**Nesterov Accelerated Gradient (NAG)**

In Batch sgd+momentum what we do is it go in the direction of gradient from Wt-1 and it also go in the direction of momentum from Wt-1 but eventually it goes in the direction of Wt which is nothing but sum of those two (momentum and gradient) and if look from vector then it is the resultant vector of that two vector.

Now In **NAG** what happen is we go in the direction of momentum from Wt-1 and suppose we reach at W’, now take gradient from here and go in that direction and we reach at Wt.

i.e eventually we reach from Wt-1 to Wt.

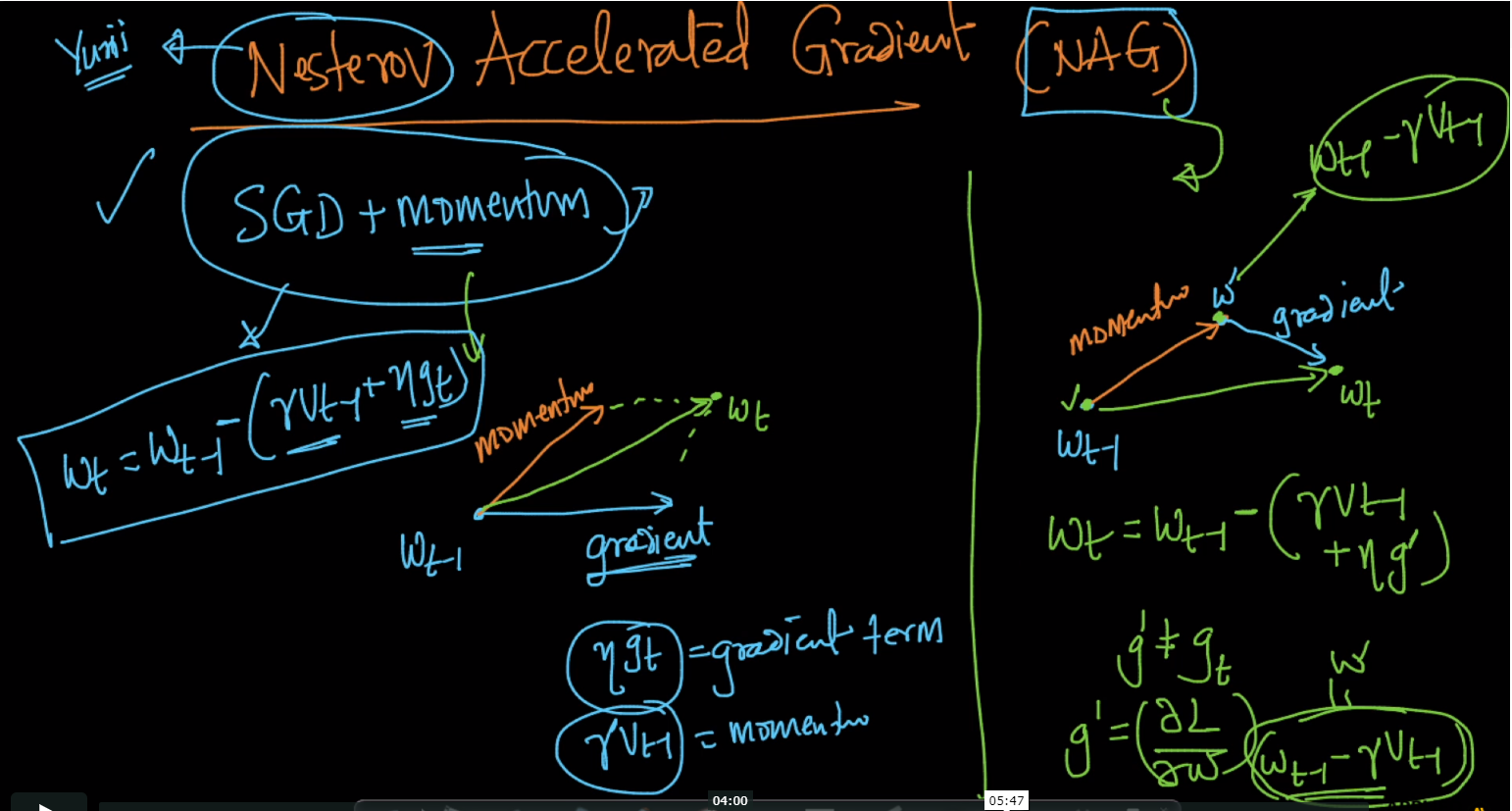
But the gradient g’ we took at W’ in NAG is different than we took gradient gt at Wt-1 in SGD + momentum

So in NAG as we go from Wt-1 in the direction of momentum we reach at W’ its value is

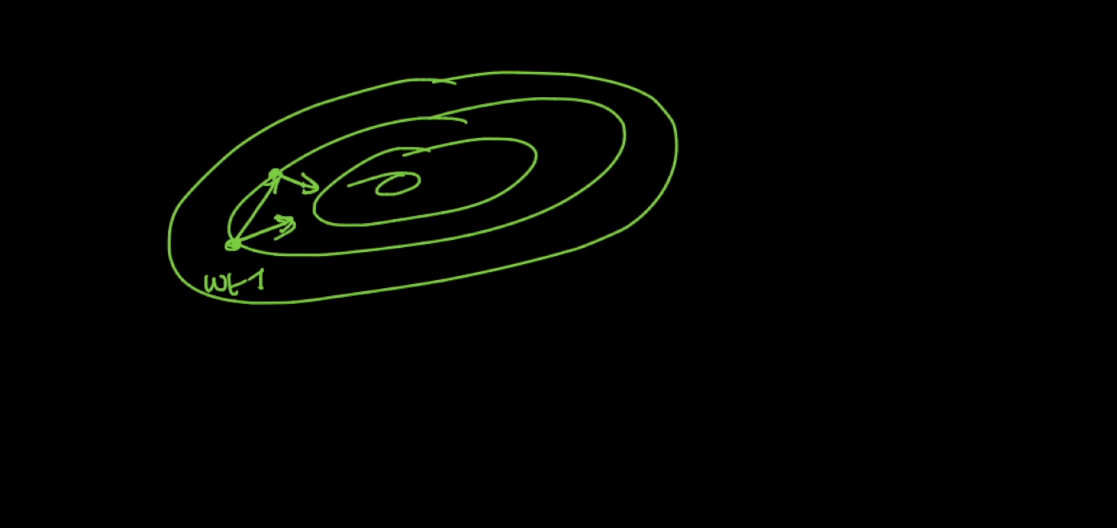
W’ = Wt-1 – gamma\*Vt-1 , now we take gradient here i.e g’ = del\_l/del\_w at w’ (Wt-1 – gamma\*Vt-1)

Now by this our update eq. looks like

Wt = Wt-1 – (gamma\*Vt-1 + eta \* g’)

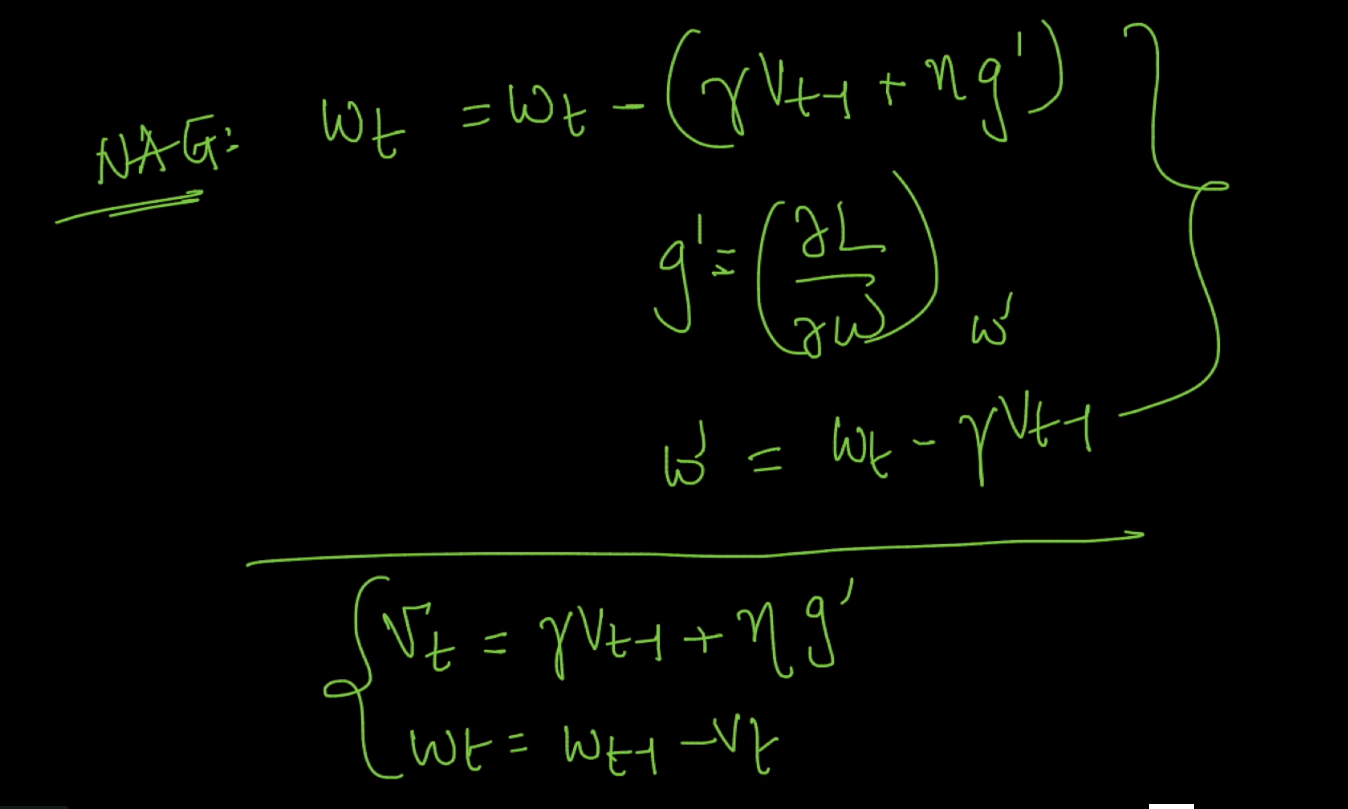


Below image shows that g’ and gt ­both are different



Below is the clear equation of NAG

Note: in below correct W’ = Wt-1 – gamma\*Vt-1



Links : <http://cs231n.github.io/neural-networks-3/>

Comments :

