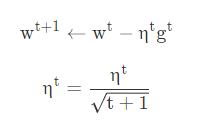
P6-梯度下降

## 自适应学习率

### 1.1 基本原理

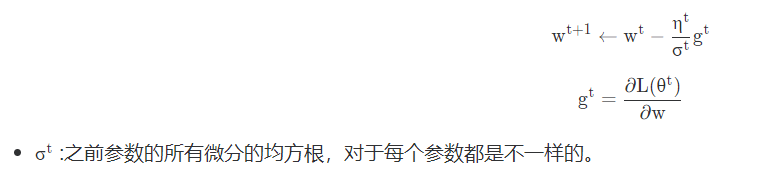


### 1.2 局限性

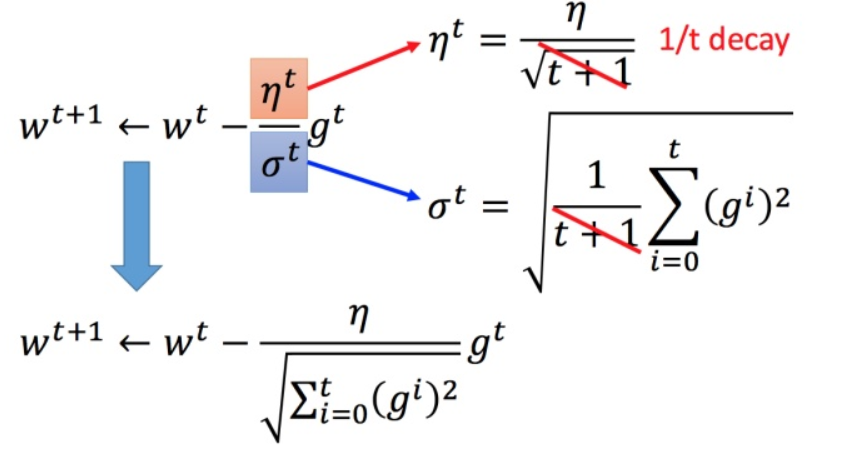
上面的t-dependent 形式的学习率其实是对任意的参数都使用相同的衰减率，应该对不同的参数使用不用的学习率

### 1.3adagrad

#### 1.3.1基本原理

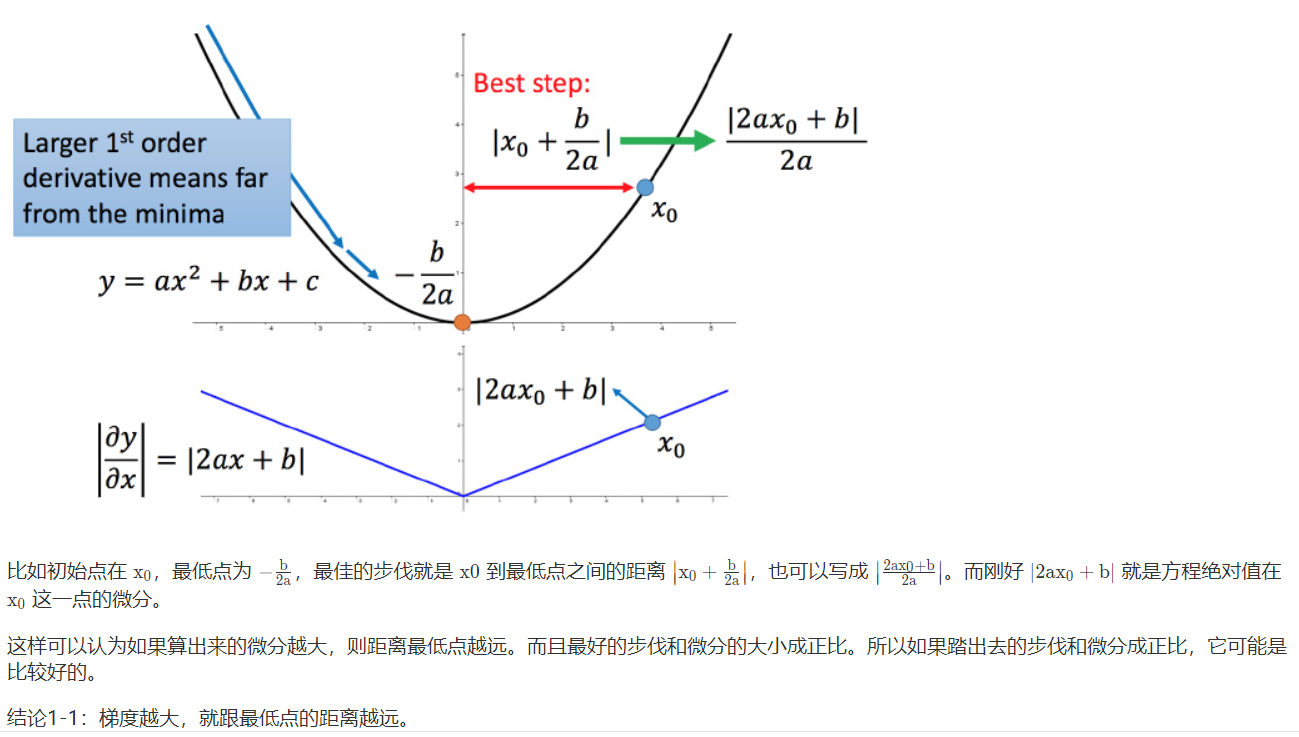


最终的表达形式：



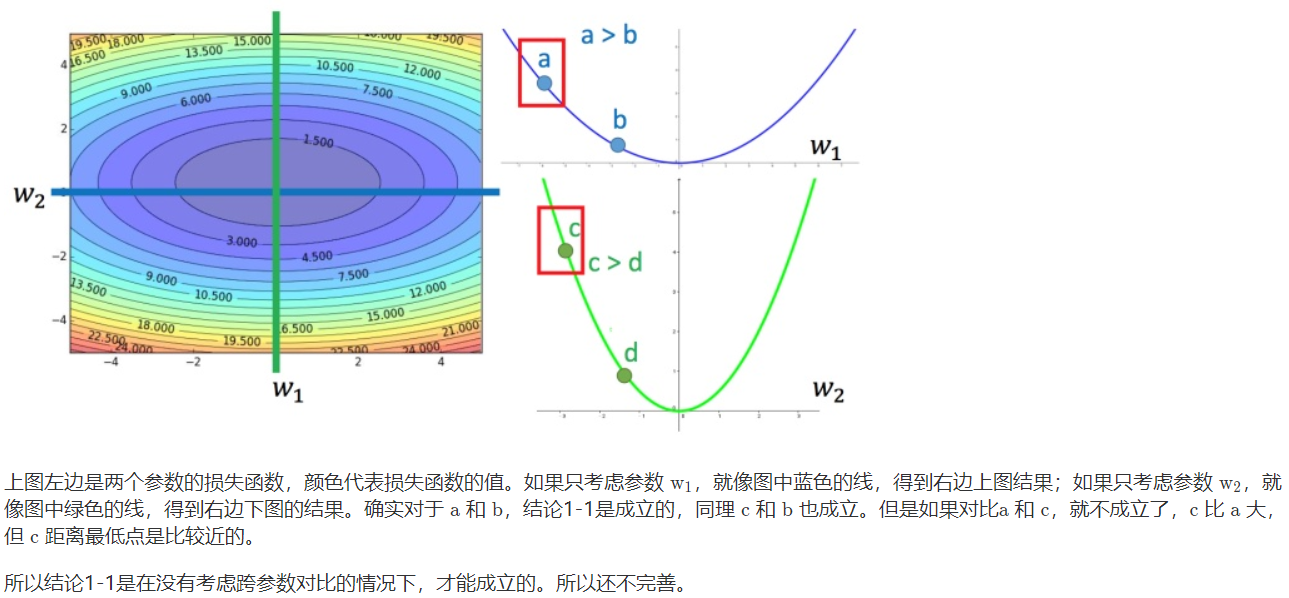
#### 1.3.2解释

##### 1.3.2.1 只看分子的一阶梯度



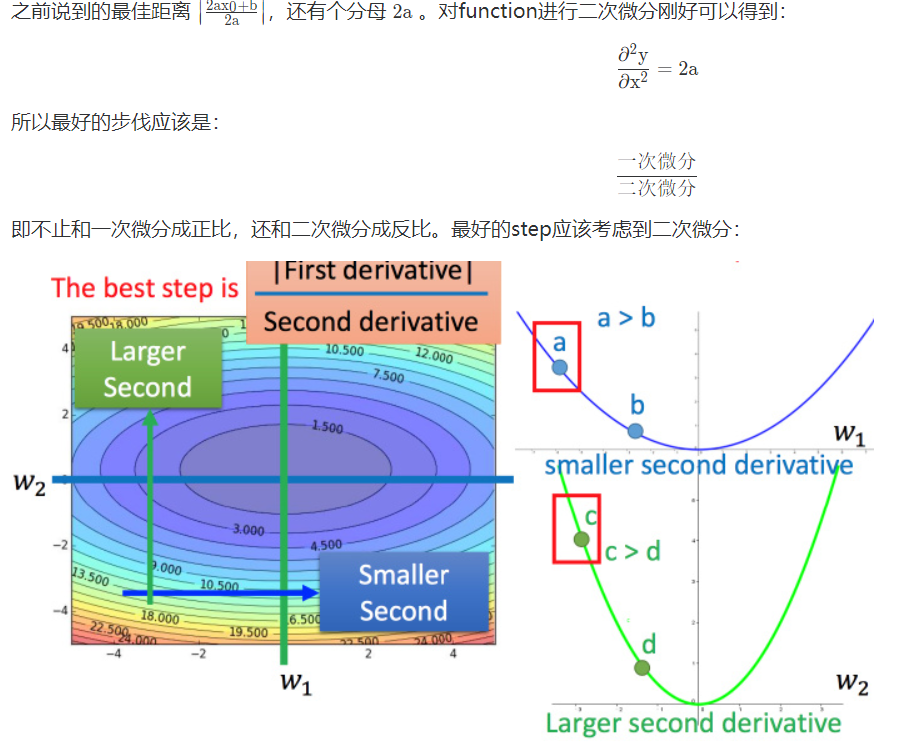
##### 1.3.2.2 只看分子一阶梯度存在的问题

虽然分子的一阶梯度与最佳步长是正相关的，但是这个结局仅能针对单一参数进行比较，不同的参数之间，一阶梯度和最佳步长并不能保证这样的关系



也就是上图的a的一阶梯度比c小，但是事实上a距离最优点更远

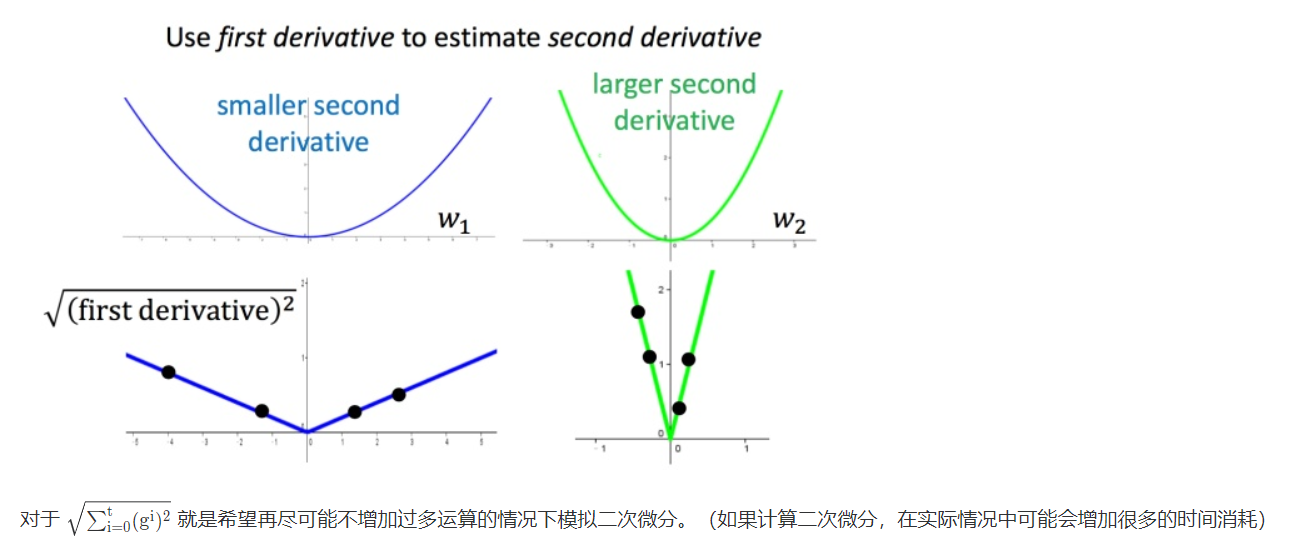
##### 1.3.2.3 再看二阶梯度



结论就是步长应该和一次微分成正比，还和二次微分成反比

##### 1.3.2.4 但是怎么计算二次微分呢，adagrad提供的方式是用过去所有一次微分平方和代替二级微分

为什么可以这么做呢？



从下面的图中可以看出，相同的采样数目，如果一阶微分的平方和比较小，则其二阶微分也是比较小的

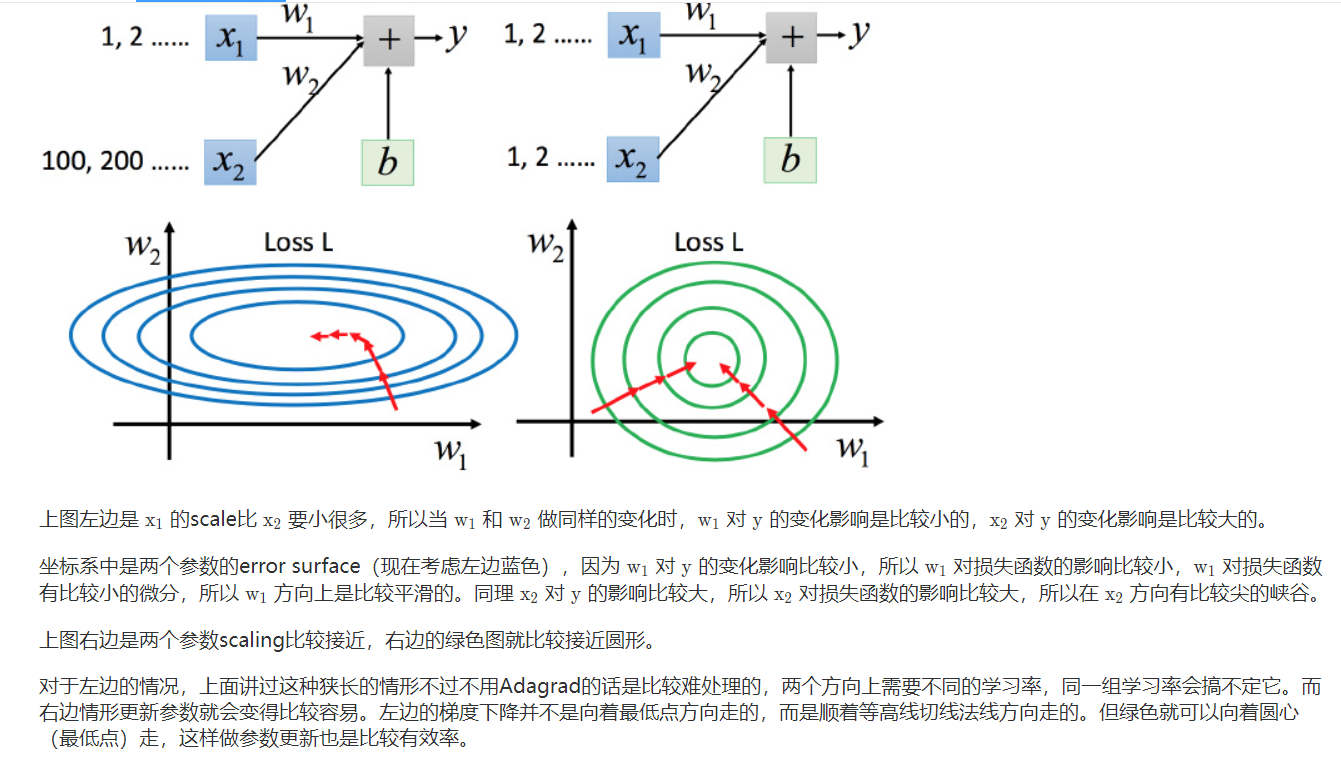
## 2.随机梯度下降SGD

因为一个sample就计算一次loss就更新一次参数，因此计算更快，但是容易振荡不够稳定，而计算所有的sample组成的loss再更新参数，比较慢，但是向最佳点的收敛很稳定

这种方案是mini-batch的GD

## 3.特征缩放（归一化）

### 3.1 为什么需要进行特征缩放？



因为梯度下降总是沿着等高线下降的方向进行，但是在没有归一化的时候，等高线下降的方向不能指向最优点的方向，就需要很多次的更新以及比较合适的自适应学习率才能够保证优化得到最佳点

但是归一化之后，以二维特征为例，等高线是圆，那么等高线的下降方向就是向着最低点的，因此能够很快的收敛

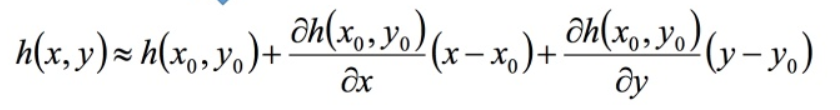
### 3.2 特征缩放的方法

Min-max

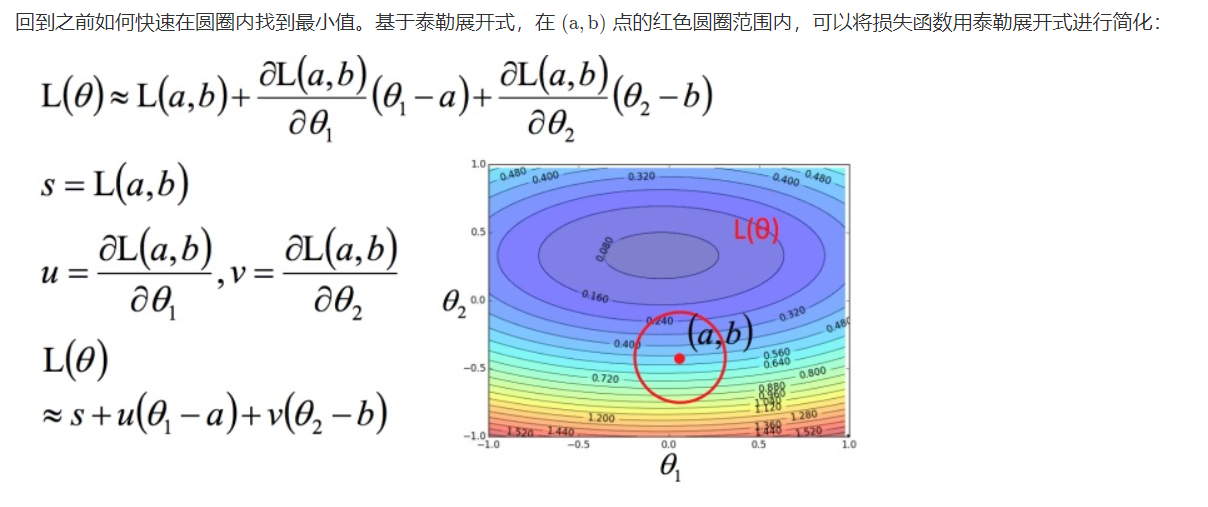
z-score

## 4.梯度下降的数学理论

### 4.1 多元泰勒展开式

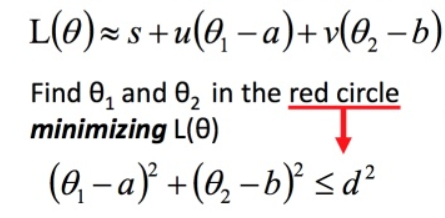


### 4.2 从泰勒展开式出发，推导梯度下降的基本原理

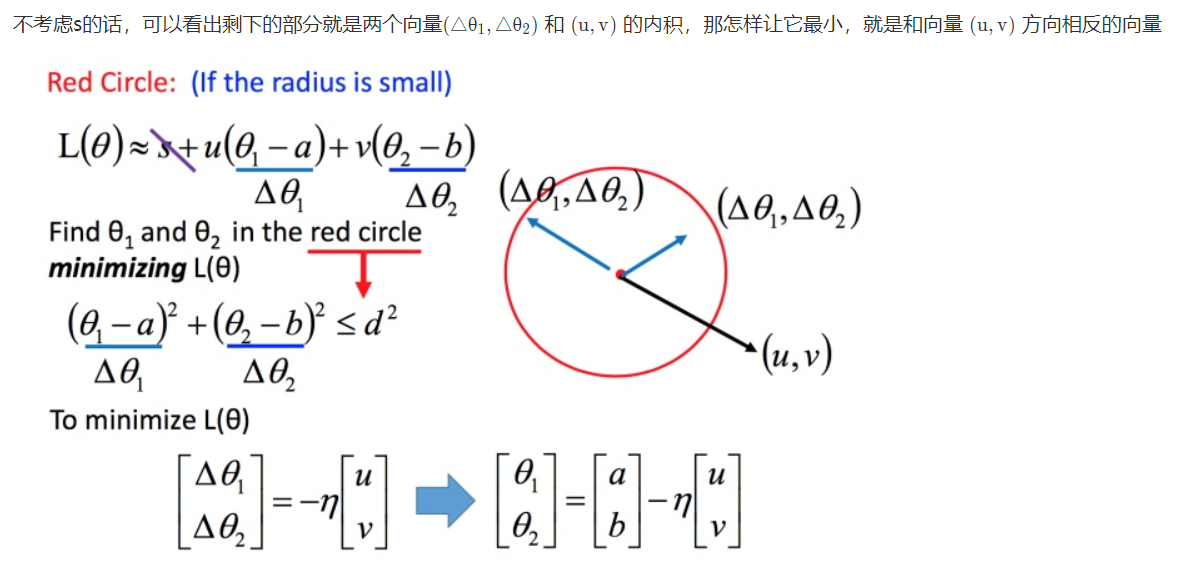


初始点就是(a,b)，然后在步长（η）范围内（红色圆圈）内向着最优点靠近

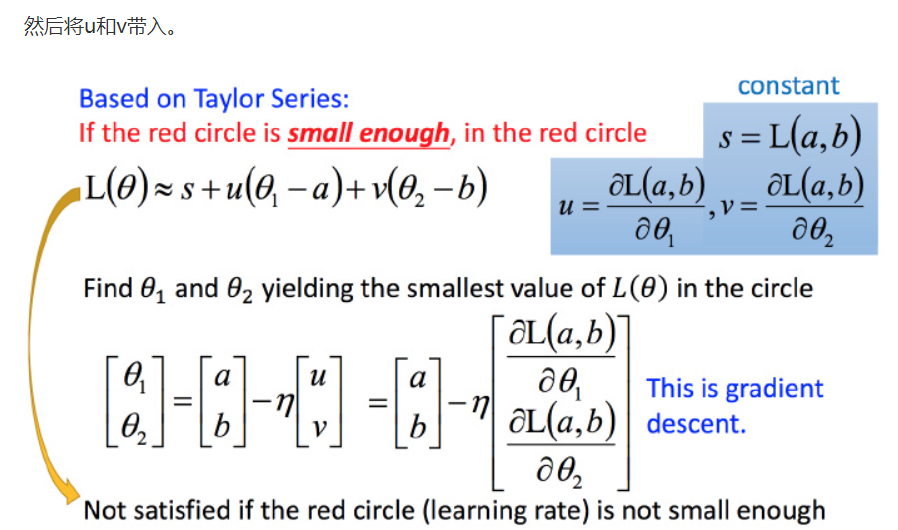
那么问题就转换成：



上面的优化问题，很容易求解：



那么：



但是要求红圈足够小，才能实现泰勒近似，因此就需要对学习率进行调整，特别是adagrad

## 5梯度下降的局限性

容易陷入局部极值

还有可能卡在不是极值，但微分值是0的地方，称为鞍点

还有可能实际中只是当微分值小于某一个数值就停下来了，但这里只是比较平缓，并不是极值点，也就是早停可能就停止了