

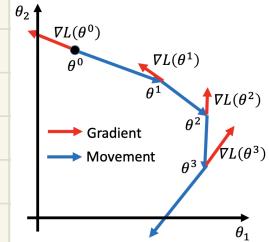
Chapter 3 Gradient Descent

1. Review GD

DGD的基本过程：已知模型 $f(\theta)$ ，损失函数 $L(\theta)$ ，目标 $\theta^* = \arg \min L(\theta)$

假设参数 θ 有 2 个变量 $\{\theta_1, \theta_2\}$ ，则梯度 $\nabla L(\theta) = \begin{bmatrix} \frac{\partial L(\theta)}{\partial \theta_1} \\ \frac{\partial L(\theta)}{\partial \theta_2} \end{bmatrix}$

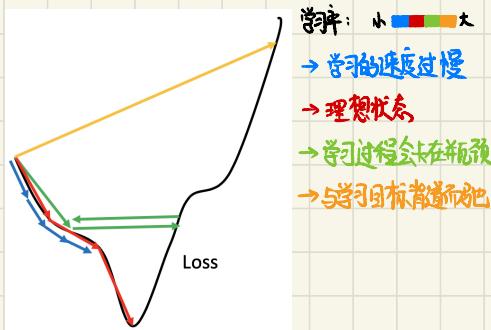
$$\theta' = \theta^0 - \eta \nabla L(\theta^0) \quad \theta^1 = \theta' - \eta \nabla L(\theta') \quad \dots$$



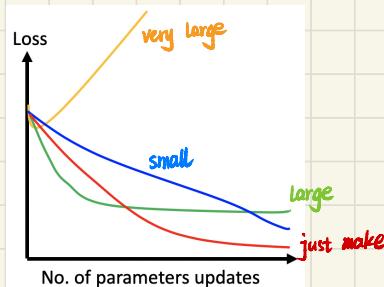
2. 调整 learning rate

① learning rate 对梯度下降的影响：

假设参数只有 2 个变量时，Loss Function 是二维可见的



当参数的维度大于 2 个时，损失函数不易绘制，可以观察误差随参数更新的变化规律



② 原则一：learning rate 随着参数的迭代不断减小（离目标远就大步前进，离目标近就小步前进）

$$E.g. \eta^t = \frac{1}{\sqrt{t+1}}$$

③ 原则二：因材施教，对每一个参数给予不同的 learning rate；以众多参数中的某个参数 w 为例：

$$\eta^t = \frac{1}{N^{t+1}} \quad g^t = \frac{\partial L(\theta^t)}{\partial w}$$

$$\text{Vanilla GD: } w^{t+1} \leftarrow w^t - \eta^t g^t$$

$$\text{Adagrad: } w^{t+1} \leftarrow w^t - \frac{\eta^t}{\sigma^t} g^t,$$

其中 σ^t 表示关于 w 所有微分的均方根

$$\begin{aligned} w^1 &\leftarrow w^0 - \frac{\eta^0}{\sigma^0} g^0 & \sigma^0 &= \sqrt{(g^0)^2} \\ w^2 &\leftarrow w^1 - \frac{\eta^1}{\sigma^1} g^1 & \sigma^1 &= \sqrt{\frac{1}{2}[(g^0)^2 + (g^1)^2]} \\ w^3 &\leftarrow w^2 - \frac{\eta^2}{\sigma^2} g^2 & \sigma^2 &= \sqrt{\frac{1}{3}[(g^0)^2 + (g^1)^2 + (g^2)^2]} \\ &\vdots \\ w^{t+1} &\leftarrow w^t - \frac{\eta^t}{\sigma^t} g^t & \sigma^t &= \sqrt{\frac{1}{t+1} \sum_{i=0}^t (g^i)^2} \end{aligned}$$

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

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

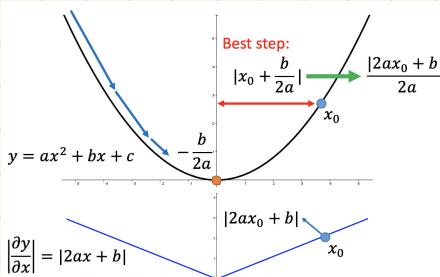
$$\sigma^t = \sqrt{\frac{1}{t+1} \sum_{i=0}^t (g^i)^2}$$

large gradient, large step
large gradient, small step } 相互矛盾?

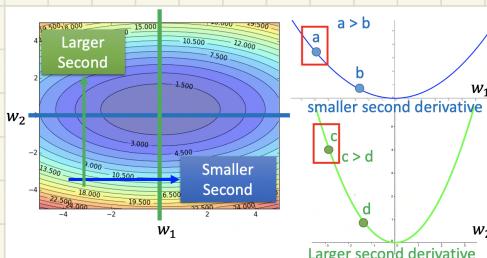
在 Adagrad 中，两项梯度的直观看起来是出现强弱梯度的“反差”



Large gradient, Large step 的适用范围：



- 对于一个参数而言，越大的一阶微分值代表离 Minima 越远
- 分子 $2ax$ 恰好等于 $\frac{\partial^2 y}{\partial x^2}$
- Best step : $\frac{\text{First derivative}}{\text{Second derivative}}$



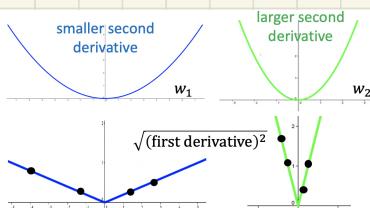
- 对于 w_1 , $g'(a) > g'(b) \rightarrow a$ 离此点更远离 minima
- 对于 w_2 , $g'(c) > g'(d) \rightarrow c$ 离此点更远离 minima
- 对于 w_1 和 w_2 , $g'(c) > g'(a) \rightarrow$ 但 a 离更远离 minima
- 一阶微分 / = 二阶微分有关

Adagrad 如何使用 “best step” :

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

best step: $\frac{|-\text{一阶微分}|}{\text{二阶微分}}$

在不增加额外计算量的情况下，使用一阶微分估计二阶微分



3. Stochastic Gradient Descent

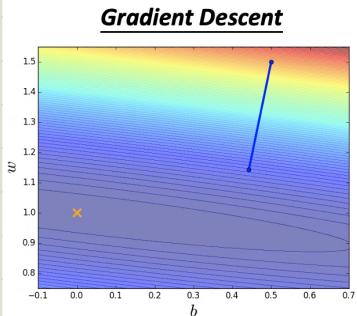
① 基本原理

在最基础的梯度下降过程中损失函数的定义包含了所有样本，即 $L = \sum_i (\hat{y}^i - (b + \sum_j w_j x_j^i))^2$, $\theta^i = \theta^{i-1} - \eta \nabla L(\theta^{i-1})$

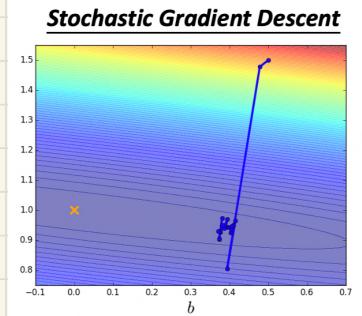
在随机梯度下降中，只选取一个样本计算损失函数：

$$\text{选取样本 } i, L^i = (\hat{y}^i - (b + \sum_j w_j x_j^i))^2 \quad \theta^i = \theta^{i-1} - \eta \nabla L^i(\theta^{i-1})$$

② 例子



使用20个样本，按照标准梯度建议的方向完成了1次下降

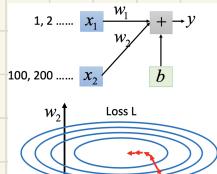


对于20个样本，每次使用1个样本进行一次下降，共完成了20次下降。在GD完成一次下降时，SGD完成了20次（天下武功，唯快不破）

4. Feature Scaling

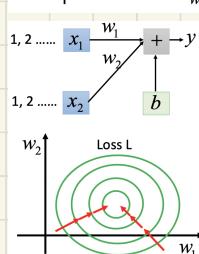
① 特征缩放的基本方法：尽可能的让不同维度的特征具有相同规模的幅度

② 进行特征缩放的原因：以 $y = w_1 x_1 + w_2 x_2 + b$ 为例

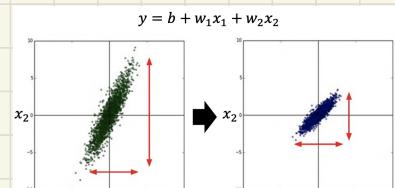


• 如果特征 x_2 的尺度远大于 x_1 ，那么对于一个 Δw 来说，

w_2 对 y 产生的影响也会远大于 w_1 对 y 产生的影响，所以损失函数在 w_2 方向上会更平滑，在 w_1 方向上会更陡峭，梯度下降初期会偏向于 w_2 的方向，而不是朝着 minima



• 如果特征 x_2 的尺度与 x_1 一致，则 x_1 和 x_2 对 y 的作用是相同的，在梯度下降时两者作用相当，则正好朝向 minima 的方向

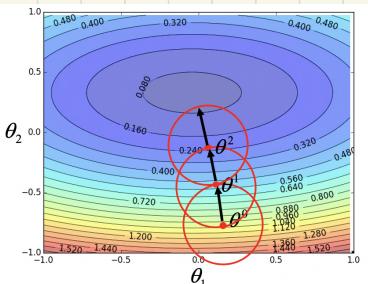


③梯度下降的常用策略：见 chapter 1 的 Homework 部分 (max-min 标准化 / Z-score 标准化)

5. Gradient Descent Theory

① 寻找极小值的形式化方法

假设损失函数是关于 θ 的函数，而 θ 有两个变量 $\{\theta_1, \theta_2\}$ ，下图是 $L(\theta)$ 的等值线，给定初始点 θ^0 ，寻找 $\theta^* = \arg \min L(\theta)$



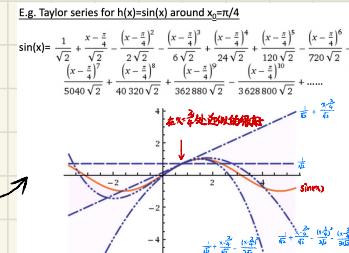
- 以 θ^0 为圆心确定一个范围的圆，寻找圆中可以使 $L(\theta)$ 变小的值 θ' ，并步进至 θ'
- 重复上述步骤直至某一次以 θ^n 为圆心的圆中没有可以使 $L(\theta)$ 更小的 θ ，此时 $\theta^* = \theta^n$

补充：泰勒级数

假设 $h(x)$ 在 $x=x_0$ 附近无限可微，今求 $h(x)$ 在 x_0 处的阶数，则

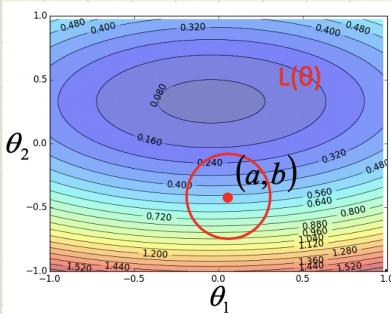
$$h(x) = \sum_{k=0}^{\infty} \frac{h^{(k)}(x_0)}{k!} (x-x_0)^k = h(x_0) + h'(x_0)(x-x_0) + \frac{h''(x_0)}{2!} (x-x_0)^2 + \dots$$

当 x 很接近 x_0 时，因为 $(x-x_0) \gg (x-x_0)^2$ ，所以 $h(x) \approx h(x_0) + h'(x_0)(x-x_0)$



对于多变量的泰勒级数： $h(x,y) = h(x_0, y_0) + \frac{\partial h(x_0, y_0)}{\partial x} (x-x_0) + \frac{\partial h(x_0, y_0)}{\partial y} (y-y_0) + (x-x_0)^2 + (y-y_0)^2$ 相关项 + ...

当 x 和 y 接近 x_0 和 y_0 时， $h(x,y) = h(x_0, y_0) + \frac{\partial h(x_0, y_0)}{\partial x} (x-x_0) + \frac{\partial h(x_0, y_0)}{\partial y} (y-y_0)$



假设在 (a,b) 点周围围绕一个足够小的圆，则损失函数可展开为泰勒级数：

$$L(\theta) \approx L(a,b) + \frac{\partial L(a,b)}{\partial \theta_1} (\theta_1 - a) + \frac{\partial L(a,b)}{\partial \theta_2} (\theta_2 - b)$$

令 $S = L(a,b)$ ， $U = \frac{\partial L(a,b)}{\partial \theta_1}$ ， $V = \frac{\partial L(a,b)}{\partial \theta_2}$ ， $S/U/V$ 为梯度

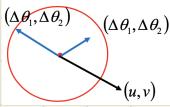
$$L(\theta) \approx S + U(\theta_1 - a) + V(\theta_2 - b)$$

问题转换为在红圈中找出 θ_1 和 θ_2 使得 $L(\theta)$ 最小，其中 $(\theta_1 - a)^2 + (\theta_2 - b)^2 \leq d$

令 $\Delta \theta_1 = \theta_1 - a$, $\Delta \theta_2 = \theta_2 - b$ ，则 $L(\theta)$ 简化为 $L(\theta) \approx U \Delta \theta_1 + V \Delta \theta_2 \approx (U, V) \cdot (\Delta \theta_1, \Delta \theta_2)$

令 $(\Delta \theta_1, \Delta \theta_2)$ 与 (U, V) 方向相反，长度尽可能的长，两个量内积最小即

令 $\Delta\theta_1 = \theta_1 - a$, $\Delta\theta_2 = \theta_2 - b$, 则 $L(\theta)$ 可简化为 $L(\theta) = u\Delta\theta_1 + v\Delta\theta_2 \approx (u, v) \cdot (\Delta\theta_1, \Delta\theta_2)$, 即两个向量的内积



当 $(\Delta\theta_1, \Delta\theta_2)$ 与 (u, v) 方向相反, 长度大于圆的半径时, 两向量的内积最小, 即 $L(\theta)$ 最小。

$$\begin{bmatrix} \Delta\theta_1 \\ \Delta\theta_2 \end{bmatrix} = -\eta \begin{bmatrix} u \\ v \end{bmatrix} \rightarrow \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} - \eta \begin{bmatrix} u \\ v \end{bmatrix}, \text{ 其中 } \eta \text{ 尽可能的大, 使 } \begin{bmatrix} \Delta\theta_1 \\ \Delta\theta_2 \end{bmatrix} \text{ 长度大于圆半径}$$

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} - \eta \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} - \eta \begin{bmatrix} \frac{\partial L(a, b)}{\partial \theta_1} \\ \frac{\partial L(a, b)}{\partial \theta_2} \end{bmatrix}$$

此公式恰好是梯度下降的思路

上述公式恰好是梯度下降的基本思想, 但前提是让圆足够小, 即 η 足够小。所以可能出现 η 比较大的情况, 随着梯度下降误差却在增大。理论上, 如果在泰勒展开时, 将二阶微分考虑进来, 可以适当增大, 例如牛顿法。但在 Deep Learning 中常用 GD, 不考虑二阶微分, 因为计算代价过大。

6. More Limitation about GD

① 梯度下降的异常情况

