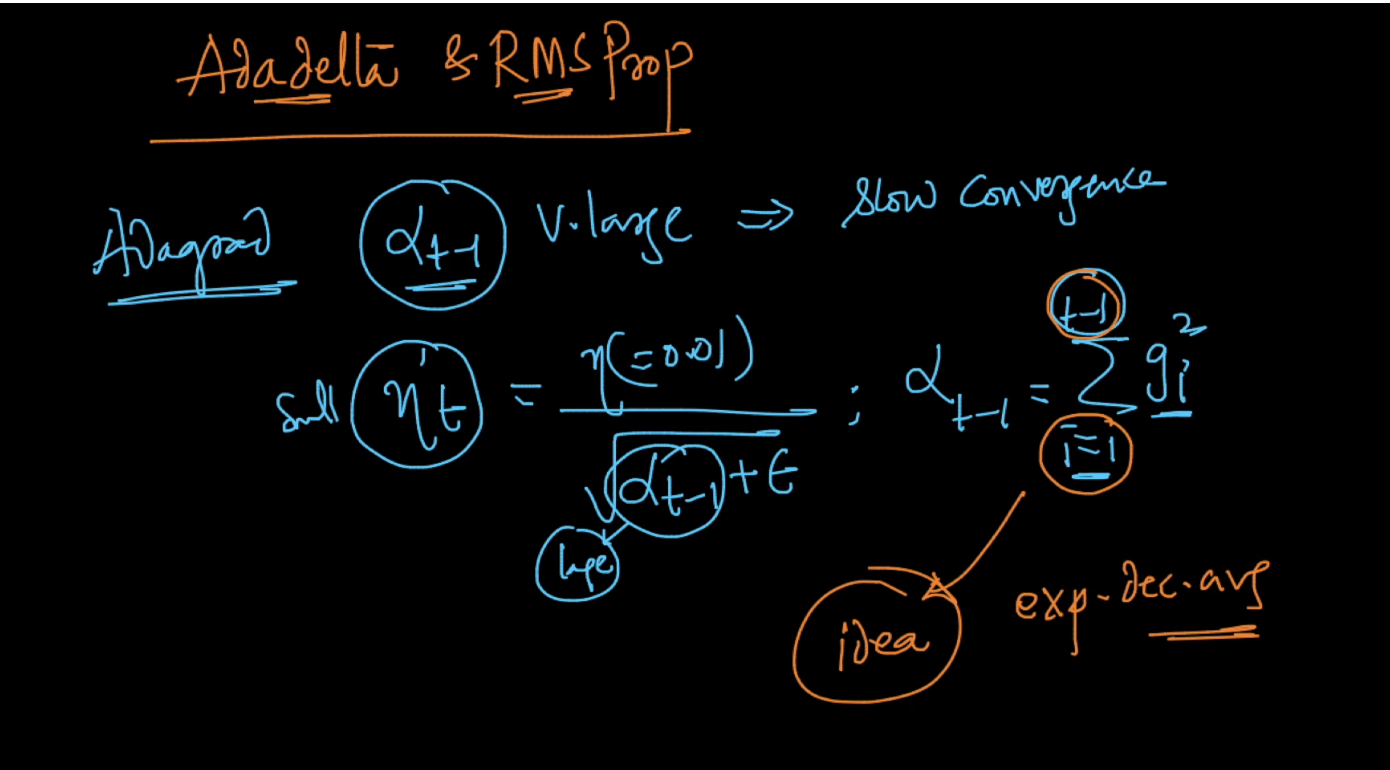
**Optimizers : Adadelta andRMSProp**

Adadelta is almost similar as Adagrad except it uses exponential weighted average gi2 instead of sum of gi2 we use in Adagrad so as to avoid large denominator in eta’t and thus avoid problem of slow convergence.

Below image shows adagrad



Now in Adadelta instead of alpha in eta we use eda(exponential decaying average which is shown in below image)

So eta’ for adadelta is eta’ = eta/ sqrt(edat-1 + epsilon)

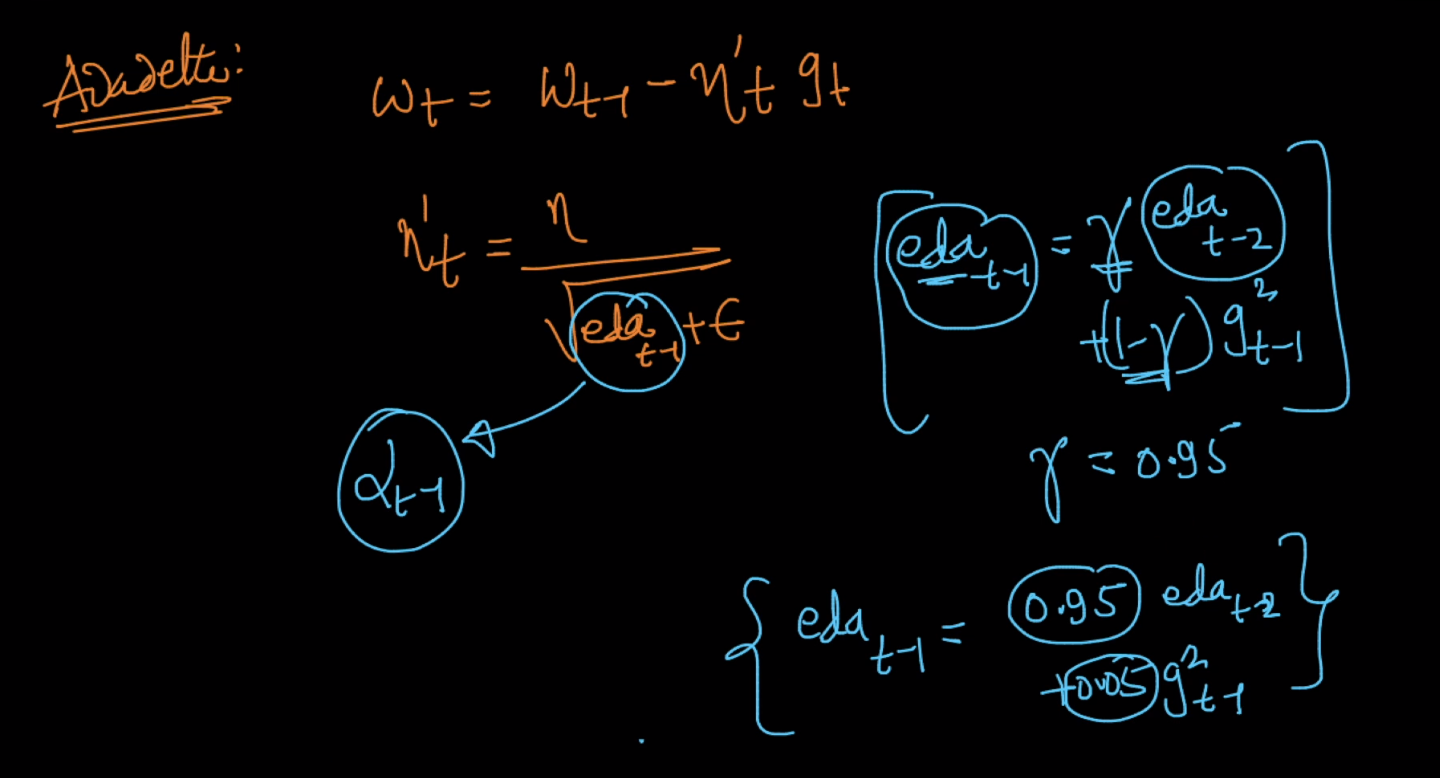
Here edat-1 = gamma\*edat-2 + (1-gamma) \* g2t-1

And typically we use gamma = 0.95

And the equation of edat-1 is recursive

### NOTE:  In Adadelta, eda(0) = 0.

### the denominator should contain eda\_t and not eda\_(t-1).

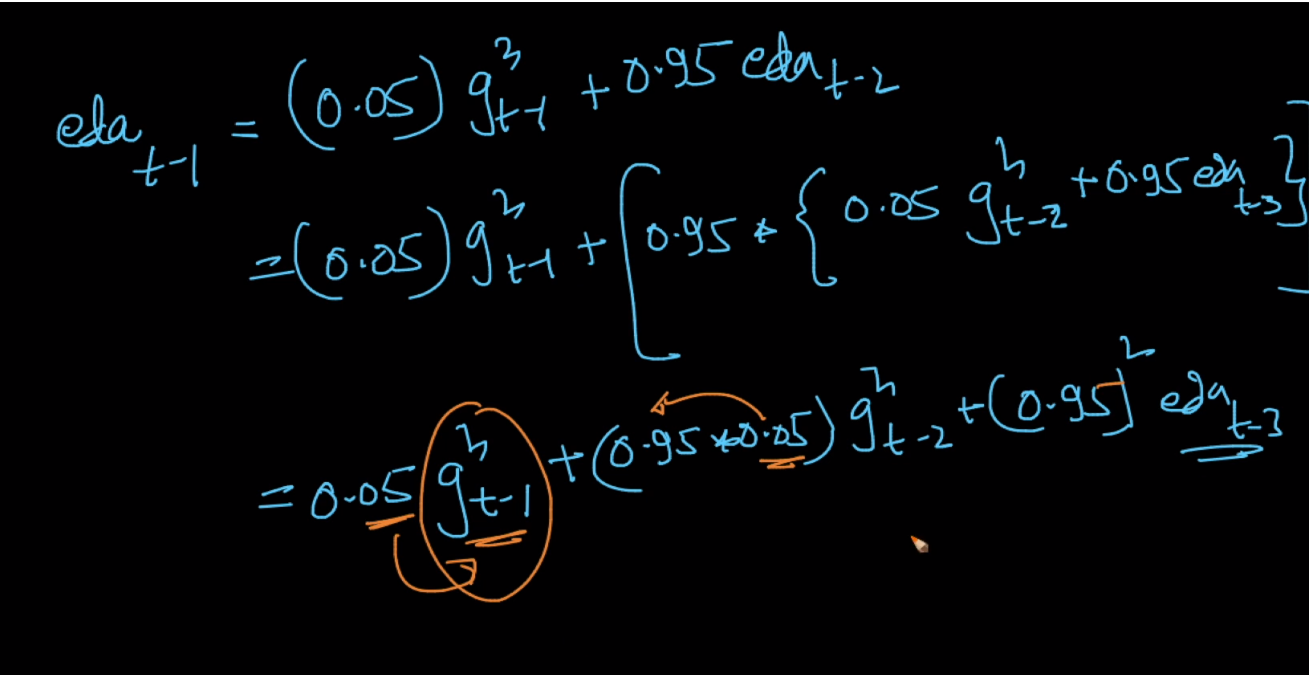


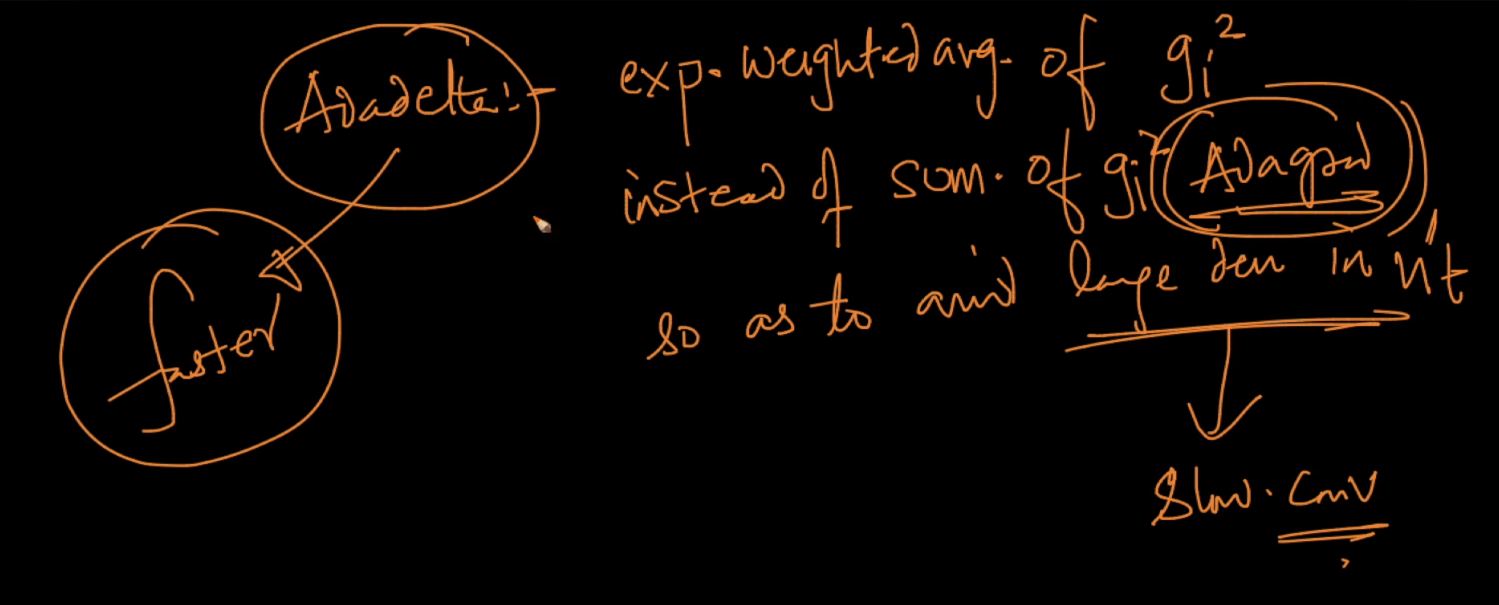
Below image shows the equation it is recursive but main thing to focus is that it gives exponential weighted average to gradient term i.e max gamma to gt-1 term, less to gt-2 and so on.

And by this we eda doesn’t increase too much and it resolves the problem of slow convergence which comes in adagrad.

Adadelta have same advantages as adagrad and adadelta is faster than adagrad.

RMS prop is little diff. than Adadelta but it works similar to adadelta





Link : <http://ruder.io/optimizing-gradient-descent/index.html#rmsprop>

Comments :

