凸优化初步

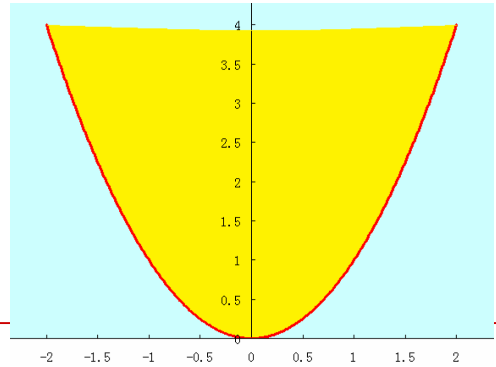
七月算法 邹博

2015年3月31日

主要内容

- 凸集基本概念
  - 凸集保凸运算
  - 分割超平面
  - 支撑超平面
- 凸函数基本概念
  - 上境图
  - Jensen不等式
  - 凸函数保凸运算
- 凸优化一般提法
  - 对偶函数
  - 鞍点解释
  - 用对偶求解最小二乘问题
  - 强对偶KKT条件

思考凸集和凸函数



- $y=x^2$是凸函数，函数图像上位于$y=x^2$上方的区域构成凸集。
  - 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为凸集，则该函数是凸函数。
  - 稍后给出上述表述的形式化定义。
- 因此，学习凸优化，考察凸函数，先从凸集及其性质开始。

凸集

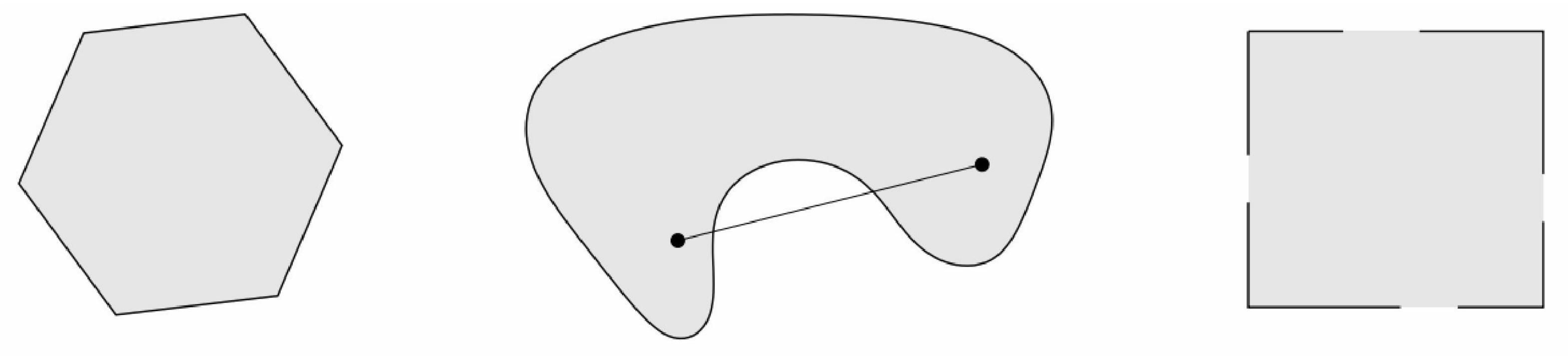
- 集合C内任意两点间的线段均在集合C内，则称集合C为凸集。

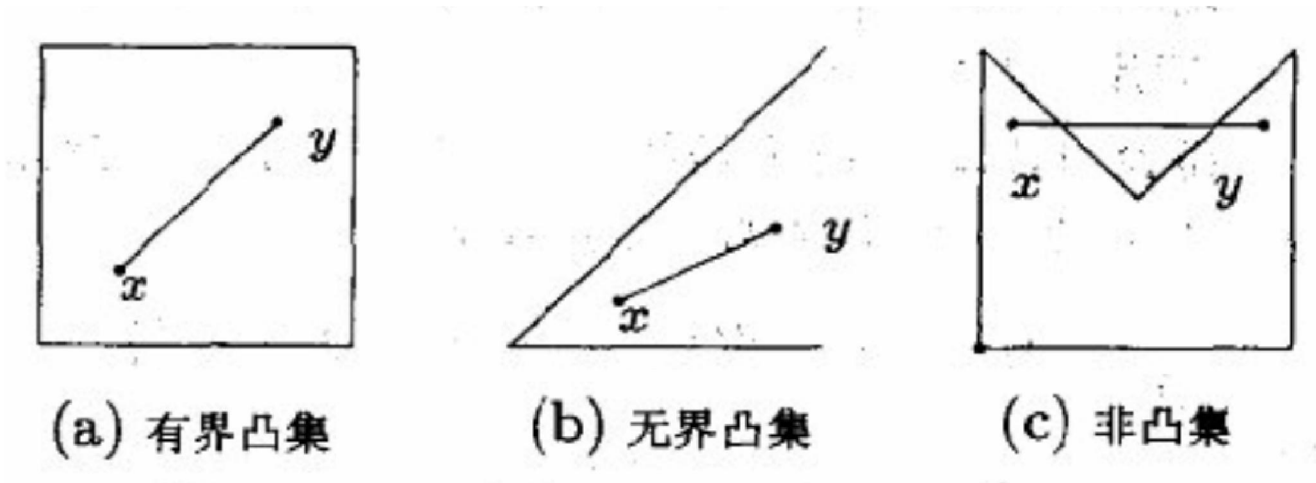
$$\forall x_1, x_2 \in C, \theta \in [0,1], \text{则} \theta x_1 + (1-\theta) x_2 \in C$$

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

$$\text{则} \sum_{i=1}^{k} \theta_i x_i \in C$$

凸集





超平面和半空间

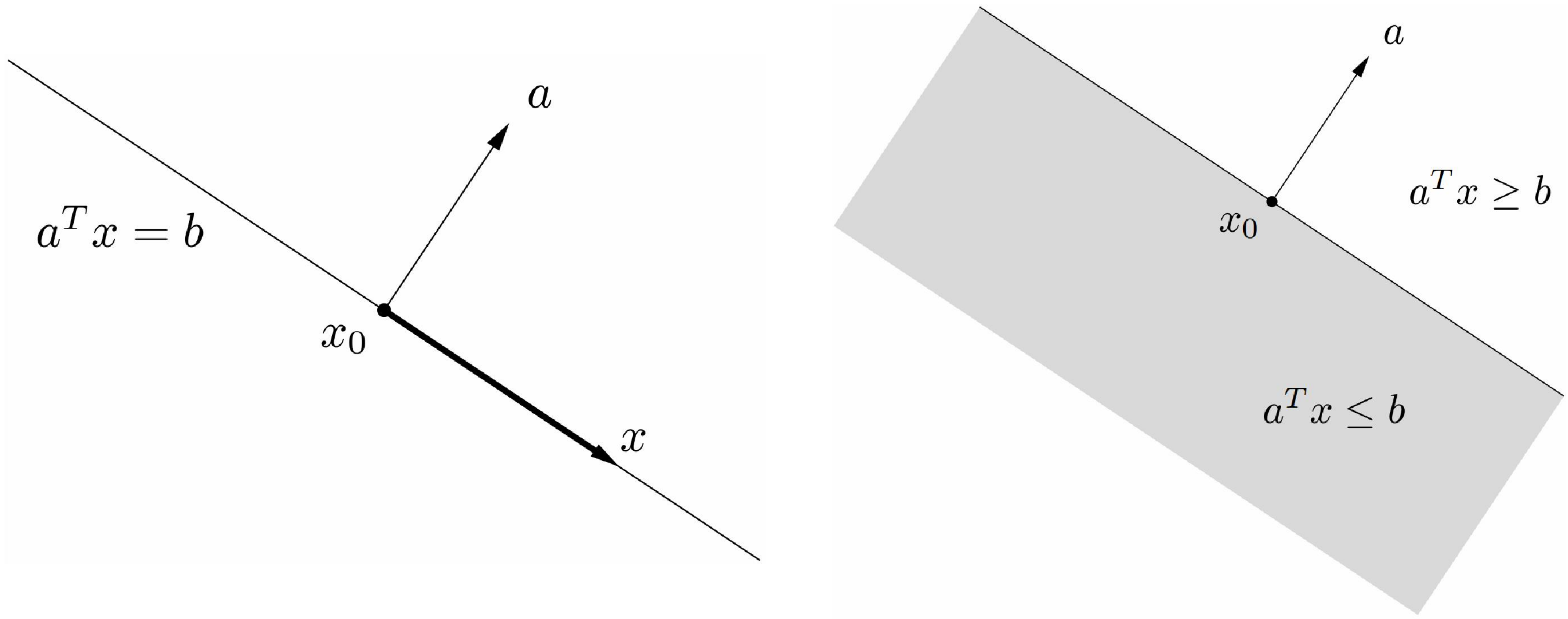
- 超平面hyperplane

$$\{x \mid a^T x = b\}$$

- 半空间halfspace

$$\{x \mid a^T x \le b\} \qquad \{x \mid a^T x \ge b\}$$

超平面和半空间



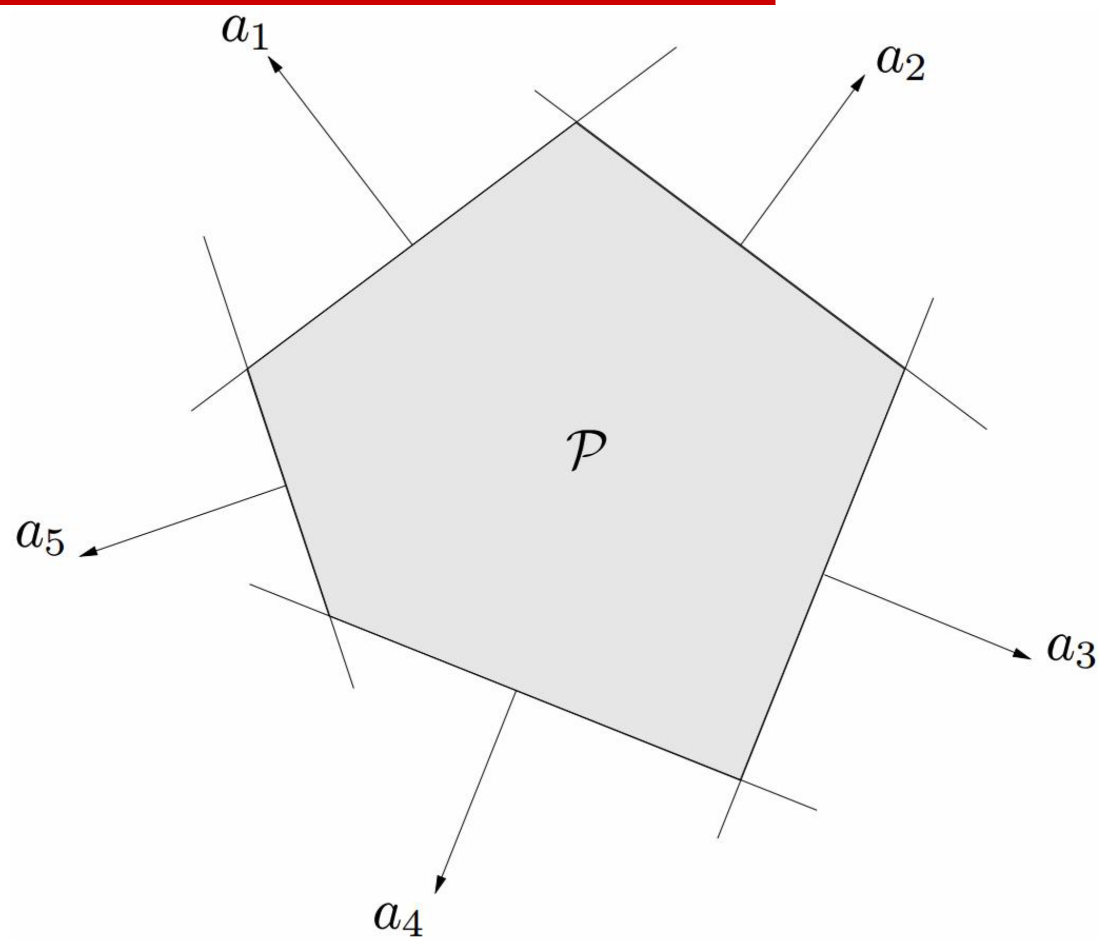
多面体

- 多面体有限个半空间和超平面的交集。

$$P=\{x \mid a_j^T x \le b_j, c_i^T x = d_i\}$$

- 仿射集(如超平面、直线)、射线、线段、半空间都是多面体。
- 多面体是凸集。
- 此外：有界的多面体有时称作多胞形(polytope)。
  - 注：该定义略混乱，不同文献的含义不同。

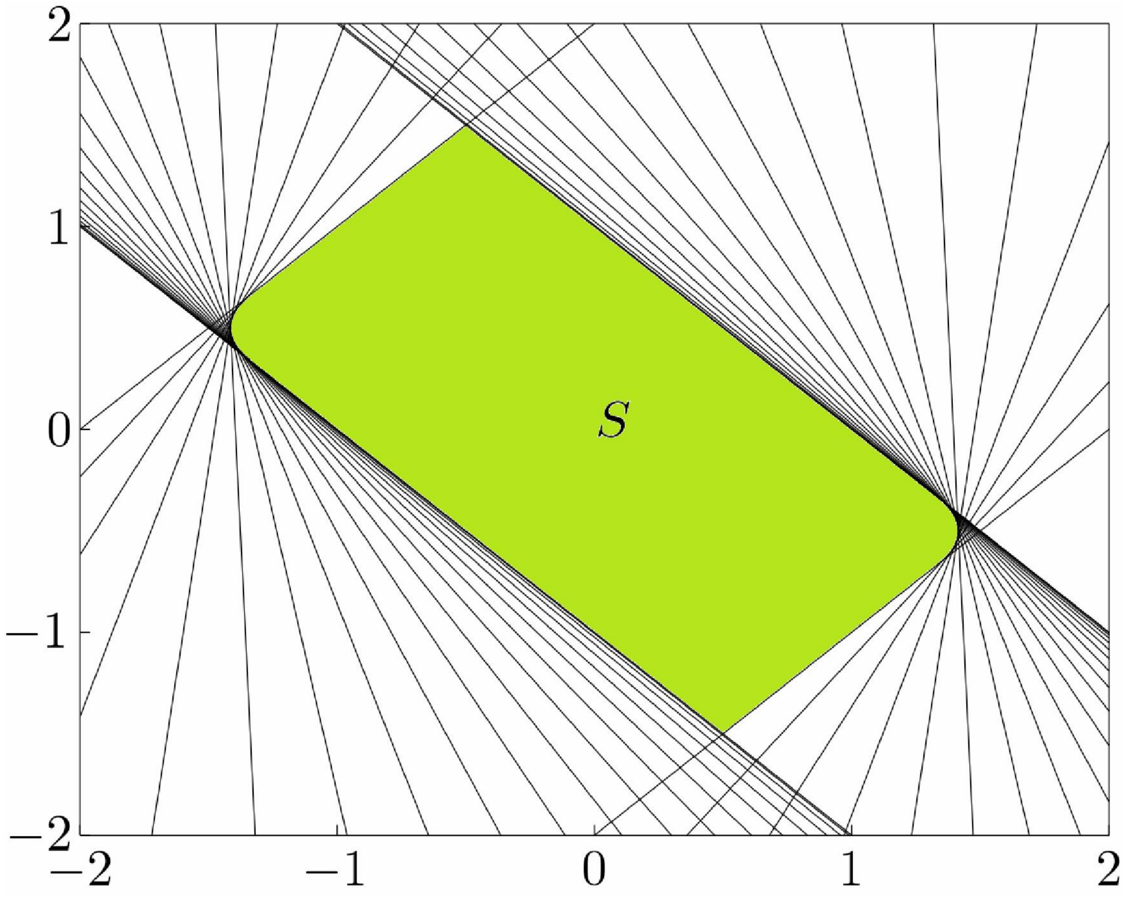
多面体



保持凸性的运算

- 集合交运算
  - 思考：如何证明？(提示：根据定义)
- 仿射变换
  - 函数f=Ax+b的形式，称函数是仿射的：即线性函数加常数的形式
- 透视变换
- 投影变换(线性分式变换)

集合交运算：半空间的交



仿射变换

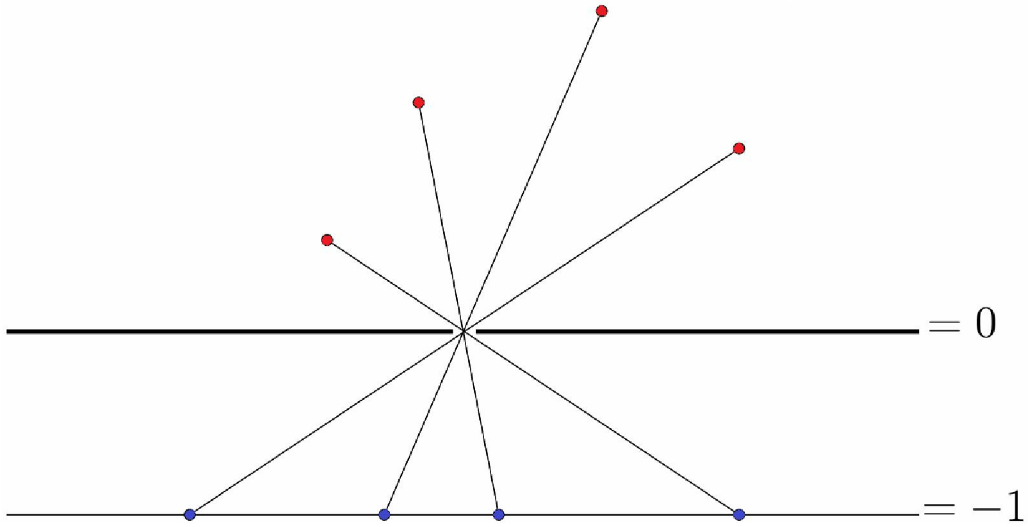
- 仿射变换 $f(x) = Ax + b,\ A \in \mathbf{R}^{m\times n}, b \in \mathbf{R}^m$
  - 伸缩、平移、投影
- 若f是仿射变换，$f : \mathbf{R}^n \to \mathbf{R}^m \quad f(S) = \{f(x) \mid x \in S\}$
  - 若S为凸集，则f(S)为凸集；
  - 若f(S)为凸集，则S为凸集。

透视变换

- 透视函数对向量进行伸缩(规范化)，使得最后一维的分量为1并舍弃之。

$$P:\mathbf{R}^{n+1}\rightarrow\mathbf{R}^{n},\ P(z,t)=z/t$$

- 透视的直观意义
  - 小孔成像



透视变换的保凸性

- 凸集的透视变换仍然是凸集。
- 思考：反过来，若某集合的透视变换是凸集，这个集合一定是凸集吗？

投射函数(线性分式函数)

- 投射函数是透视函数和仿射函数的复合。
- g为仿射函数：

$$g : \mathbf{R}^n \to \mathbf{R}^{m+1}$$
$$g(x) = \begin{bmatrix} A \\ c^T \end{bmatrix} x + \begin{bmatrix} b \\ d \end{bmatrix}$$
$$A \in \mathbf{R}^{m \times n},\ b \in \mathbf{R}^m,\ c \in \mathbf{R}^n,\ d \in \mathbf{R}$$

- 定义f为线性分式函数

$$f(x) = (Ax + b)/(c^T x + d),\ \mathbf{dom}\, f = \{x \mid c^T x + d > 0\}$$

- 若c=0,d>0，则f即为普通的仿射函数。

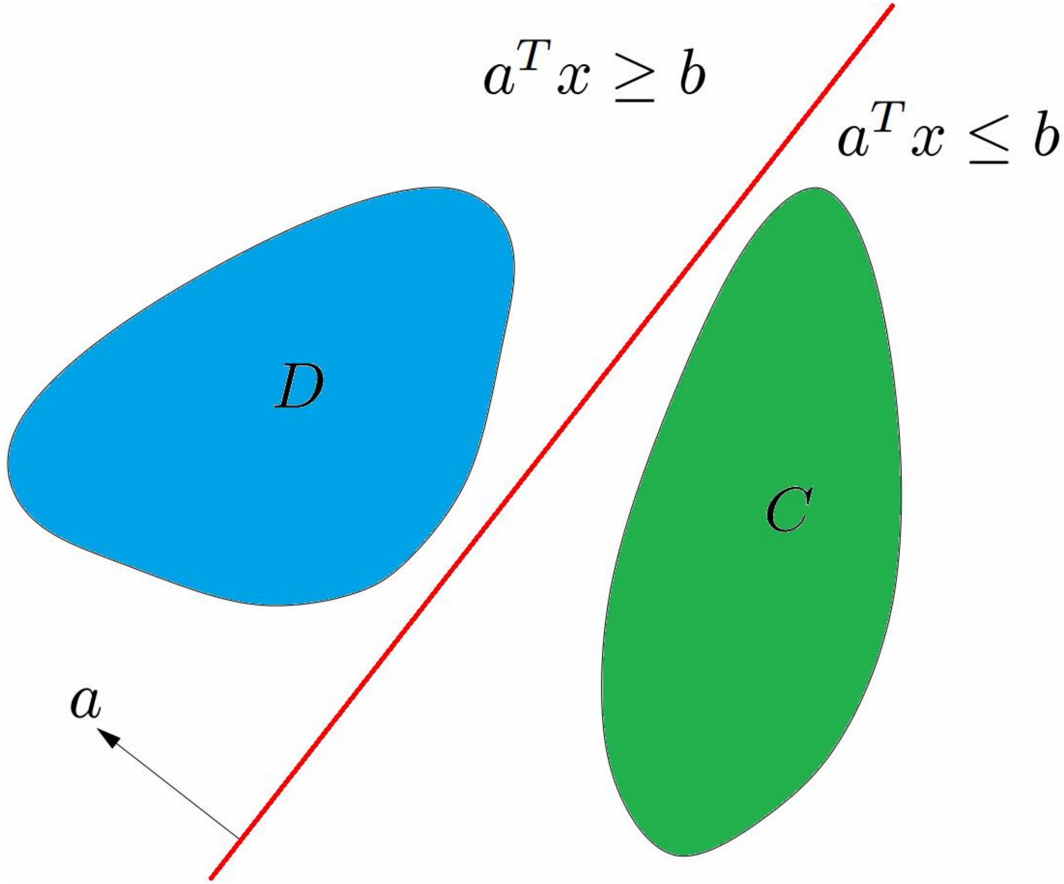
分割超平面

- 设C和D为两不相交的凸集，则存在超平面P，P可以将C和D分离。

$$\forall x \in C, a^T x \le b \text{且} \forall x \in D, a^T x \ge b$$

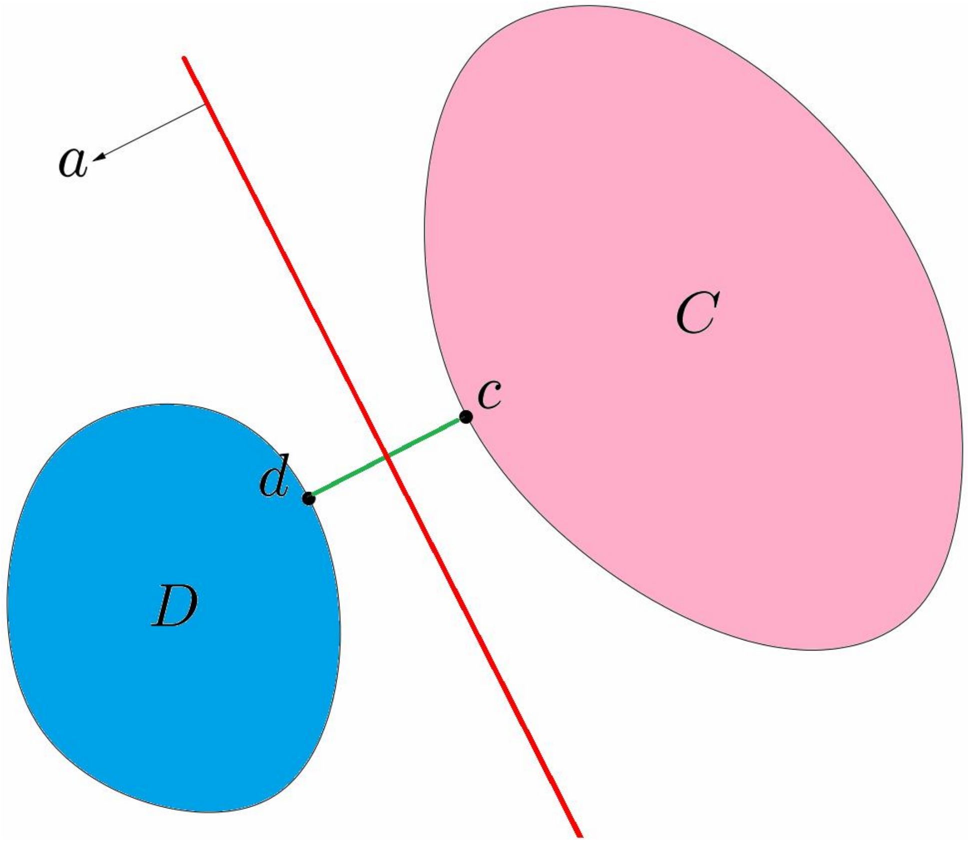
- 注意上式中可以取等号：
  - 所以：逆命题："若两个凸集C和D的分割超平面存在，C和D不相交"为假命题。
  - 加强条件：若两个凸集至少有一个是开集，那么当且仅当存在分割超平面，它们不相交。

分割超平面

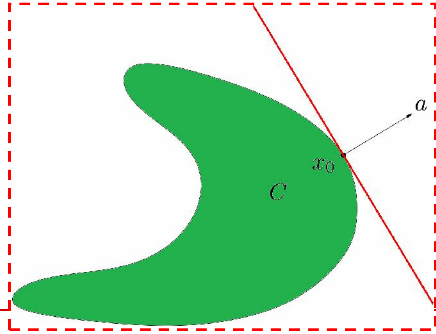


分割超平面的构造

- 两个集合的距离，定义为两个集合间元素的最短距离。
- 做集合C和集合D最短线段的垂直平分线。

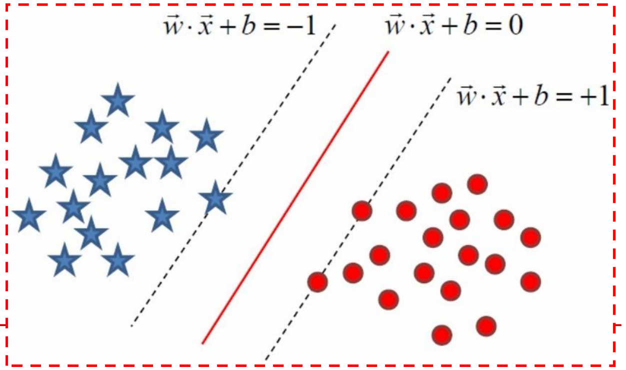


支撑超平面



- 设集合C，x0为C边界上的点。若存在a≠0，满足对任意x∈C，都有 $a^T x \le a^T x_0$ 成立，则称超平面 $\{x \mid a^T x = a^T x_0\}$ 为集合C在点x0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。

思考



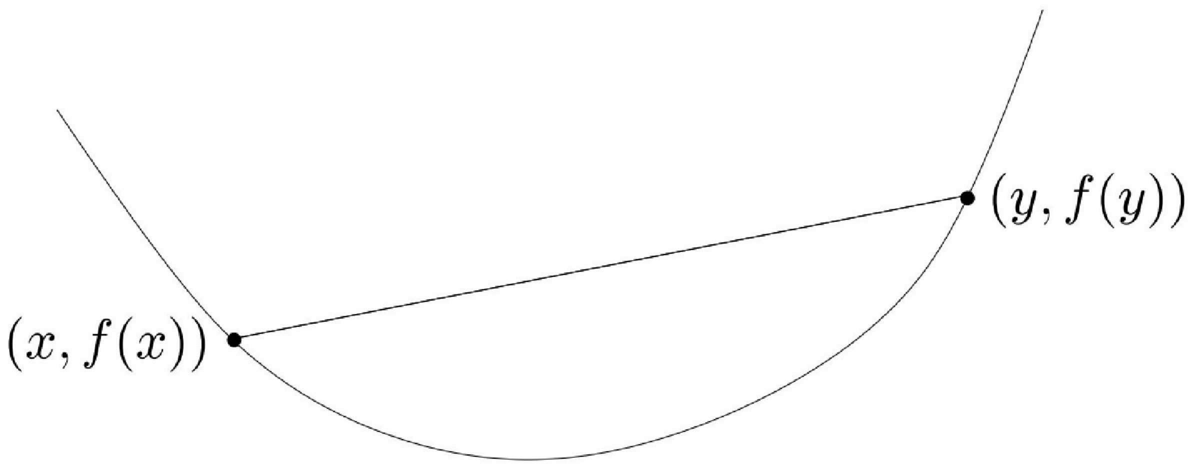
- 如何定义两个集合的"最优"分割超平面？
  - 找到集合"边界"上的若干点，以这些点为"基础"计算超平面的方向；以两个集合边界上的这些点的平均作为超平面的"截距"
  - 支持向量：support vector
- 若两个集合有部分相交，如何定义超平面，使得两个集合"尽量"分开？
  - 注：上述"集合"不一定是凸集，可能是由若干离散点组成。若一组集合为(**x**,1)，另一组集合为(**x**,2)，则为机器学习中的分类问题。

凸函数

- 若函数f的定义域domf为凸集，且满足

$$\forall x, y \in \operatorname{dom} f, 0 \le \theta \le 1,\text{ 有}$$

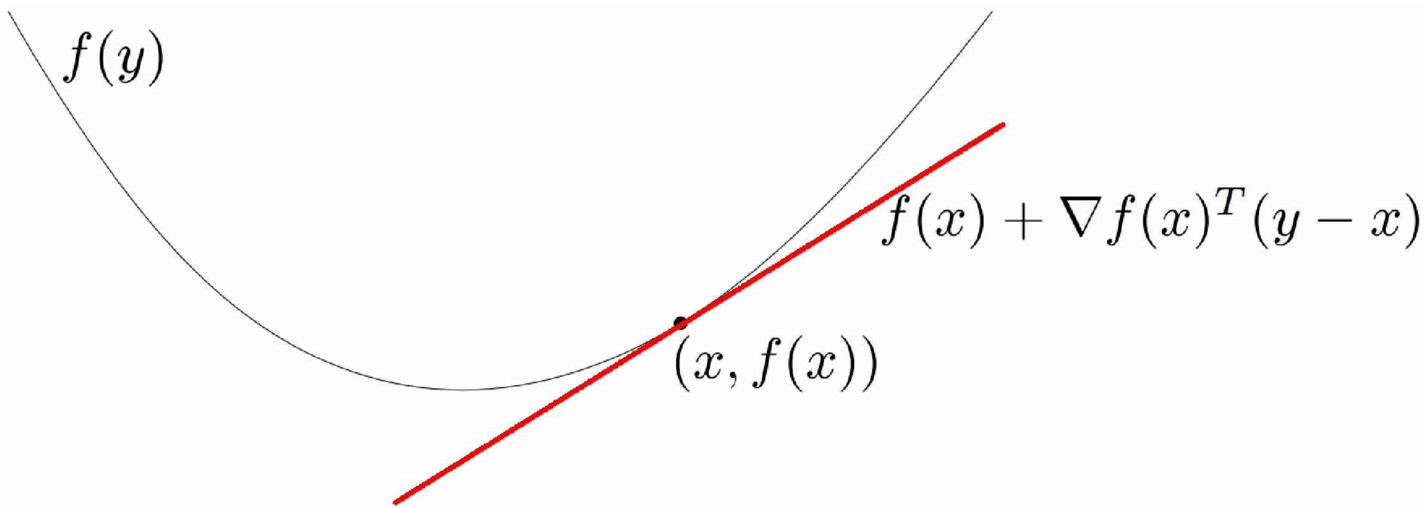
$$f(\theta x + (1-\theta) y) \le \theta f(x) + (1-\theta) f(y)$$



一阶可微

- 若f一阶可微，则函数f为凸函数当前仅当f的定义域domf为凸集，且

$$\forall x, y \in \mathbf{dom} f,\ f(y) \geq f(x) + \nabla f(x)^T (y - x)$$



进一步的思考 $f(y) \geq f(x) + \nabla f(x)^T (y - x)$

- 结合凸函数图像和支撑超平面理解该问题
- 对于凸函数，其一阶Taylor近似本质上是该函数的全局下估计。
- 反之，如果一个函数的一阶Taylor近似总是起全局下估计，则该函数是凸函数。
- 该不等式说明从一个函数的局部信息，可以得到一定程度的全局信息。

二阶可微

- 若函数f二阶可微，则函数f为凸函数当前仅当dom为凸集，且

$$\nabla^2 f(x) \succ= 0$$

- 若f是一元函数，上式表示二阶导大于等于0
- 若f是多元函数，上式表示二阶导Hessian矩阵半正定。

凸函数举例

- 指数函数 e^{ax}
- 幂函数 $x^{a}, x \in R_{+}, a \geq 1 \text{ or } a \leq 0$
- 负对数函数 $-\log x$
- 负熵函数 $x \log x$
- 范数函数 $\|x\|_{p}$

$$f(x) = \max(x_1, \ldots, x_n)$$

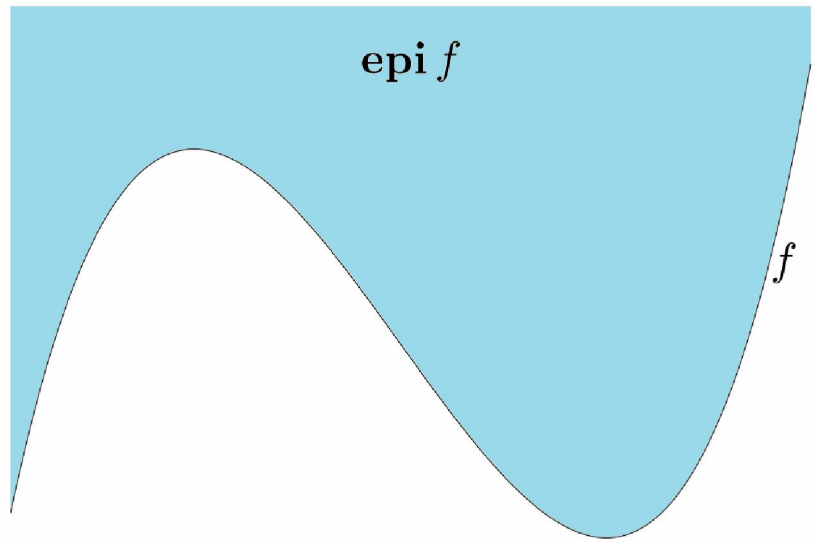
$$f(x) = x^2 / y, y > 0$$

$$f(x) = \log(e^{x_1} + \ldots + e^{x_n})$$

上境图

- 函数f的图像定义为：$\{(x, f(x)) \mid x \in \mathbf{dom}\, f\}$
- 函数f的上境图(epigraph)定义为：

$$\mathbf{epi}\, f = \{(x, t) \mid x \in \mathbf{dom}\, f,\ f(x) \leq t\}$$



凸函数与凸集

- 一个函数是凸函数，当且仅当其上境图是凸集。
  - 思考：如何证明？(提示：定义)
- 进一步，一个函数是凹函数，当且仅当其亚图(hypograph)是凸集。

$$\textbf{hypo}\, f = \{(x,t) \mid t \le f(x)\}$$

Jensen不等式：若f是凸函数

- 基本Jensen不等式

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

- 若 $\theta_1, \ldots, \theta_k \ge 0,\ \theta_1 + \cdots + \theta_k = 1$
- 则 $f(\theta_1 x_1 + \cdots + \theta_k x_k) \le \theta_1 f(x_1) + \cdots + \theta_k f(x_k)$

- 若 $p(x) \ge 0$ on $S \subseteq \mathbf{dom}\, f$, $\int_S p(x)\, dx = 1$
- 则 $f\left(\int_S p(x) x\, dx\right) \le \int_S f(x) p(x)\, dx$

$$f(\mathbf{E}\, x) \le \mathbf{E}\, f(x)$$

Jensen不等式是几乎所有不等式的基础

- 利用f(E(x))≤E(f(x))，(f是凸函数)，证明下式D≥0

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

- 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，θ =0.5带入基本Jensen不等式

保持函数凸性的算子

- 凸函数的非负加权和

$$f(x) = \omega_1 f_1(x) + \ldots + \omega_n f_n(x)$$

- 凸函数与仿射函数的复合

$$g(x) = f(Ax + b)$$

- 凸函数的逐点最大值、逐点上确界

$$f(x) = \max(f_1(x), \ldots, f_n(x))$$

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

凸函数的逐点最大值

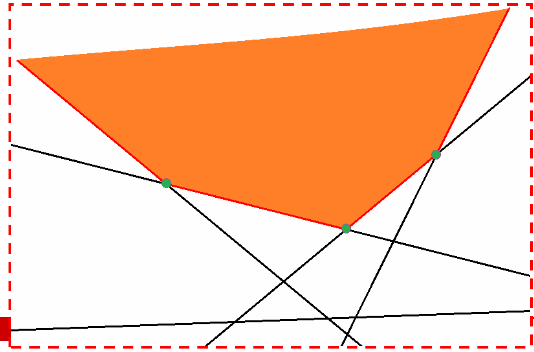
- f1,f2均为凸函数，定义函数f：

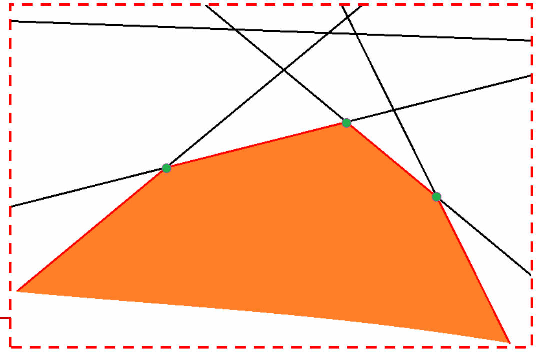
$$f(x) = \max\{f_1(x), f_2(x)\}$$

- 则函数f为凸函数。

$$\begin{aligned}
& f(\theta x + (1-\theta)y) \\
= \ & \max\{f_1(\theta x + (1-\theta)y), f_2(\theta x + (1-\theta)y)\} \\
\le \ & \max\{\theta f_1(x) + (1-\theta) f_1(y), \theta f_2(x) + (1-\theta) f_2(y)\} \\
\le \ & \theta \max\{f_1(x), f_2(x)\} + (1-\theta) \max\{f_1(y), f_2(y)\} \\
= \ & \theta f(x) + (1-\theta) f(y)
\end{aligned}$$

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(直线是凸的——虽然直线也是凹的)，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，是凹函数；
    - 在Lagrange对偶函数中会用到该结论。

凸优化

- 优化问题的基本形式

$$
\begin{aligned}
&\text{minimize} && f_0(x), \quad x \in \mathbf{R}^n \\
&\text{subject to} && f_i(x) \le 0, \quad i = 1, \dots, m \\
& && h_j(x) = 0, \quad j = 1, \dots, p
\end{aligned}
$$

优化变量	$x \in R^n$
不等式约束	$f_i(x) \le 0$
等式约束	$h_j(x) = 0.$
无约束优化	$m = p = 0$

优化问题的基本形式

- 优化问题的域

$$D=\bigcap_{i=0}^{m}\operatorname{dom} f_i \cap \bigcap_{j=1}^{p}\operatorname{dom} h_j$$

- 可行点(解)(feasible)
  - x∈D，且满足约束条件
- 可行域(可解集)
  - 所有可行点的集合
- 最优化值

$$p^* = \inf\{f_0(x) \mid f_i(x)\le 0, i=1,\dots,m, h_j(x)=0, j=1,\dots,p\}$$

- 最优化解

$$p^* = f_0(x^*)$$

凸优化问题的基本形式

$$
\begin{aligned}
\text{minimize} \quad & f_0(x), \quad x \in \mathbf{R}^n \\
\text{subject to} \quad & f_i(x) \le 0, \quad i = 1, \dots, m \\
& h_j(x) = 0, \quad j = 1, \dots, p
\end{aligned}
$$

- $f_i(x)(0 \leqslant i \leqslant m)$为凸函数，$h_j(x)(1 \leqslant i \leqslant p)$为仿射函数
- 凸优化问题的重要性质
  - 可行域为凸集
  - 局部最优解即为全局最优解

对偶问题

- 一般优化问题的Lagrange乘子法

$$\begin{aligned} \text{minimize} \quad & f_0(x), \quad x \in \mathbf{R}^n \\ \text{subject to} \quad & f_i(x) \le 0, \quad i = 1, \dots, m \\ & h_j(x) = 0, \quad j = 1, \dots, p \end{aligned}$$

- Lagrange函数

$$L(x, \lambda, v) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} v_j h_j(x)$$

  - 对固定的x，Lagrange函数L(x, λ ,v)为关于λ和v的仿射函数

Lagrange对偶函数(dual function)

- Lagrange对偶函数

$$g(\lambda,\nu)=\inf_{x\in D} L(x,\lambda,\nu)=\inf_{x\in D}\left(f_0(x)+\sum_{i=1}^{m}\lambda_i f_i(x)+\sum_{i=1}^{p}\nu_i h_i(x)\right)$$

- 若没有下确界，定义：

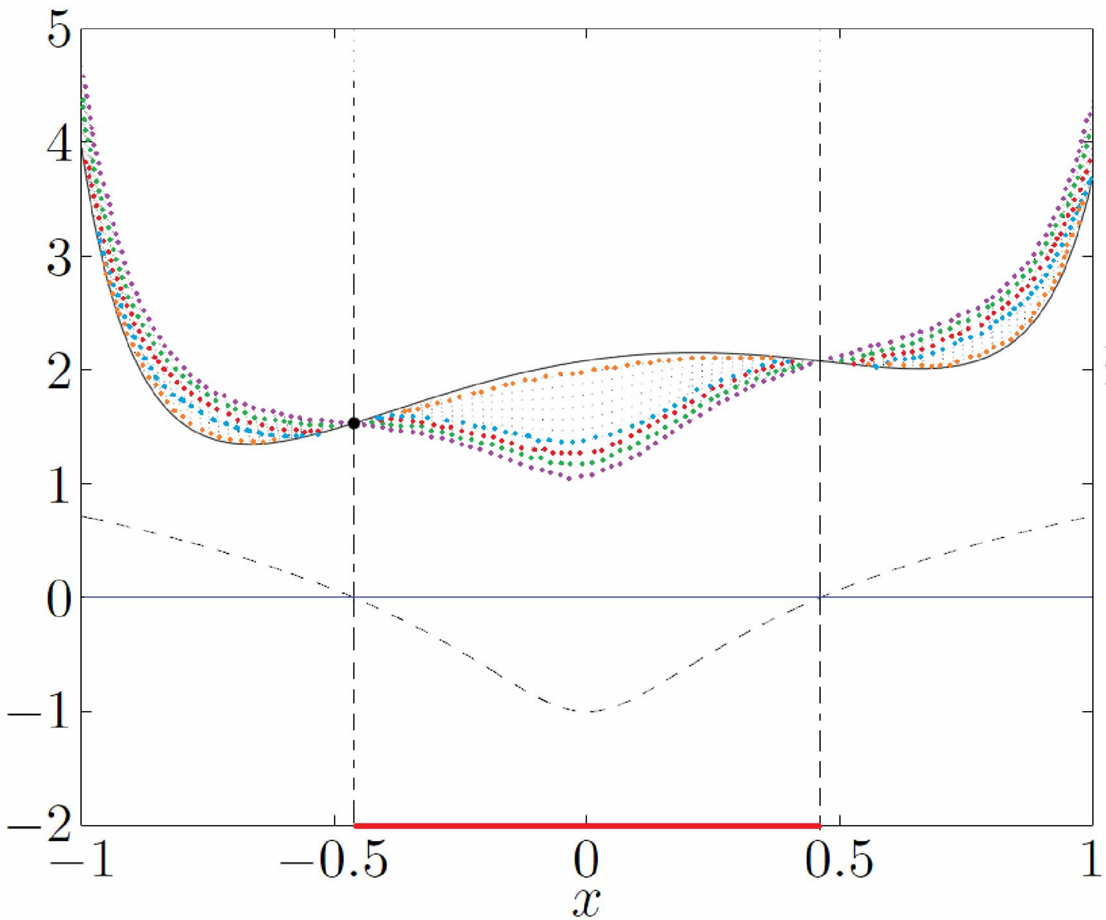
$$g(\lambda,\nu)=-\infty$$

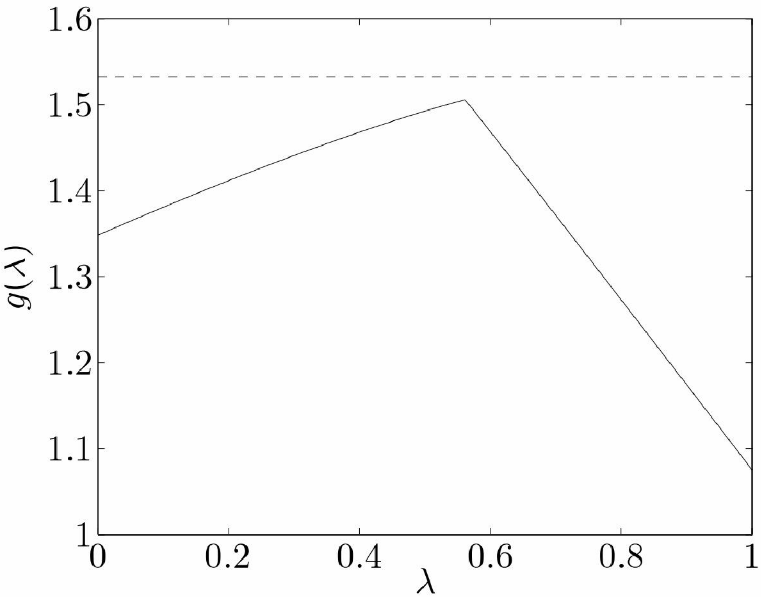
- 根据定义，显然有：对∀ λ ≥0，∀v，若原优化问题有最优值p*，则

$$g(\lambda,\nu)\le p^*$$

- 进一步：Lagrange对偶函数为凹函数。

左侧为原函数，右侧为对偶函数





鞍点解释

- 为表述方便，假设没有等式约束，只考虑不等式约束，结论可方便的扩展到等式约束。
- 假设x0不可行，即存在某些i，使得$f_i(x)>0$。则选择 $\lambda_i \to \infty$，对于其他乘子，$\lambda_j = 0, j \neq i$
- 假设x0可行，则有$f_i(x) \leqslant 0, (i=1,2,\ldots,m)$，选择

$$\lambda_i = 0, i = 1,2,\cdots,m$$

- 有：

$$\sup_{\lambda \ge 0} L(x,\lambda) = \sup_{\lambda \ge 0}\left(f_0(x) + \sum_{i=1}^{m}\lambda_i f_i(x)\right) = \begin{cases} f_0(x) & f_i(x) \le 0, i = 1,2\cdots,m \\ \infty & otherwise \end{cases}$$

鞍点：最优点

- 而原问题是：$\inf_x f_0(x)$
- 从而，原问题的本质为：$\inf_x \sup_{\lambda \geq 0} L(x,\lambda)$
- 而对偶问题，是求对偶函数的最大值，即：

$$\sup_{\lambda \geq 0} \inf_x L(x,\lambda)$$

- 而：

$$\sup_{\lambda \geq 0} \inf_x L(x,\lambda) \leq \inf_x \sup_{\lambda \geq 0} L(x,\lambda)$$

证明：$\max_x \min_y f(x,y) \le \min_y \max_x f(x,y)$

- 对于任意的(x,y)∈domf

$$f(x,y) \le \max_x f(x,y)$$

$$\Rightarrow \min_y f(x,y) \le \min_y \max_x f(x,y)$$

$$\Rightarrow \max_x \min_y f(x,y) \le \min_y \max_x f(x,y)$$

线性方程的最小二乘问题

- 原问题 $$\text{minimize} \quad x^T x, \quad x \in \mathbf{R}^n$$ $$\text{subject to} \quad Ax = b$$
- Lagrange函数
$$L(x,v) = x^T x + v^T (Ax - b)$$
- Lagrange对偶函数
$$g(v) = -\frac{1}{4} v^T A A^T v - b^T v$$
  - 对L求x的偏导，带入L，得到g
  - 对g求v的偏导，求g的极大值，作为原问题的最小值

求L的对偶函数 $L(x,\nu)=x^T x+\nu^T(Ax-b)$

$$\frac{\partial L}{\partial x}=\frac{\partial\left(x^T x+\nu^T(Ax-b)\right)}{\partial x}=2x+A^T\nu \overset{令}{=} 0 \Rightarrow x^*=-\frac{1}{2}A^T\nu$$

$$\begin{aligned}
L(x,\nu)&=x^T x+\nu^T(Ax-b)\\
&=\left(-\frac{1}{2}A^T\nu\right)^T\left(-\frac{1}{2}A^T\nu\right)+\nu^T\left(A\left(-\frac{1}{2}A^T\nu\right)-b\right)\\
&=\frac{1}{4}\nu^T AA^T\nu-\frac{1}{2}\nu^T AA^T\nu-\nu^T b\\
&=-\frac{1}{4}\nu^T AA^T\nu-\nu^T b\\
&\overset{\Delta}{=} g(\nu)
\end{aligned}$$

求对偶函数的极大值 $g(\nu) = -\frac{1}{4}\nu^T AA^T \nu - \nu^T b$

$$\frac{\partial g}{\partial \nu} = \frac{\partial\left(-\frac{1}{4}\nu^T AA^T \nu - \nu^T b\right)}{\partial \nu} = -\frac{1}{2}AA^T \nu - b \overset{令}{=} 0$$

$$\Rightarrow AA^T \nu = -2b$$

$$\Rightarrow A^T AA^T \nu = -2A^T b$$

$$\Rightarrow A^T \nu = -2\left(A^T A\right)^{-1} A^T b$$

$$\Rightarrow -\frac{1}{2}A^T \nu = \left(A^T A\right)^{-1} A^T b$$

$$\Rightarrow x^* = \left(A^T A\right)^{-1} A^T b$$

极小值点 $x^* = \left(A^T A\right)^{-1} A^T b$

- 极小值： $\min\left(x^T x\right)$

$$= \left(\left(A^T A\right)^{-1} A^T b\right)^T \left(\left(A^T A\right)^{-1} A^T b\right)$$

$$= b^T A\left(A^T A\right)^{-1}\left(A^T A\right)^{-1} A^T b$$

$$= b^T A\left(A^T A\right)^{-2} A^T b$$

- 极小值点的结论，和通过线性回归计算得到的结论是完全一致的。
  - 线性回归问题具有强对偶性。

强对偶条件

- 若要对偶函数的最大值即为原问题的最小值，考察需要满足的条件：

$$
\begin{aligned}
f_0(x^\star) &= g(\lambda^\star, \nu^\star) \\
&= \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^\star f_i(x) + \sum_{i=1}^{p} \nu_i^\star h_i(x) \right) \\
&\le f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star f_i(x^\star) + \sum_{i=1}^{p} \nu_i^\star h_i(x^\star) \\
&\le f_0(x^\star).
\end{aligned}
$$

Karush-Kuhn-Tucker (KKT)条件

$$\begin{aligned} f_i(x^\star) &\leq 0, \quad i = 1, \ldots, m \\ h_i(x^\star) &= 0, \quad i = 1, \ldots, p \\ \lambda_i^\star &\geq 0, \quad i = 1, \ldots, m \\ \lambda_i^\star f_i(x^\star) &= 0, \quad i = 1, \ldots, m \end{aligned}$$

$$\nabla f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \nabla f_i(x^\star) + \sum_{i=1}^{p} \nu_i^\star \nabla h_i(x^\star) = 0$$

参考文献

- Convex Optimization, Stephen Boyd,Lieven Vandenberghe, Cambridge University Press, 2004
  - 中译本：王书宁，许鋆，黄晓霖 译，凸优化，清华大学出版社，2013
- 同济大学数学教研室 主编，高等数学，高等教育出版社，1996
- 同济大学数学系 编，工程数学线性代数(第五版)，高等教育出版社，2007

我们在这里

- 七月算法官网：http://www.julyedu.com/
  - 免费视频
  - 直播课程
  - 问答社区
- 联系我们：微博
  - @研究者July
  - @七月问答
  - @邹博_机器学习

感谢大家！

恳请大家批评指正！