# 4.凸优化

## 4.1.无约束优化

### 4.1.1.无约束优化问题

无约束优化问题是机器学习中最普遍、最简单的优化问题。



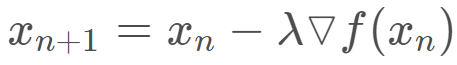
求最大值也可以 在前面加上负号，变成上面求最小的形式。

求一个函数f(x)的最小值可以对函数f(x)求导并使其等于0(或者说使得梯度▽f(x)等于0)，但是很多复杂的函数求导后没法求出解，所以这种方法实际上很少用。

常用梯度下降法、牛顿法或者拟牛顿法求解。

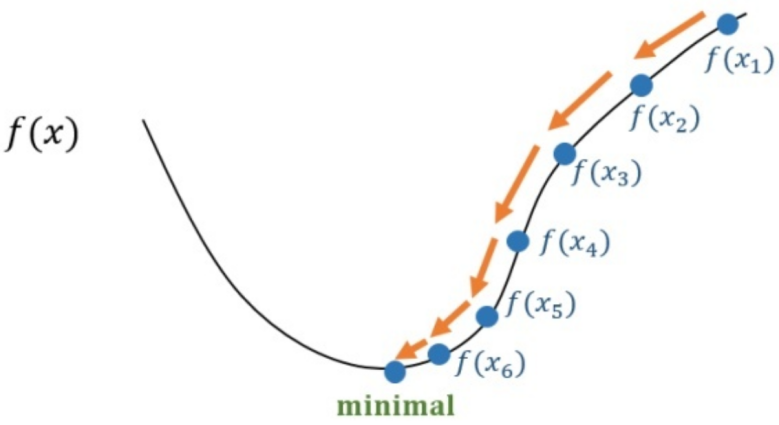
### 4.1.2.梯度下降法

基于迭代的方法，从某个点开始找很多点，使得这些点满足：，且有，这里表示单位梯度，经常写作，λ表示步长，所以通项是：



实际上λ也不会取很大，一般是

其过程为：



梯度下降法的种类：

①批量梯度下降法（BGD）

更新系数时，所有样本都参与计算

优点：需要个很少的迭代次数就可以收敛

缺点：当样本量很大时，更新一次的时间很长

②随机梯度下降法（SGD）

更新系数时，从n个样本中随机选择一个样本参与计算，

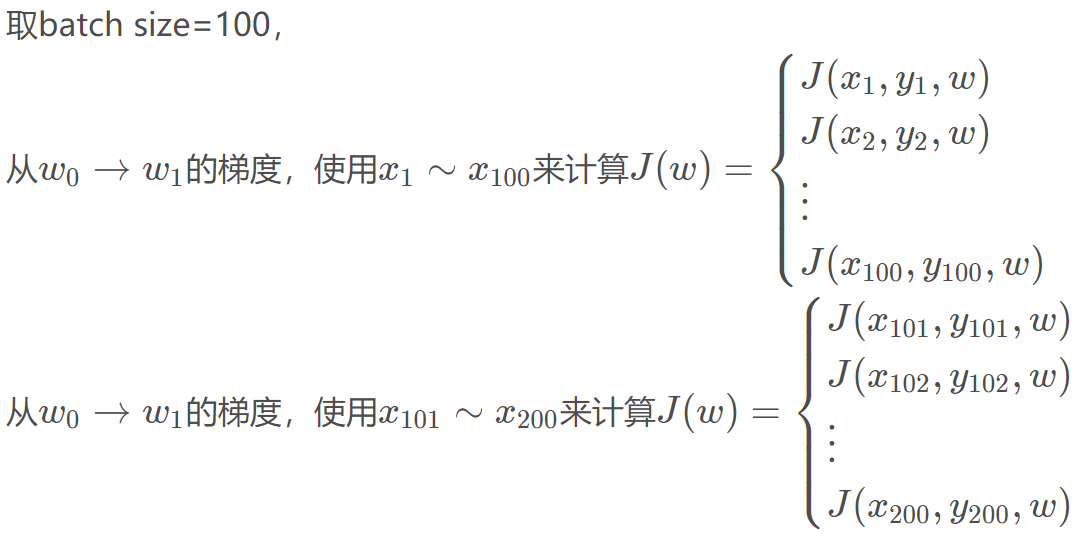
优点：更新一次的时间很短，所以大样本时有优势

缺点：会受到每一个样本的影响会很大，不稳定，需要更多的迭代次数才能收敛

③小批量梯度下降法（MBGD）

结合了批量梯度下降法和随机梯度下降法，选择一小部分样本参与计算

例如：

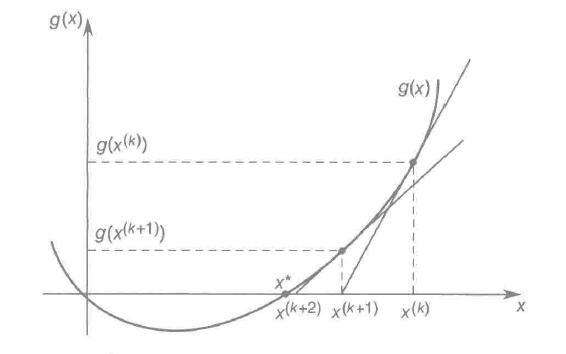


所有的样本都算完，就是一个epoch

### 4.1.3.牛顿法

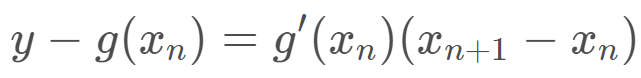
求一个函数的最小值可以对函数求导并使其等于0(或者说使得梯度等于0)：，把函数的导数看做一个函数，令

牛顿法求的过程也是迭代过程

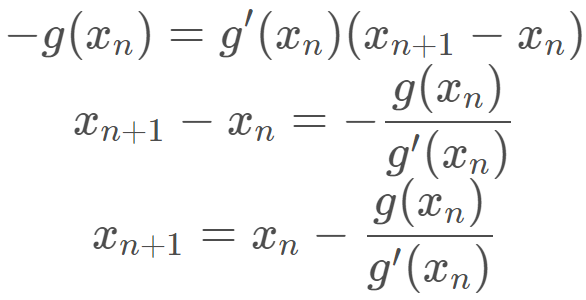


假设的函数曲线是这个样子，要找到那个的点，先做某个的切线，然后找到切线与x轴相交的点然后再做的切线，以此类推，不断逼近的点。

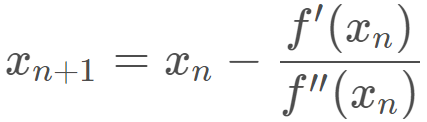
先来求第一条切线的方程：



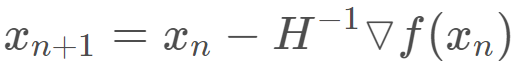
令y=0（就是上图中的点）得：



再把带入得：



这是二维的情况，如果是多维的情况：



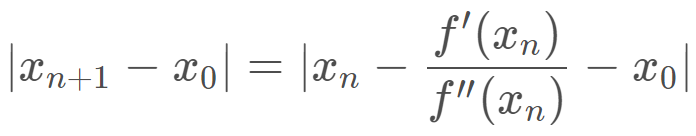
其中H是海森矩阵，除以海森矩阵就是乘以它的逆矩阵。

为什么这里是海森矩阵？因为是的n维向量，是n维向量，二次求导就是海森矩阵。

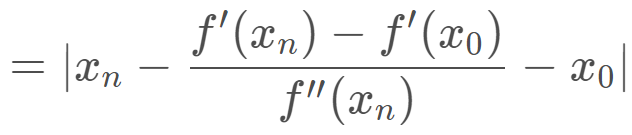
在机器学习中，要算海森矩阵的逆矩阵很麻烦，于是就引申出了很多种拟牛顿法BFGS（用另外一个矩阵来逼近海森矩阵的逆矩阵）。

**牛顿法收敛速度：**

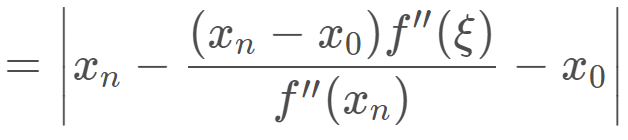
按这个迭代原理，就应该是函数的局部最优点，也就是有最小值，且有要弄明白这个收敛速度，就是要比较下到的距离和到的距离的区别，由上述结论得：

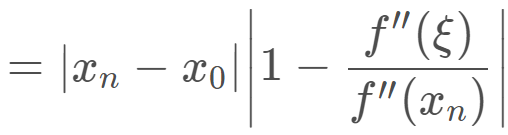


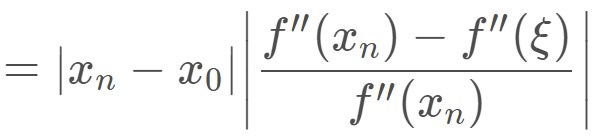
由于，所以分子加上得：



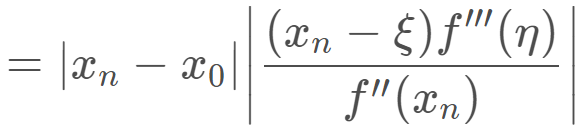
根据中值定理f(b)−f(a)=(b−a)f′(ξ),a<ξ<b，得：

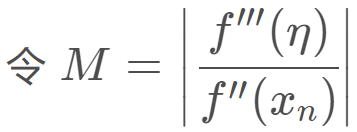






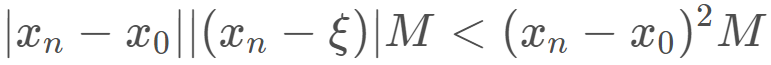
再利用拉格朗日中值定理得：







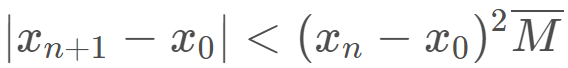
ξ是在之间的，所以



由于M的分子分母都是导数，导数都是有界的，所以M是有界的，用表示其上界。



即：



当和的距离小于1：，则，这里是按照平方的速度进行收敛的，收敛速度更快，注意这里有条件：x和的距离小于1，如果距离大于1，上界会越来越大，没法收敛。

综上，牛顿法要拟合，不能离最小值太远的地方拟合，越接近极小值再拟合收敛的效果越好。因此可以先用梯度下降，到了局部极小值附近后再用牛顿法。

## 4.2.带约束优化

### 4.2.1.等式约束-拉格朗日乘子

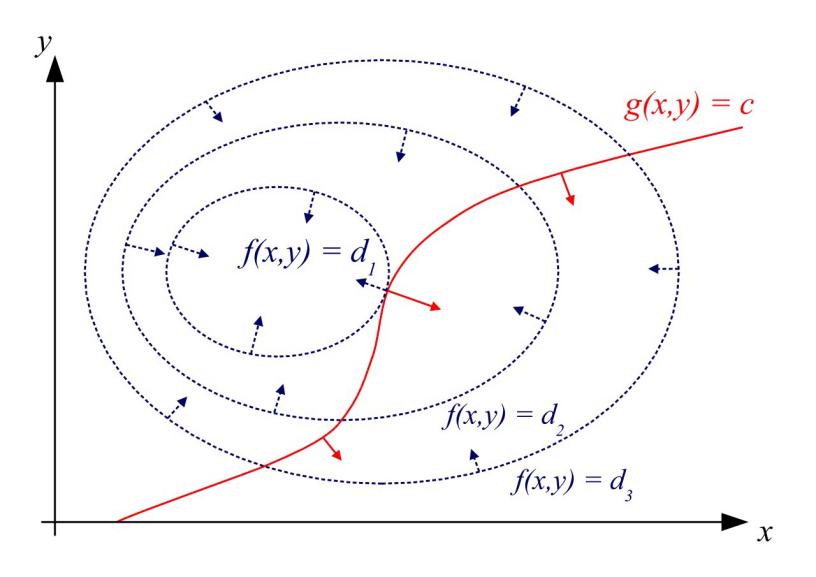
经典拉格朗日乘子法是下面的优化问题（注：x是一个向量）



等式约束：



以二维函数f(x,y)为例，这里画出z=f(x,y)的等高线越往中心，越接近最小值（越往中心，越接近最小值）但是又有等式约束g(x)=0（下图的约束曲线是g(x,y)=c,当然也可以写成g(x,y)−c=0）



也就是说这个时候的极值点应该在曲线g(x,y)上，又要尽量接近等高线中心，所以等高线和曲线应该是相切的，就是切线方向是共线的，也就是法线方向也是共线的，（法线和切线是垂直的）。

那么等高线的法线怎么求？实际上就是某点的梯度。

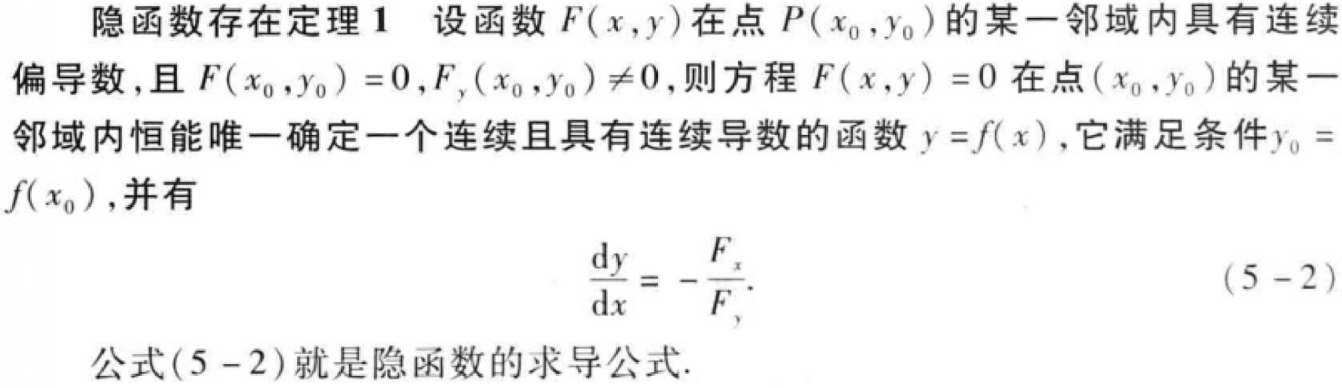
证明如下：

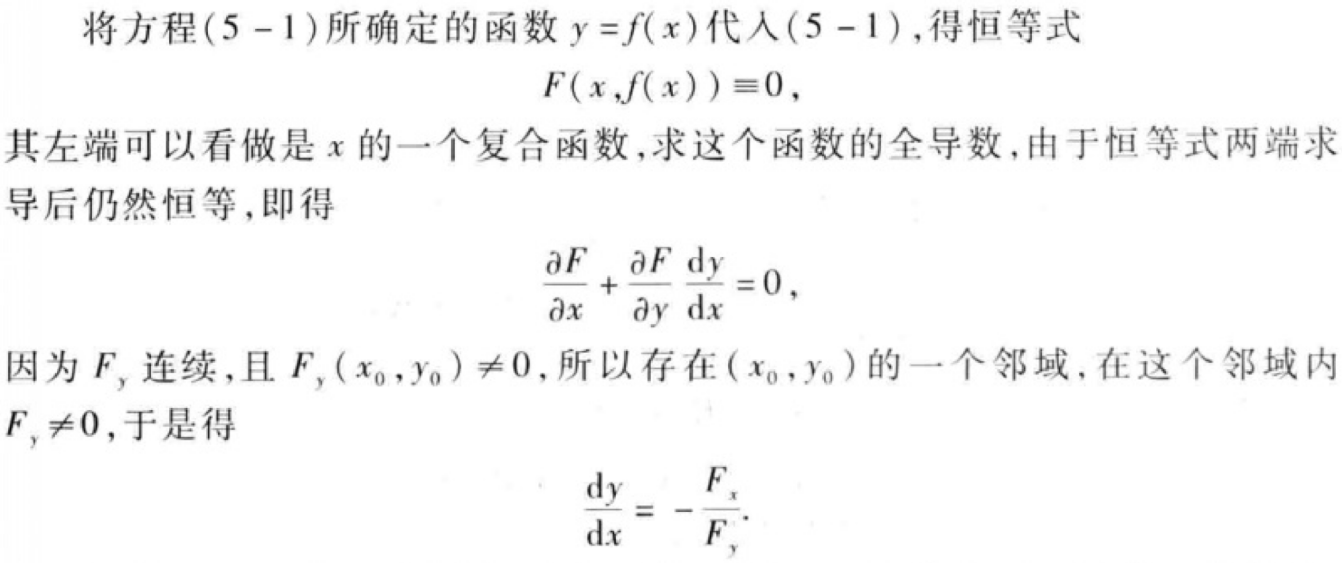
假如二维函数是z=f(x,y)，那么某个等高线也就是z为某个常数c，可以写为：f(x,y)=c

曲线的切线方程为

由隐函数求导公式得

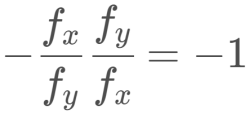
隐函数定理及求导公式如下





由于梯度可以用各个偏导排列组成的向量来表示：，其方向为：

用切线乘以梯度方向：



也就意味着梯度和切线是垂直关系。也就是说梯度方向和法线方向平行。

所以就有：



二维函数与约束函数的梯度要平行。（平行关系所以λ是正是负都可以）

由上述条件得：



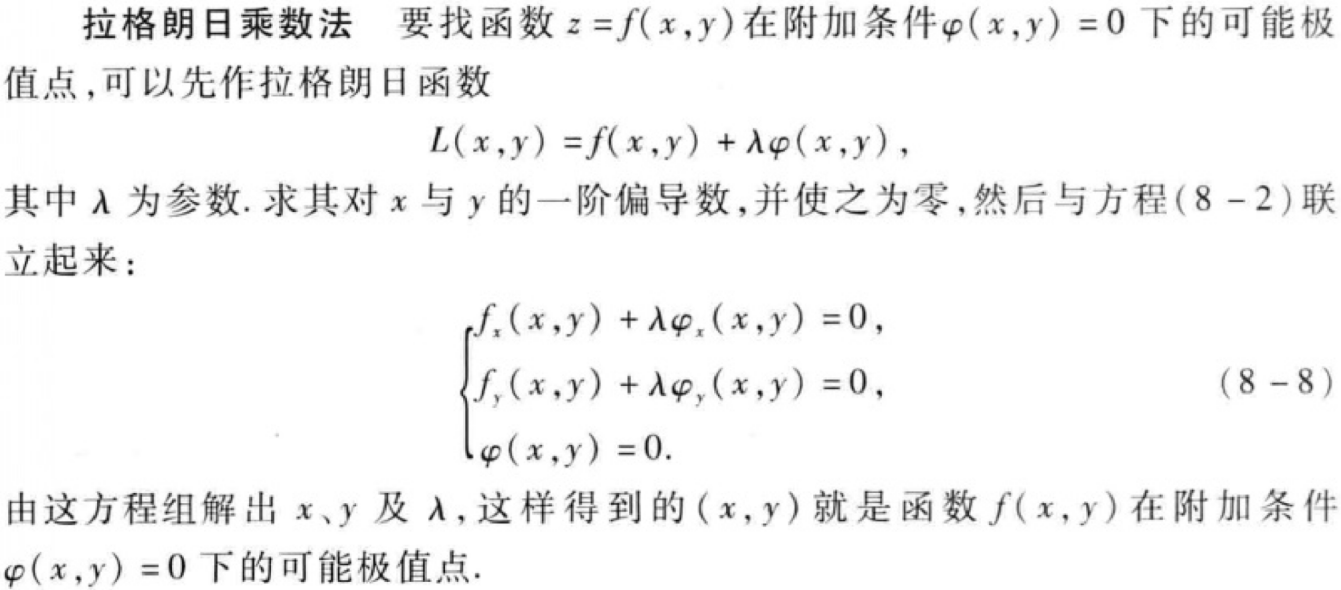
若引进辅助函数



可得

，

**函数称为拉格朗日函数，参数λ称为拉格朗日乘子**

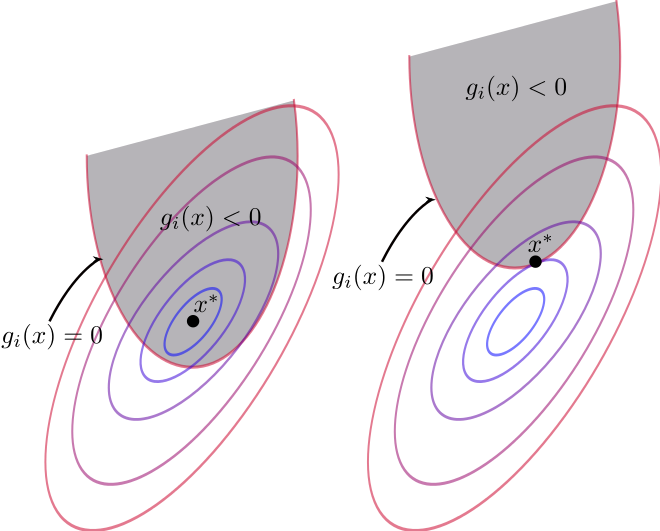


### 4.2.2.不等式约束

不等式约束问题：



仍然考虑向量x为二维空间中的点的场景，图中阴影部分为g(x)<0区域



这里依然引入拉格朗日函数：



但是这里需要分情况讨论：

（1）当最优点在的区域中，这个点其实就是目标函数自身的极值点，有没有约束条件都一样，所以这个没什么用，直接对目标函数求导并令其为0 ，就可以求出极值点。这种情况相当于在的情况下，令， 所以；

（2）当最优点在这个边界上，就需要注意的符号了。这里和等式约束不同，等式约束不关心的部分在在哪里，而像这里的不等式约束，是约束条件的要求，所以我们可以知道约束条件的梯度，也就是的方向是指向外面的，因为梯度的指向是值增长的方向，即是可行域外。而目标函数的梯度不管是等式约束还是不等式约束，都是指向等势面增大的方面，所以是指向的。所以此时，目标函数的梯度方向和约束条件的梯度方向之间是相反的，也就是说存在， 使得。这种情况相当于在的情况下，令， 所以。

那么对以上两个部分的条件可以转化为以下函数，也就是经常见到的 KKT 条件：



通过以上KKT条件，就可以对该不等式约束问题进行优化求解了。

KKT 只是取极值的必要条件，不是充分条件。

**等式与不等式约束混合的极值问题：**

对于以下的混合优化问题：



s.t. 



那么可以构造拉格朗日函数：

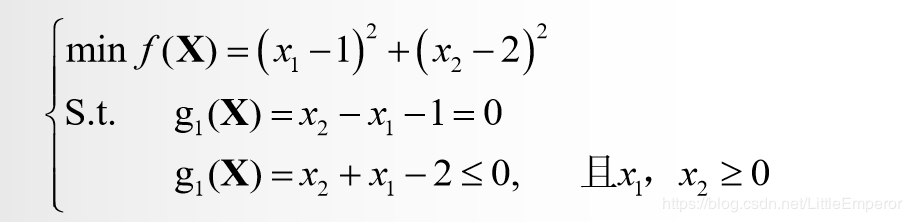


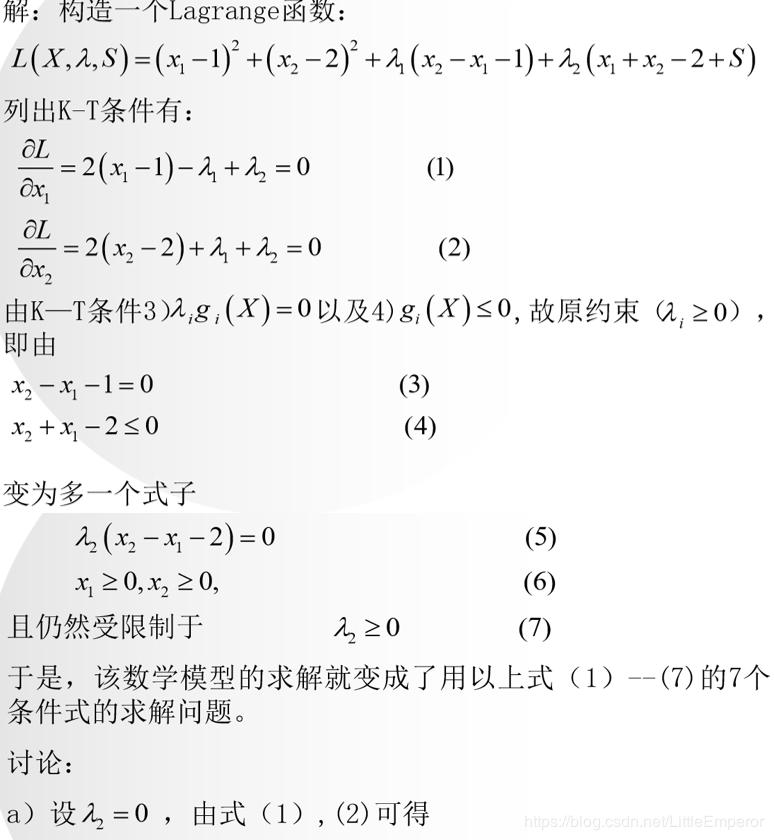
KKT 条件：

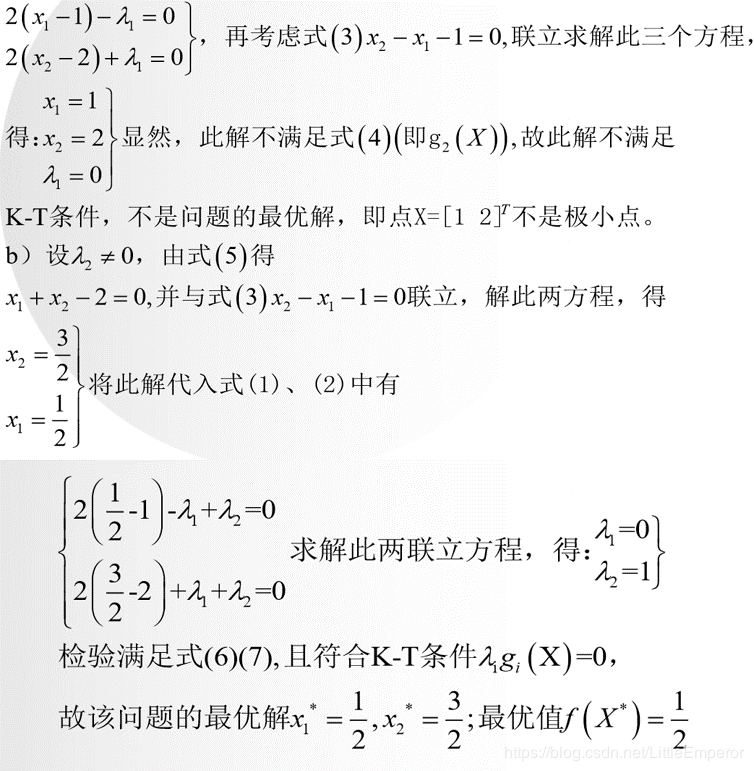


通过以上KKT条件，就可以对该不等式约束问题进行优化求解了。

例：求解如下非线性规划问题



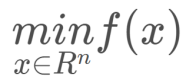


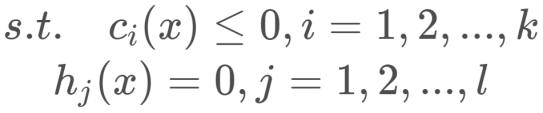


### 4.2.3.优化的对偶理论

**1.原始问题**

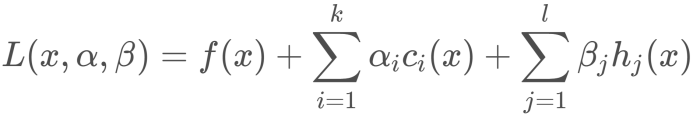
假设是定义在上的连续可微函数。考虑约束最优化问题





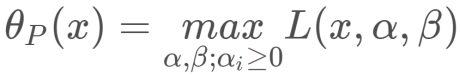
称此约束最优化问题为原始最优化问题或原始问题.

首先，引进广义拉格朗日函数





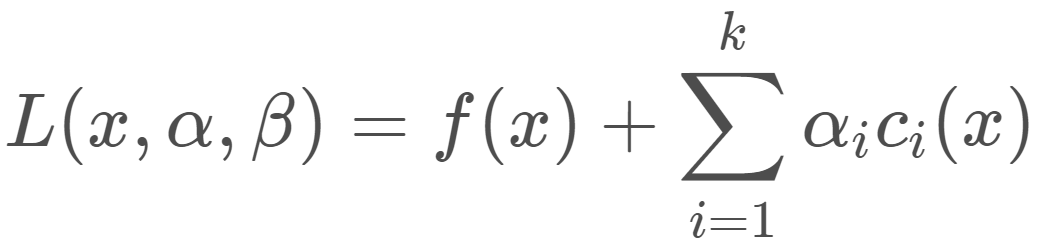
引入如下函数：



计算此函数，分情况讨论

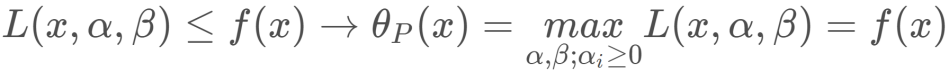
**情况一：**

当x满足原始问题约束，首先，所以有：





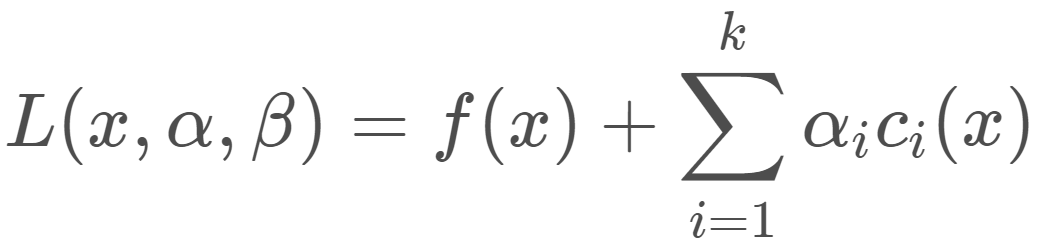
因此函数的最大值就是当的时候，所以有：



**情况二：**

x不满足原始问题约束，又分两种情况讨论：

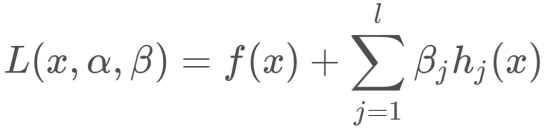
（1）当时,得：





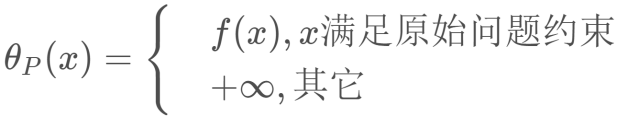
所以后面项的最大值为，即

（2）当时，取得：

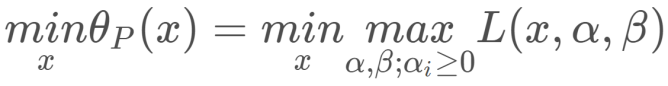


后面那项的最大值也是，即

综合情况一和情况二就可以把写为：



现在考虑极小值，由于是分段函数，第二部分是不会有最小值的，所以整个的最小值肯定是在第一部分，写成：

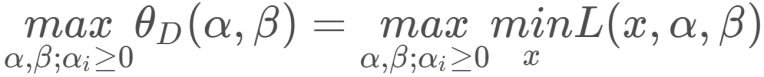


可以看到，如果要满足上述公式，那么x就是既要满足情况一中的原始问题约束，又要使得f(x)最小，这个就是和原始问题等价的，也就是把原始问题转换成了上述的形式。这个形式没有约束的存在了。

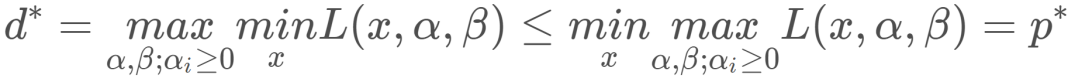
**2.对偶问题**

之前的一般问题是先极小值再极大值，对偶问题是先极小值再极大值。下面直接给定义。  
 

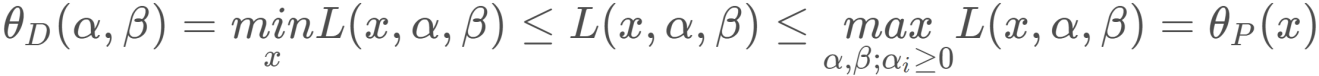




**弱对偶定理：**若原始问题和对偶问题都有最优值，则



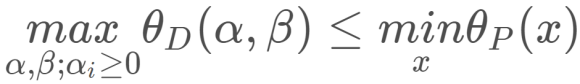
证明：对任意的α，β和x，有



即：



由于原始问题和对偶问题均有最优值，所以：



以上这种形式是弱对偶形式，如果能把不等号变成等号，那么就变成强对偶形式。

**强对偶定理：**考虑原始问题和对偶问题。假设函数和是凸函数，是仿射函数（最高次数为1的多项式函数）；并且假设不等式约束是严格可行的，即存在，对所有有，则存在使是原始问题的解，是对偶问题的解，并且



KTT定理：对原始问题和对偶问题，假设函数和是凸函数，是仿射函数，并且不等式约束是严格可行的，则分别是原始问题和对偶问题的解的充分必要条件是满足下面的Karush-Kuhn-Tucker（KKT）条件：

