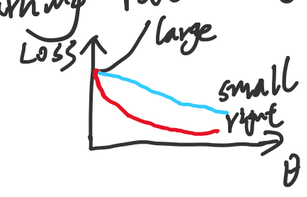
# 3 Gradient Descent

2020年5月20日 星期三　　下午2:42

梯度下降：　$\nabla L(\theta) = \begin{bmatrix} \partial L(\theta_1)/\partial \theta_1 \\ \partial L(\theta_2)/\partial \theta_2 \end{bmatrix}$

Tips：1. Learning rate　$\theta^i = \theta^{i-1} - \eta \nabla L(\theta^{i-1})$

Visualize Loss：




Q → 可以自动调整学习率吗？
1. 刚开始 η ↓ 很大
2. 之后 η 不太减了

技巧：自动梯度 adgrad

$$\begin{cases} w^{t+1} \leftarrow w^t - \eta^t g^t \\ w^{t+1} \leftarrow w^t - \frac{\eta^t}{\sigma^t} g^t \end{cases}$$

(adgrad — $\sigma^t$：root mean square 的 w 的偏导数的值

g：微分值：$\frac{\partial w^t}{\partial w}$

example：

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

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

⋮　　⋮

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

$\eta^t = \frac{\eta}{\sqrt{t+1}}$　decay

$w^{t+1} \leftarrow w^t - \frac{\eta}{\sqrt{\sum_{i=0}^{t}(g^i)^2}} g^t$　— 强调反差效果

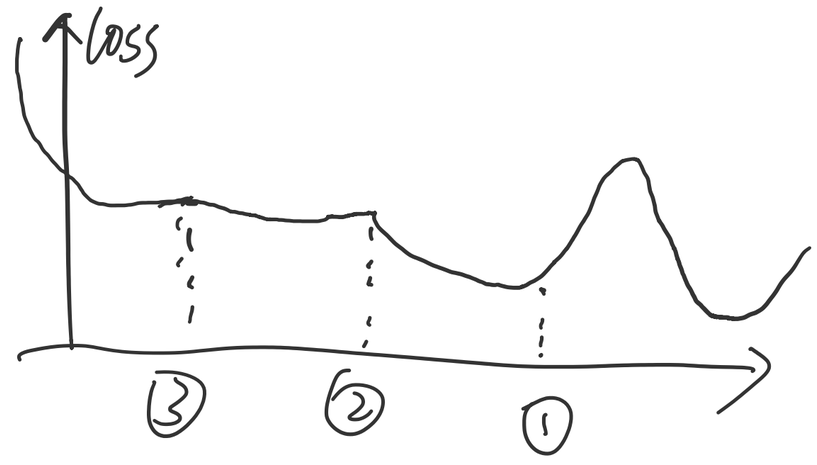
best step：$\left|\frac{\text{first derivative}}{\text{second derivative}}\right|$

问题：

1. Stuck local minima
2. Stuck saddle point

$\Big\} \frac{\partial L}{\partial w} = 0$

3. very slow at plateau　$\frac{\partial L}{\partial w} \approx 0$

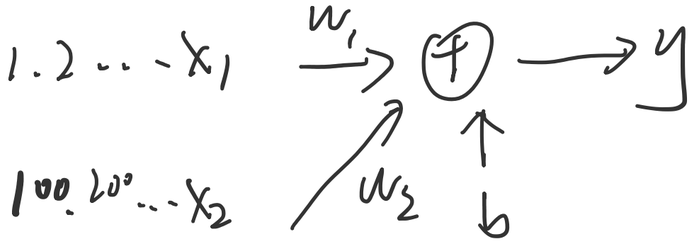



## 2. Stochastic Gradien Descene 随机梯度下降

pick an example：$x^n$

$L^n = (\hat{y}^n - (b + \sum w_i x_i^n))^2$　　$\theta^i = \theta^{i-1} - \eta \nabla L^n(\theta^{i-1})$

## 3. Feature Scaling 特征编码




将数据转换到相同维度

$$\begin{cases} x_i^r = \frac{x_i^r - m_i}{\sigma_i} \\ x_i^r = \frac{x_i - \min}{\max - \min} \end{cases}$$

$m_i$：均值

$\sigma_i$：标准差

最大最小化