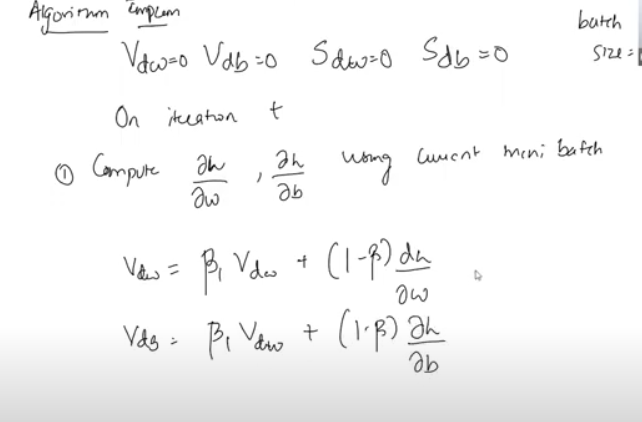
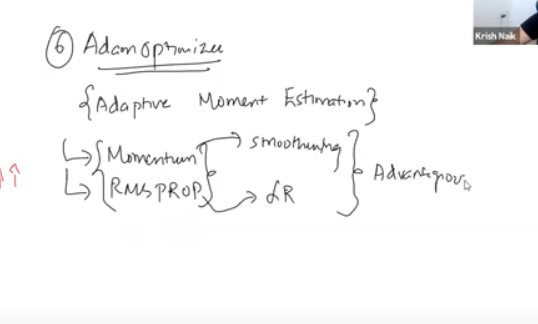


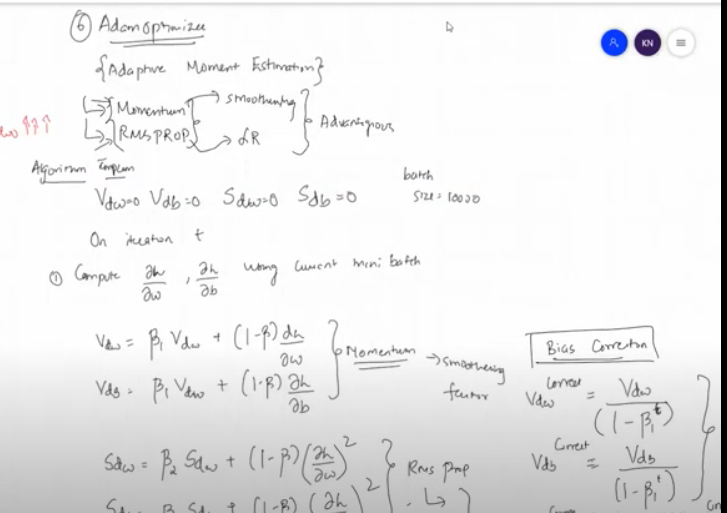
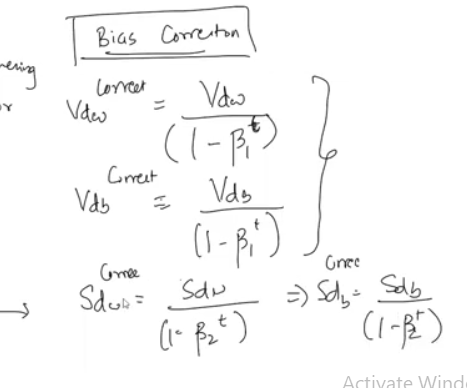
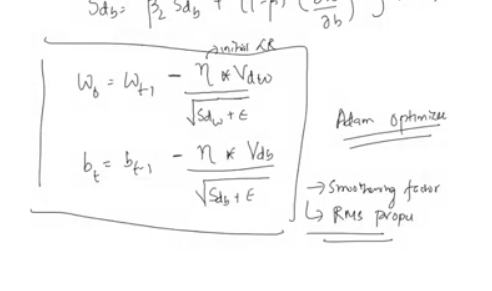
These equations gives a smoothing curver same concept is applied in case of weights in case of SGD optimizers to smoothen the zigzag curve this is called as Exponential Weighted Average





In this optimizer we change the learning rate dynamically to reach the global minima fastly





Adam optimizer is well used optimizer every where in this we are combining the momentum smoothing factor and Rms prompt (learning rate change in dynamical) hence it is well used optimizer