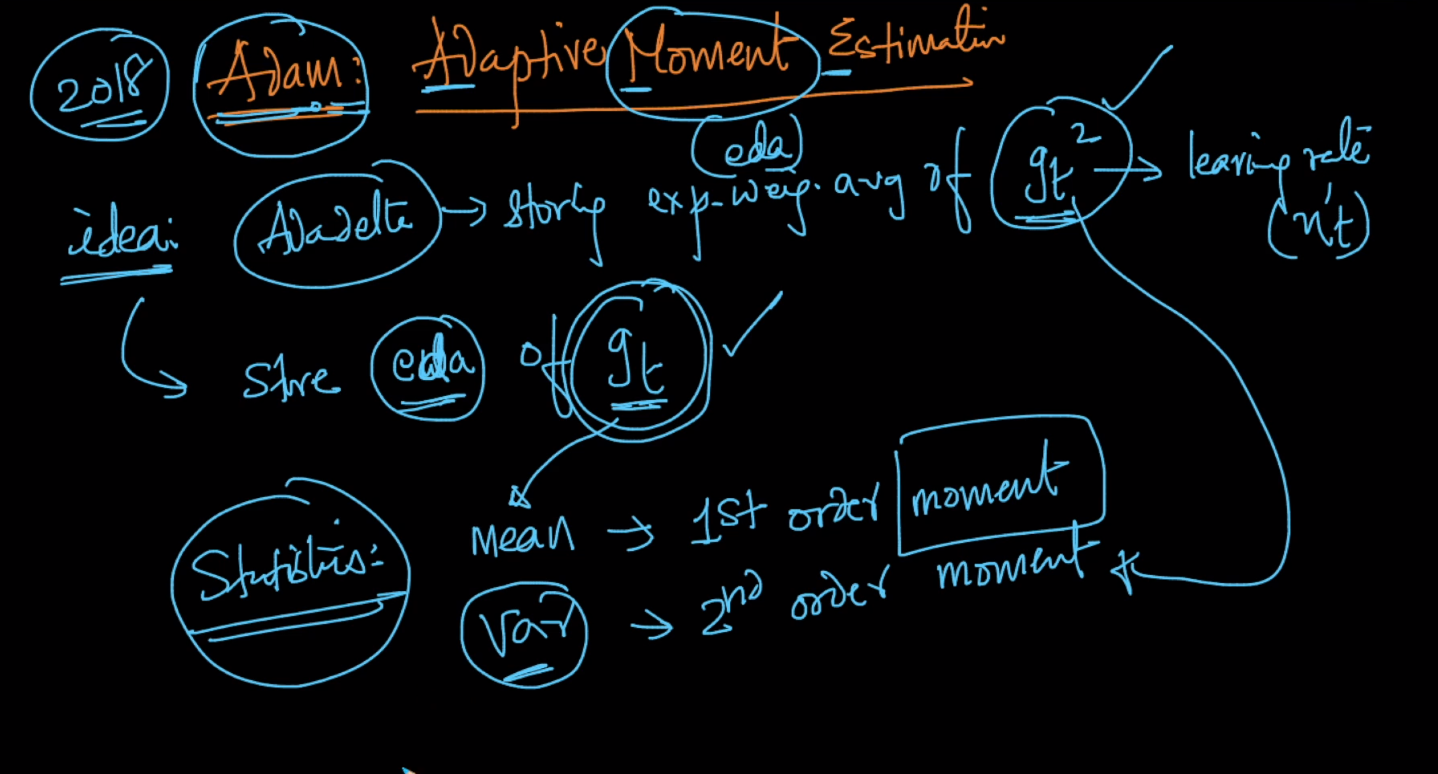
**Adam**

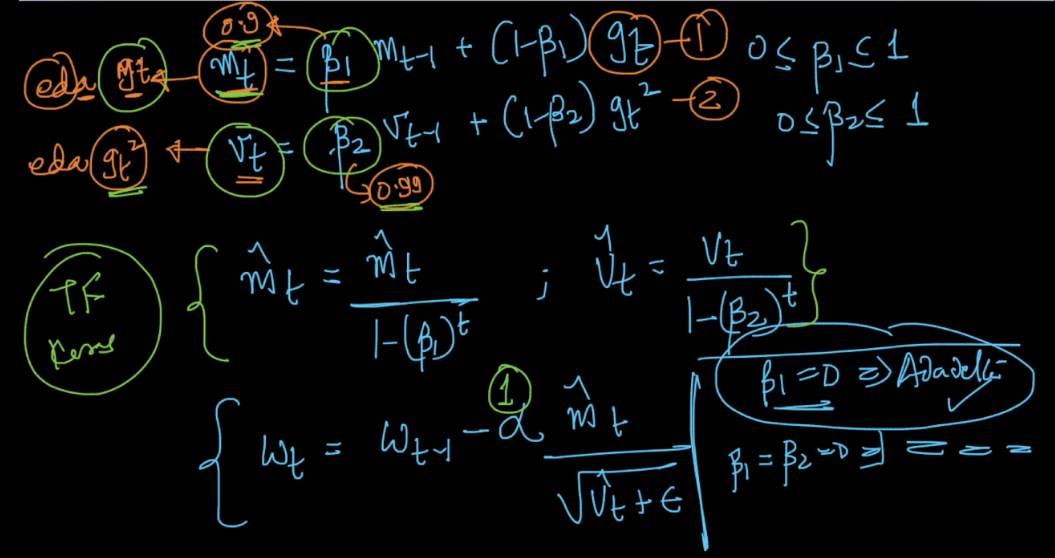
Adaptive Moment Estimation (Adam)  is another method that computes adaptive learning rates for each parameter. In addition to storing an exponentially decaying average of past squared gradients vt like Adadelta and RMSprop, Adam also keeps an exponentially decaying average of past gradients mt, similar to momentum

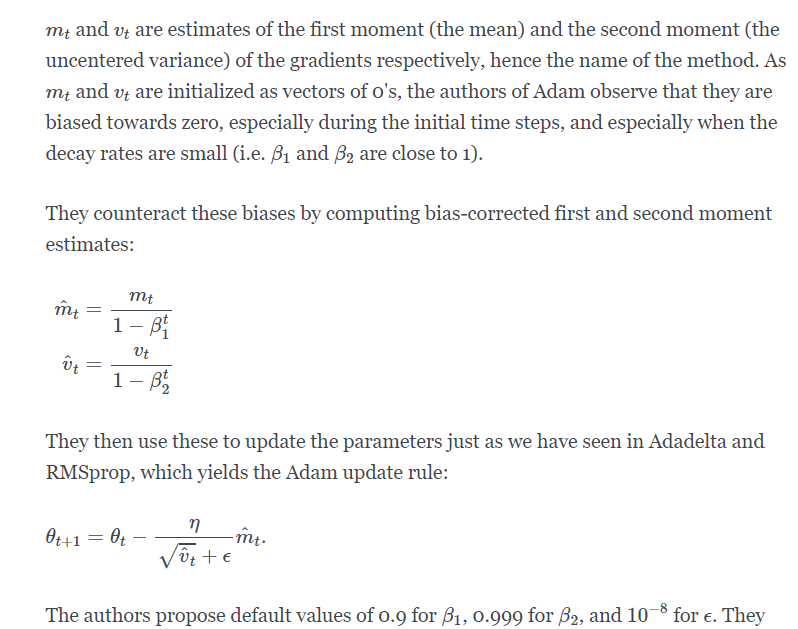
Here vt is similar to eda in adadelta(which is called 2nd order momentum) and mt is similar to momentum(which is called 1st order momentum).

We can think of Adam as combination of adadelta and SGD with momentum.

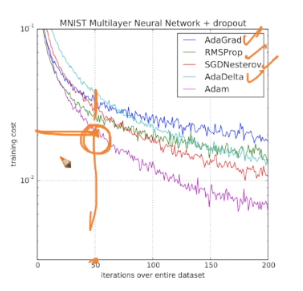
If we remove 1st order momentum then it become adadelta and if we remove 2nd order momentum then it become SGD with momentum.



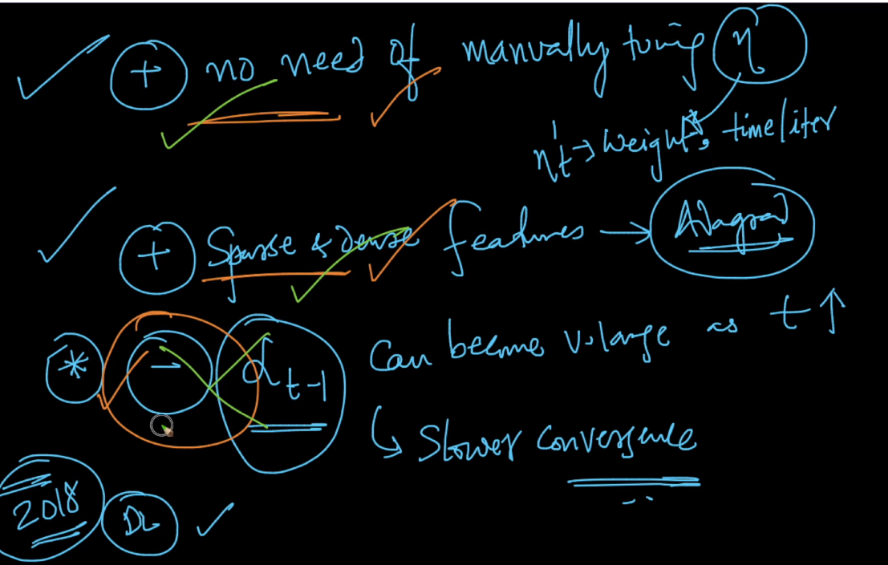




ADAM works well as compare to other adaptive learning methods.



And it covers all the below issues



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