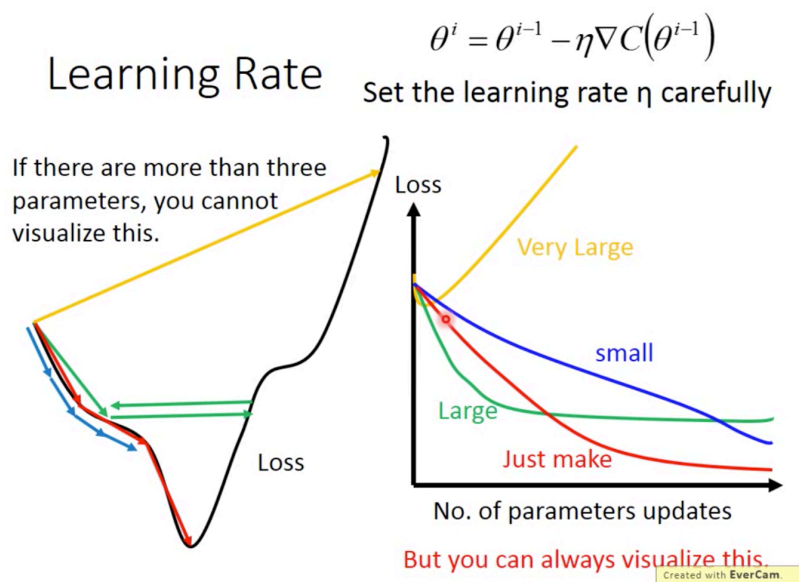


#学习率太大，可能错过最低点



#Adagrad (Adaptive Gradient) 自适应梯度下降

Adagrad

σ^t : root mean square of the previous derivatives of parameter w

$$w^1 \leftarrow w^0 - \frac{\eta^0}{\sigma^0} g^0 \quad \sigma^0 = \sqrt{(g^0)^2}$$

$$w^2 \leftarrow w^1 - \frac{\eta^1}{\sigma^1} g^1 \quad \sigma^1 = \sqrt{\frac{1}{2} [(g^0)^2 + (g^1)^2]}$$

$$w^3 \leftarrow w^2 - \frac{\eta^2}{\sigma^2} g^2 \quad \sigma^2 = \sqrt{\frac{1}{3} [(g^0)^2 + (g^1)^2 + (g^2)^2]}$$

$$\vdots$$

$$w^{t+1} \leftarrow w^t - \frac{\eta^t}{\sigma^t} g^t \quad \sigma^t = \sqrt{\frac{1}{t+1} \sum_{i=0}^t (g^i)^2}$$

Created with EverCam
http://www.camdemy.com

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**

$$w^{t+1} \leftarrow w^t - \frac{\eta^t}{\sigma^t} g^t$$
$$\eta^t = \frac{\eta}{\sqrt{t+1}} \quad \text{1/t decay}$$
$$\sigma^t = \sqrt{\frac{1}{t+1} \sum_{i=0}^t (g^i)^2}$$
$$w^{t+1} \leftarrow w^t - \frac{\eta}{\sqrt{\sum_{i=0}^t (g^i)^2}} g^t$$

Created with EverCam.
<http://www.camdemy.co>

最终式，这里 η 下面加一个与 σ 中一样的系数，来抵消 $\sqrt{t+1}$

