Refer: http://ruder.io/optimizing-gradient-descent/

Correction: From 4:48 min onwards, it should be alpha_t in the formula and NOT alpha_(t-1)