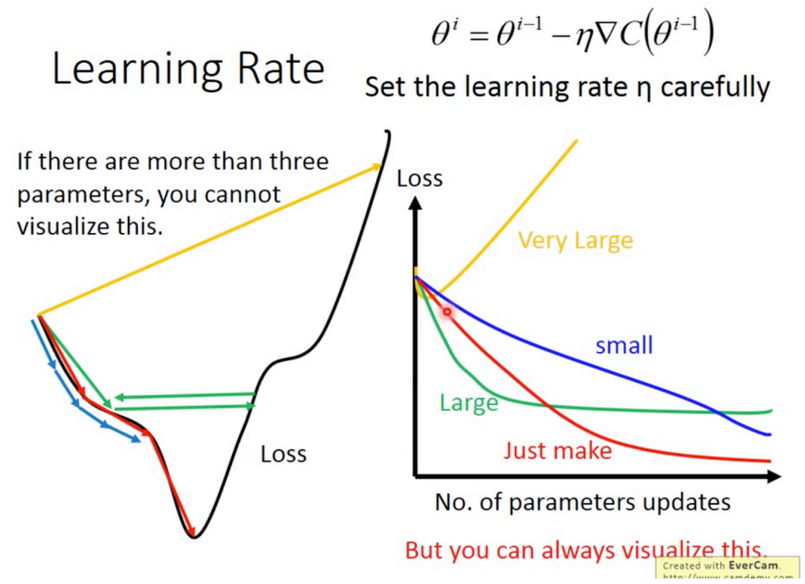
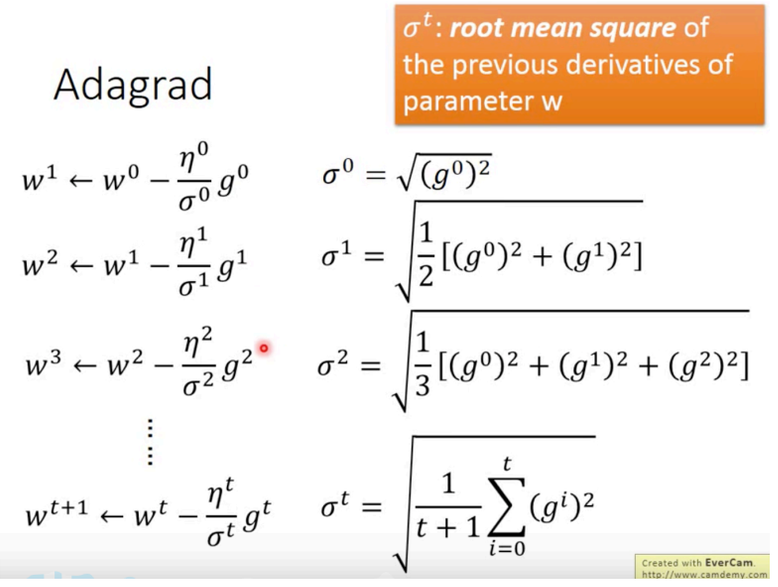
#学习率太大，可能错过最低点



#Adagrad （Adaptive Gradient） 自适应梯度下降



让学习率去除以一个数，使学习率逐渐变小

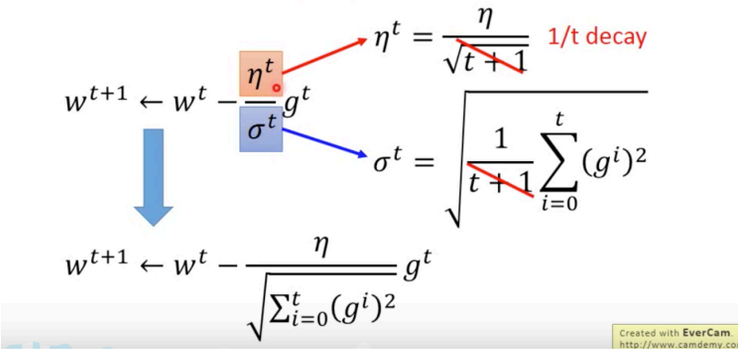
每次学习后累加梯度g，并求平均

这里除σ，σ的计算见上图

最终式

Adagrad

- Divide the learning rate of each parameter by the ***root mean square of its previous derivatives***



最终式，这里η下面加一个与σ中一样的系数，来抵消sqrt(t+1)