4. 凸优化

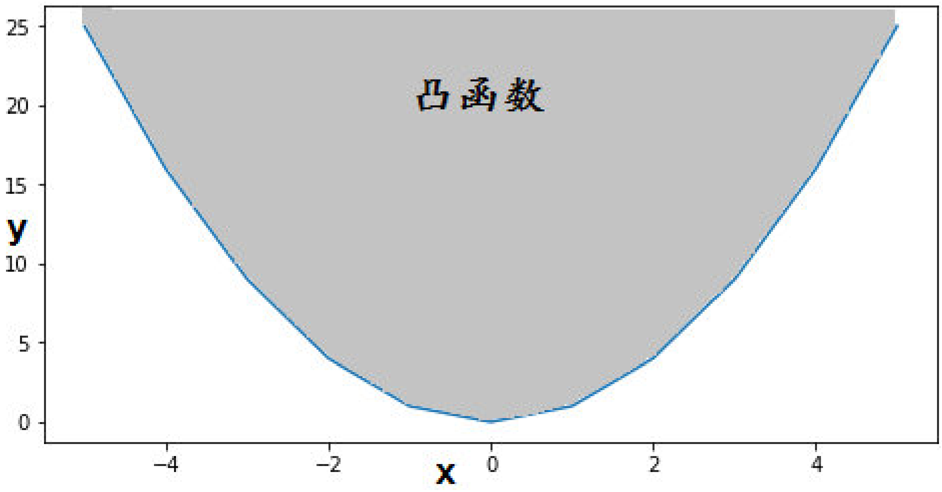
4.1 定义

4.1.1 凸函数

若$f(x)$是凸函数，（如$y=x^2$）则函数图像位于凸函数上方的区域构成凸集。

有：

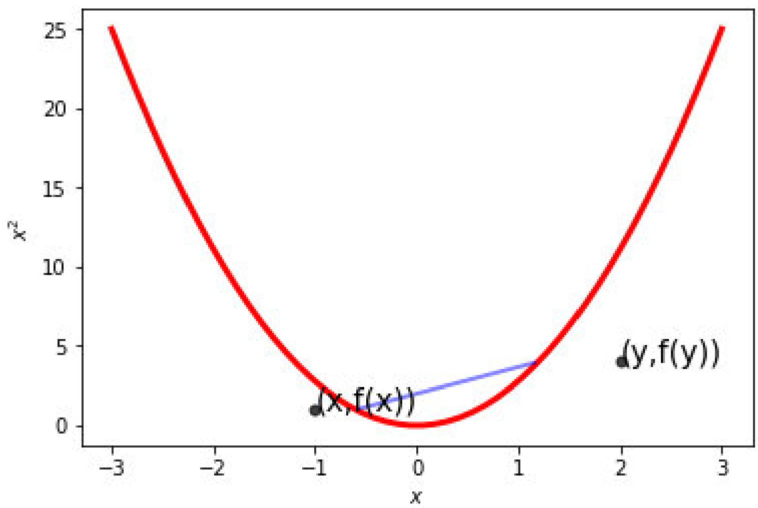
- 凸函数图像的上方区域，一定是凸集。
- 一个函数图像的上方区域是凸集，则该函数就是凸函数。
- 凸函数的局部最小值就是全局最小值。



正式定义：

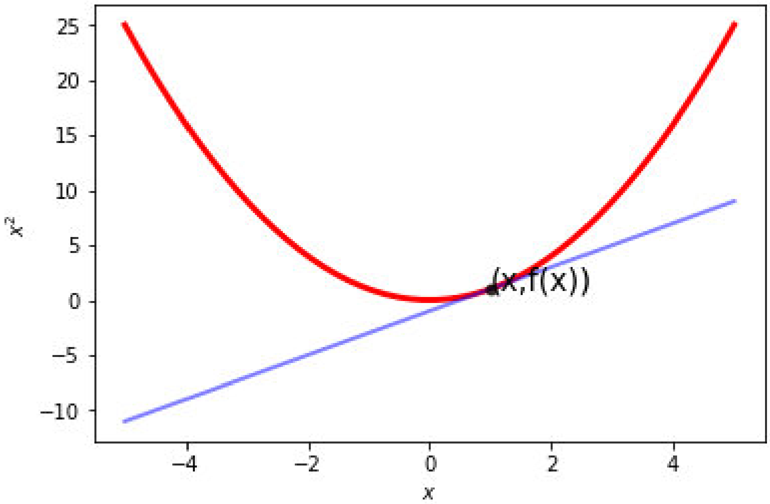
（1）若函数 f 的定义域 $dom f$ 为凸集，且满足

$$\forall x_1, x_2 \in \operatorname{dom} f, 0 \leq \theta \leq 1, \text{有 } f(\theta x+(1-\theta) x) \leq \theta f(x)+(1-\theta) f(x)$$



（2）若函数 f 一阶可微，则函数 f 为凸函数当且仅当 f 的定义域domf为凸集，且：

$$\forall x_1, x_2 \in \operatorname{dom} f, f\left(x_2\right) \geq f\left(x_1\right)+\nabla f\left(x_1\right)^T\left(x_2-x_1\right)$$



（3）若函数 f 二阶可微，则函数 f 为凸函数当且仅当 f 的定义域domf为凸集，且：

$$\nabla^2 f(x) \geq 0$$

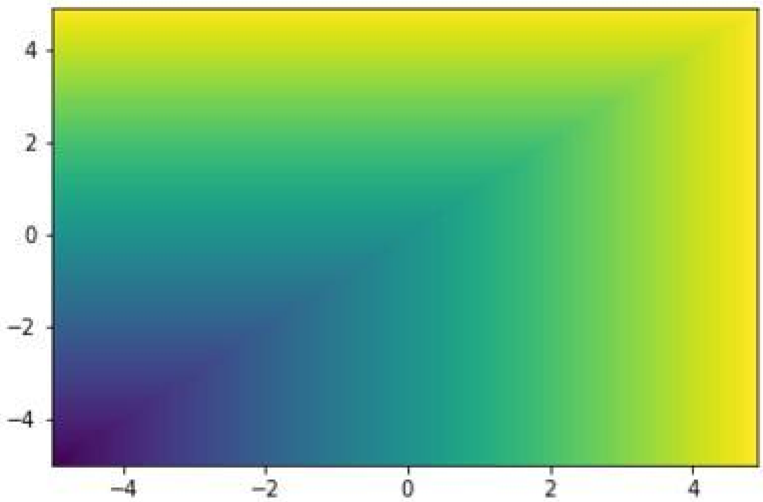
若 f 是一元函数，上式表示二阶导大于等于零。

若 f 是多元函数，上式表示二阶导 $Hessian$ 矩阵半正定。

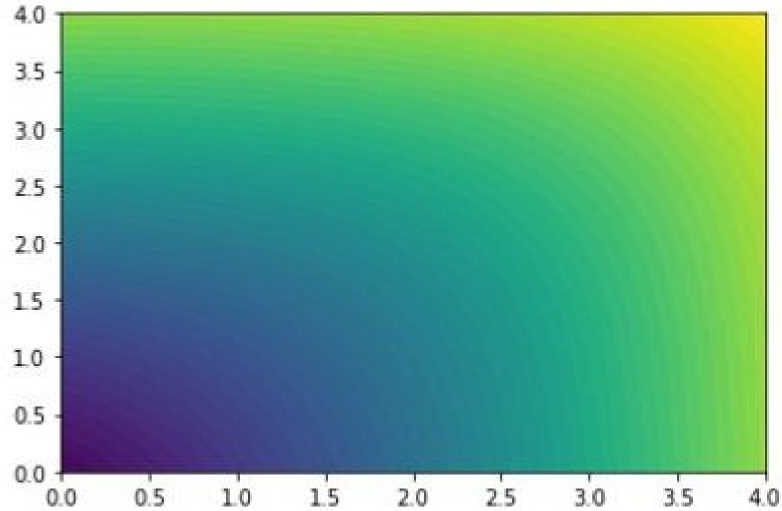
凸函数的举例：

- 指数函数：$f(x) = e^{ax}$
- 幂函数：$f(x) = x^a, x \in R^+, a \geq 1$ 或 $a \leq 0$
- 负对数函数：$f(X) = -lnx$
- 负熵函数：$f(X) = xlnx$
- 范数函数：$f(\vec{x}) = \|x\|$
- 最大值函数：$f(\vec{x}) = max(x_1, x_2, \ldots, x_n)$
- 指数线性函数：$f(\vec{x}) = \log(e^{x_1} + e^{x_2} + \ldots + e^{x_n})$

  最大值函数：



指数线性函数：



4.1.2 仿射集

若通过集合 C 中任意两个不同点的直线仍然在集合 C 内，则称集合 C 为仿射集。

$$\forall x_1, x_2 \in C, \forall \theta \in R\text{，则 } x = \theta \cdot x_1 + (1-\theta) \cdot x_2 \in C$$

直线，平面和超平面都属于仿射集：n 维空间的 n -1维仿射集为 n -1维超平面。

4.1.3 凸集

集合 C 内任意两点间的线段均在集合 C 内，则称集合 C 为凸集：

改：$\forall x_1, x_2 \in C, \forall \theta \in [0,1]$，则 $\theta x_1 + (1-\theta) x_2 \in C$

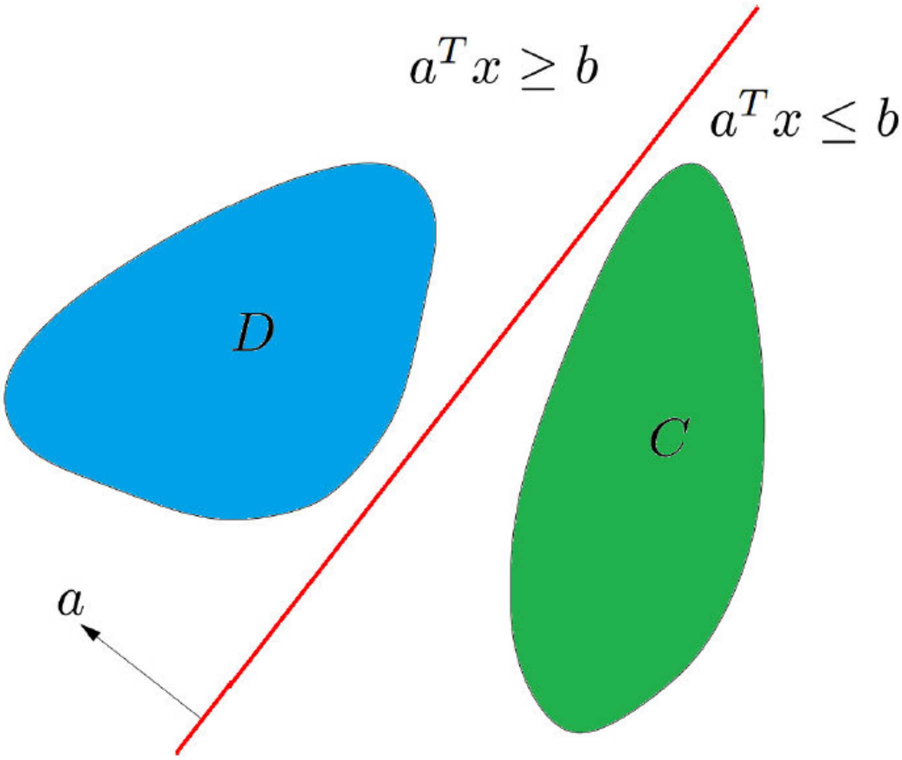
$$\forall x_1, \ldots, x_k \in C, \theta_i \in [0,1], \sum_{i=1}^{k} \theta_i = 1\text{，则 } x = \sum_{i=1}^{k} \theta_i x_i \in C$$

一般来说，仿射集的要求更高，仿射集必然是凸集，凸集未必是仿射集。

4.1.4 分割超平面

设 C 和 D 是两个不相交的凸集，则存在超平面 P，P 可以将 C 和 D 分离。

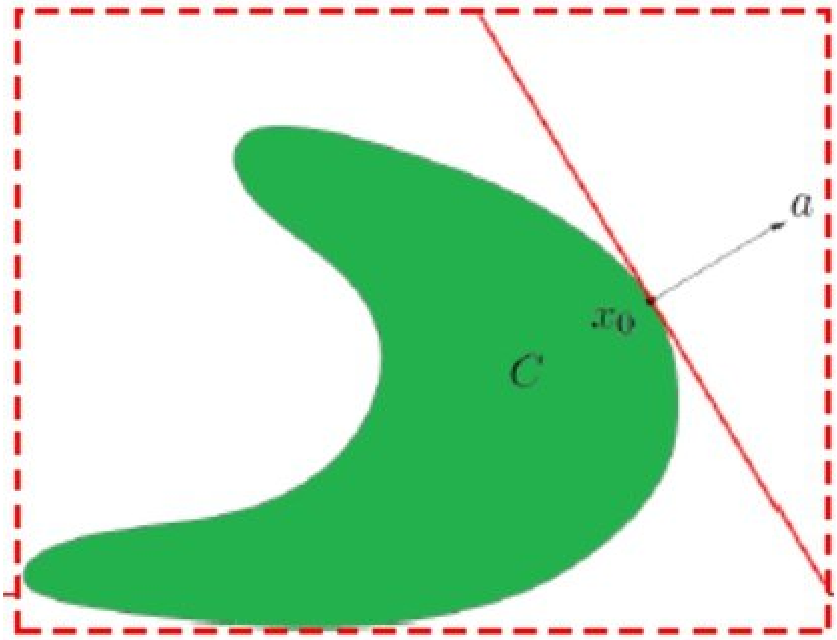
有：$\forall x \in C, a^T x \leq b$，且 $\forall x \in D, a^T x \geq b$



4.1.5 支撑超平面

设集合 C ， X_0 为 C 边界上的点。若存在 $a \neq 0$ ，满足对任意 $x \in C$ ，都有 $a^Tx \le a^Tx_0$ 成立，则称超平面 $\{x|a^Tx = a^Tx_0\}$ 为集合 C 在点 X_0 处的支撑超平面。

凸集边界上任意一点，均存在支撑超平面。反之，若一个闭的非中空（内部点不为空）集合，在边界上的任意一点存在支撑超平面，则该集合为凸集。



4.2 Jensen不等式

$Jensen$ 不等式相当于把凸函数的概念反过来说,即是如果 f 是一个凸函数,任意取一个在 f 定义域上的(x,y) 点，$\theta \in [0,1]$

若 f 是凸函数，有：

$$\forall x_1, x_2 \in \operatorname{dom} f, 0 \le \theta \le 1,\ 有\ f(\theta x_1 + (1-\theta)x_2) \le \theta f(x_1) + (1-\theta) f(x_2)$$

若 $\theta_1, \theta_2, \ldots \theta_k \geq 0, \theta_1+\theta_2+\ldots+\theta_k=1$

则 $f\left(\theta_1 x_1+\theta_2 x_2+\ldots+\theta_k x_k\right) \leq \theta_1 f\left(x_1\right)+\theta_2 f\left(x_2\right)+\ldots+\theta_k f\left(x_k\right)$

若 $p(x) \geq 0 \ \ on \ \ S \subseteq \operatorname{dom} f, \int_S p(x) d x=1$

则 $f\left(\int_S p(x) x d x\right) \leq f\left(\int_S p(x) f(x) d x \rightarrow f(E(x)) \leq E(f(x))\right.$

- $Jensen$ 不等式是几乎所有不等式的基础

1. 利用 $y=-logx$ 是凸函数，证明：$\frac{a+b}{2} \geq \sqrt{ab}, \quad a>0, b>0$ 。

   提示：任取 $a, b>0, \theta=0.5$ 代入基本 $Jensen$ 不等式。

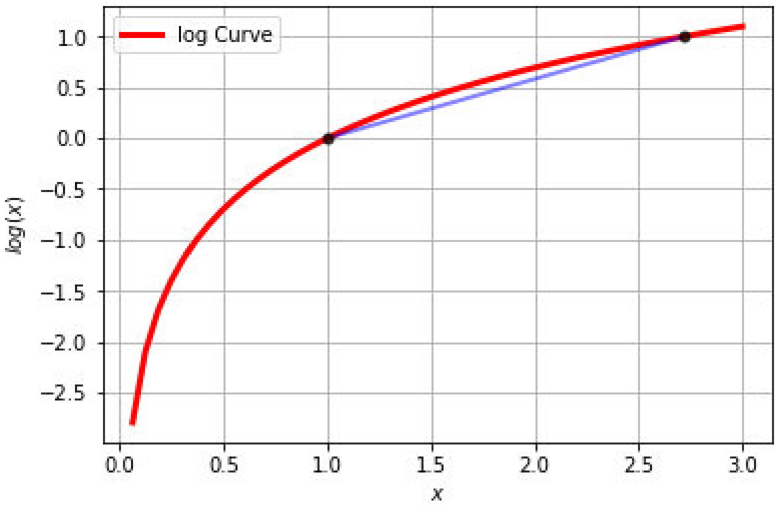
2. 利用 $f(E(x)) \leq E(f(x))$, (f 是凸函数)，证明下式 $D \geq 0$ 。其中，

   $D(p \| q)=\sum_x p(x) \log \frac{p(x)}{q(x)}=E_{p(x)} \log \frac{p(x)}{q(x)}$

   证明：

   注意到 $y=-logx$ 在定义域上是凸函数，则：

$$
\begin{aligned}
& D(p \| q) \\
= & \sum_x p(x) \log \frac{p(x)}{q(x)} \\
= & -\sum_x p(x)\left(\log \frac{q(x)}{p(x)}\right) \\
\geq & -\log \sum_x\left(p(x) \cdot \frac{q(x)}{p(x)}\right) \\
= & -\log \sum_x q(x) \\
= & -\log 1 \\
= & 0
\end{aligned}
$$



4.3 保凸性算子

（1） 保持函数凸性的算子

凸函数的非负加权和：$f(x)=\omega_1 f_1(x)+\omega_1 f_2(x)+\cdots+\omega_n f_n(x)$ 。

凸函数与仿射函数的复合：$g(x)=f(A x+b)$ 。

凸函数的逐点最大值、逐点上确界：

$$f(x)=\max \left(f_1(x), \cdots, f_n(x)\right)$$

$$f(x)=\sup _{y \in A} g(x, y)$$

（2）凸函数的逐点最大值

若 f_1，f_2 均为凸函数，定义函数 f : $f(x)=\max\{f_1(x), f_2(x)\}$ ，则函数 f 为凸函数。

证明：

$f(\theta \cdot x+(1-\theta) \cdot y)$

$=\max\{f_1(\theta \cdot x+(1-\theta) \cdot y), f_2(\theta \cdot x+(1-\theta) \cdot y)\}$

$\leq \max\{\theta \cdot f_1(x)+(1-\theta) \cdot f_1(y), \theta \cdot f_2(x)+(1-\theta) \cdot f_2(y)\}$

$\leq \theta \cdot \max\{f_1(x), f_2(x)\}+(1-\theta) \cdot \max\{f_1(y), f_2(y)\}$

$=\theta \cdot f(x)+(1-\theta) \cdot f(y)$

备注：上述证明过程中第二个不等号的证明如下：

$f_1(x) \leq \max\{f_1(x), f_2(x)\} \Rightarrow$

$\theta \cdot f_1(x) \leq \theta \cdot \max\{f_1(x), f_2(x)\} \ldots$（1）

$f_1(y) \leq \max\{f_1(y), f_2(y)\} \Rightarrow$

$(1-\theta) \cdot f(y) \leq (1-\theta) \cdot \max\{f_1(y), f_2(y)\} \cdots$（2）

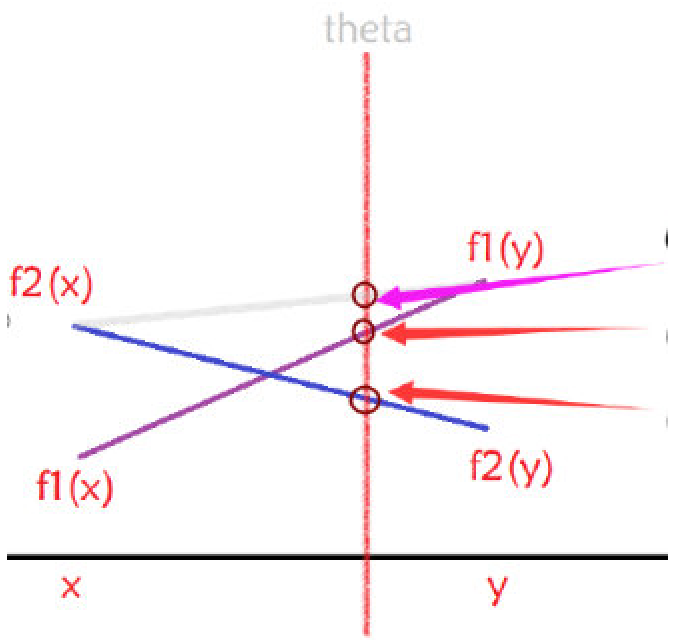
（1）＋（2）$\Rightarrow$

$\theta \cdot f_1(x)+(1-\theta) \cdot f_1(y) \leq \theta \cdot \max\{f_1(x), f_2(x)\}+(1-\theta) \cdot \max\{f_1(y), f_2(y)\}$

同理：

$\theta \cdot f_2(x)+(1-\theta) \cdot f_2(y) \leq \theta \cdot \max\{f_1(x), f_2(x)\}+(1-\theta) \cdot \max\{f_1(y), f_2(y)\}$

第二个不等式的形式化表达如下图所示。



$\theta \max\{f_1(x), f_2(x)\}+(1-\theta) \max\{f_1(y), f_2(y)\}$

$\theta f_1(x)+(1-\theta) f_1(y)$

$\theta f_2(x)+(1-\theta) f_2(y)$

（3）思考：逐点上确界和上境图的关系

一系列函数逐点上确界函数对应着这些函数上境图的交集。

直观例子：

- Oxy 平面上随意画 N 条直线（直线上是凸的——虽然直线上也是凹的），在每个 x 处取这些直线的最大的点，则构成的新函数是凸函数；
- 同时：N 条直线逐点求下界，是凸函数；

备注：在 $Lagrange$ 对偶函数中会用到该结论。

参考：https://max.book118.com/html/2016/0228/36295851.shtm

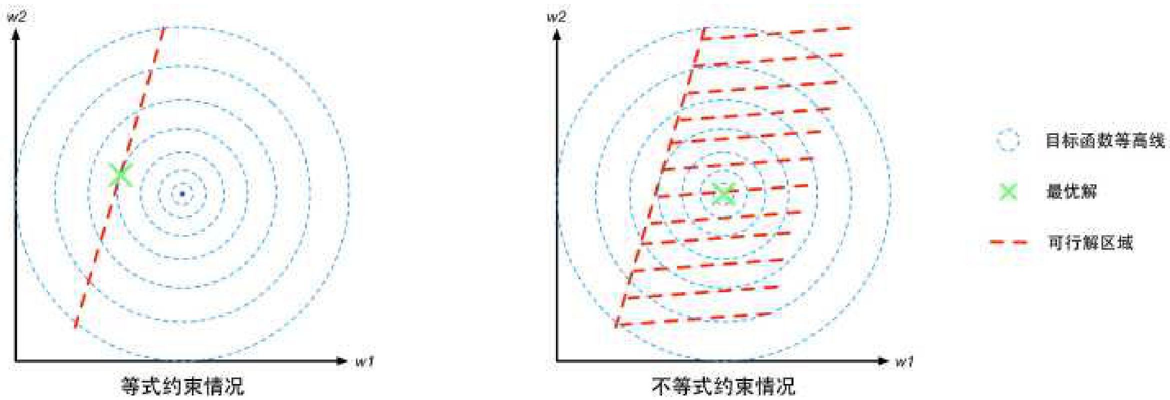
4.4 凸优化

4.4.1 约束条件下的优化问题

对于目标函数，我们限定是凸函数；对于优化变量的可行域（注意，还要包括目标函数定义域的约束），我们限定它是凸集。同时满足这两个限制条件的最优化问题称为凸优化问题，这类问题有一个非常好性质，那就是局部最优解一定是全局最优解。

约束条件一般分为等式约束和不等式约束两种，前者表示为 $g(x)=0$；后者表示为 $g(x) \le 0$。

几何图像如下：



（1） 等式约束

设目标函数为 $f(x)$，约束条件为 $h_k(x)$，形如:

$$minf(x) \quad s.t. \;\; h_k(x)=0 \;\; k=1,2,\dots,l$$

注：解决方法可以为**消元法**

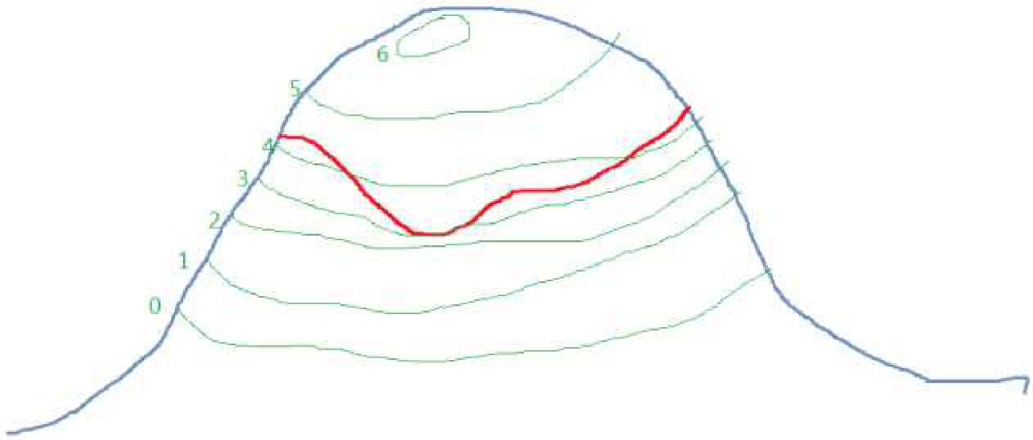
4.4.2 拉格朗日乘子法

首先定义原始目标函数 $f(x)$,拉格朗日乘子法的基本思想是把约束条件转化为新的目标函数 $L(x,\lambda)$ 的一部分，从而使得有约束优化问题变成我们习惯的无约束优化问题。问题：如何转化？

1） 最优解的特点分析（等式约束下）

等式约束下观察上左图，发现最优解恰好在可行解空间（红色虚线）和目标函数等值线（蓝色虚线）相切的地方。

想象一下目标函数 $f(x)$ 是一座山，约束 $g(x)$ 是镶嵌在山上的一条线。从最低的等高线开始往上数，满足约束条件的最低点肯定是等高线与约束条件相切的地方。



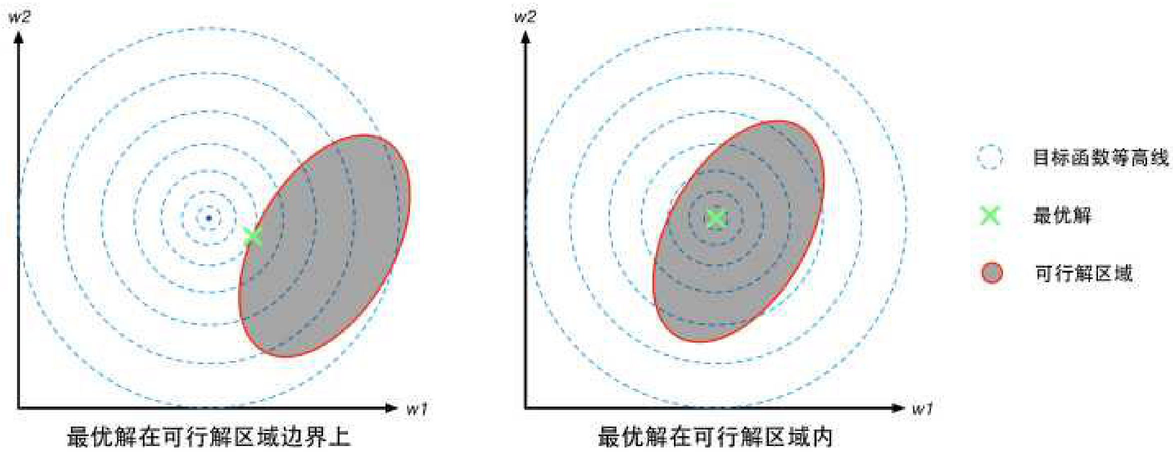
两条曲线相切，意味着他们在这点的法线平行，也就是法向量只差一个任意的常数乘子（取为 λ ）: $\nabla f(x)+\lambda \nabla g(x)=0$

因此我们定义拉格朗日函数有：

$$L(x, \lambda)=f(x)+\lambda g(x)$$

上式中对 x 求偏导即可得 $\nabla f(x)+\lambda \nabla g(x)=0$ ，而对 λ 求偏导即可的约束条件 $g(x)=0$ 。

2） 不等式约束——KKT 条件



对于不等式约束 $g(x) \leq 0$ 的情况最优解所在的位置有两种可能，或者在边界 $g(x)=0$ 上或者在可行解区域内部满足 $g(x) \leq 0$ 的地方。如果在 $g(x)=0$ 的边界上，这时约束条件起作用，并且 $\nabla f\left(x^{*}\right)$ 必定与 $\nabla g\left(x^{*}\right)$ 方向相反，可以推断 $\lambda>0$ ，如果在区域内，则相当于约束条件没有起作用，因此拉格朗日函数中的参数 $\lambda=0$ 。整合这两种情况，可以写出一个约束条件的统一表达：

$$\left\{\begin{array}{c}
g(x) \leq 0 \\
\lambda \geq 0 \\
\lambda g(\mathrm{x})=0
\end{array}\right.$$

以上公式即为 KKT 条件。

3）拉格朗日对偶

构造原始目标函数

$$\min_x f(x)$$

$$s.t.\, h_i(x)=0 \quad i=1,2,3,\ldots,m g_j(x)\leq 0 \quad j=1,2,3,\ldots,n$$

接下来构造基于拉格朗日函数的新目标函数，记为：

$$\theta_P(x)=\max_{\alpha,\beta;\beta,20} L(x,\alpha,\beta)$$

其中 $L(x,\alpha,\beta)$ 为广义拉格朗日函数，定义为

$$L(x,\alpha,\beta)=f(x)+\sum_{i=1}^{m}\alpha_i h_i(x)+\sum_{j=1}^{n}\beta_j g_j(x)$$

所以

$$\theta_p(x)=\max_{\alpha,S,\beta,\beta_2 0} L(x,\alpha,\beta)=f(x)+\max_{\alpha,\beta,\beta,z_2}\left[\sum_{i=1}^{m}\alpha h_i(x)+\sum_{j=1}^{n}\beta_j g_j(x)\right]$$

可行解区域内 $\max_{\alpha,\beta;\beta,z^0}\left[\sum_{i=1}^{m}\alpha_i h_i(x)+\sum_{j=1}^{n}\beta_j g_j(x)\right]=0$

可行解区域外可使 $\max_{\alpha,\beta;\beta,z^0}\left[\sum_{i=1}^{m}\alpha_i h_i(x)+\sum_{j=1}^{n}\beta_j g_j(x)\right]=+\infty$

所以

$$\theta_p(x)=\begin{cases} f(x),\ \boldsymbol{x}\text{在可行解区域内} \\ +\infty,\ \boldsymbol{x}\text{在可行解区域内}\end{cases}$$

接下来我们求 $\min_x \theta_P(x)=\min_x \max_{a_1,\beta_2\geqslant 0} L(x,\alpha,\beta)$，构造对偶问题

$$\max_{\alpha,\beta;\beta_j\geq 0}\theta_D(x)=\max_{\alpha,\beta;\beta_j\geq 0}\min_x L(x,\alpha,\beta)$$

问题：对偶问题何时同解？

定理1：$d^*=\max_{\alpha,\beta;\beta_j\geq 0}\min_x L(x,\alpha,\beta)\leq \min_x \max_{\alpha,\beta;\beta_j\geq 0} L(x,\alpha,\beta)=p^*$ （弱对偶性）

定理2：对于原始问题和对偶问题，假设函数 $f(x)$ 和不等式约束条件 $g_j(x)$ 为凸函数，等式约束条件中的 $h_i(x)$ 为仿射函数（即由一阶多项式构成的函数， $h_i(x)=a_i^T x+b_i, a_i, x$ 均为列向量，b 为标量）；并且至少存在一个 x 使所有不等式约束条件严格成立，则存在 $\boldsymbol{x}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*$ 使得 $\boldsymbol{x}^*$ 是原始问题的最优解，$\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*$ 是对偶问题的最优解且有：$d^*=p^*=L\left(\boldsymbol{x}^*,\boldsymbol{a}^*,\boldsymbol{\beta}^*\right)$，并其充分必要条件如下：

$\nabla_x\left(x^*,\alpha^*,\beta^*\right)=0$ (1)

$\nabla_\alpha\left(x^*,\alpha^*,\beta^*\right)=0$ (2)

$\nabla_\beta\left(x^*,\alpha^*,\beta^*\right)=0$ (3)

$g_j\left(x^*\right)\leq 0, j=1,2,\cdots,n$ (4)

$\beta_j^*\geq 0, j=1,2,\cdots,n$ (5)

$\beta_j^* g_j\left(x^*\right)=0, j=1,2,\cdots,n$ (6)

$h_i\left(x^*\right)=0, i=1,2,\cdots,n$ (7)

（1）~（3）是为了求解最优化要求目标函数相对于三个变量 $\boldsymbol{x}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*$ 的梯度为0；（4）~（6）为 KKT 条件，（7）为等式约束条件。

注：证明详解可见《Convex Optimization》，by Boyd and Vandenberghe. Page-234, 5.3.2.

http://link.zhihu.com/?target=http%3A//www.stanford.edu/%7Eboyd/cvxbook/bv_cvxbook.pdf

SMO算法：

假设 $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_N)$ 为最优解，可得到分离超平面

$$g\left(x_i\right)=\sum_j^N \alpha_j y_j x_j \cdot x_i+b$$

那么就有

$$y_i \cdot g\left(x_i\right)=\begin{cases}\geq 1, & \{x_i|\alpha_i=0\} \\ =1, & \{x_i|0<\alpha_i<C\} \\ \leq 1, & \{x_i|\alpha_i=C\}\end{cases}$$

由于

$$\sum_{n=1}^N y_n \alpha_n=0$$

$$0 \leq \alpha_n \leq C, \quad for\ \ n=1,2\cdots,N$$

所以每次优化时，必须同时优化 a 的两个分量，因为只优化一个分量的话，新的 a 就不再满足初始限制条件中的等式条件了。此外每次优化的两个分量应当是违反 $g(x)$ 目标条件比较多的。就是说，本来应当是大于等于1的，越是小于1违反 $g(x)$ 目标条件就越多，这样一来，选择优化的两个分量时，就有了基本的标准。

此时，将 a_1、a_2 看做变量，其他分量看做常数，对偶问题就是一个二次函数优化问题：

$$\min_{\alpha_1,\alpha_2} W\left(\alpha_1,\alpha_2\right)=m\alpha_1^2+n\alpha_2^2+k\alpha_1\alpha_2+q\alpha_1+p\alpha_2$$

其中：$y_1\alpha_1+y_2\alpha_2=K \quad 0\leq\alpha_1\leq C, 0\leq\alpha_2\leq C$

由于 $y_i=\pm 1$，所以变为 $a_1 \pm a_2=K$。把 $a_1=K\pm a_2$ 代入目标函数就变成关于a_2 的的一元函数。

迭代更新求出 a_2 后就求出 a_1