

法律声明

□ 本课件包括演示文稿、示例、代码、题库、视频和声音等内容，小象学院和主讲老师拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意及内容，我们保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：小象

■ 新浪微博：ChinaHadoop



凸优化



小象学院
ChinaHadoop.cn

邹博

凸优化主要内容

- 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

思考两个不等式

□ 两个正数的算术平均数大于等于几何平均数

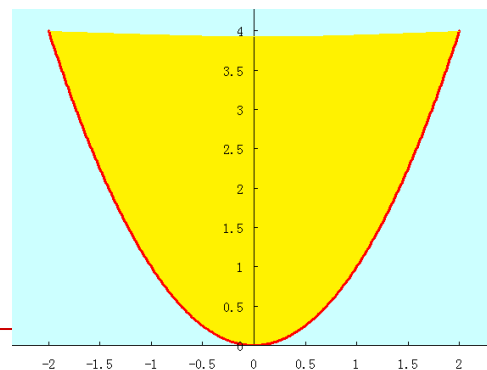
$$\frac{a+b}{2} \geq \sqrt{ab}, \quad a > 0, b > 0$$

□ 给定可逆对称阵 Q ，对于任意的向量 x 、 y ，
有：

$$x^T Q x + y^T Q^{-1} y \geq 2x^T y$$

■ 都可以在凸函数的框架下得到解决。

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。

直线的向量表达

□ 已知二维平面上两定点A(5,1)、B(2,3)，试给出经过点A、B的直线方程。

$$\begin{cases} x_1 = \theta \cdot 5 + (1 - \theta) \cdot 2 \\ x_2 = \theta \cdot 1 + (1 - \theta) \cdot 3 \end{cases}, \theta \in R$$

$$\Rightarrow 2x_1 + 3x_2 - 13 = 0$$

□ 写成向量形式： $\vec{x} = \theta \cdot \vec{a} + (1 - \theta) \cdot \vec{b}$

■ 其中， $\vec{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

几何体的向量表达

□ 给定二维平面上两个定点： $\mathbf{a}(u_1, v_1)$, $\mathbf{b}(u_2, v_2)$, 则：

■ 直线 $\vec{x} = \theta \cdot \vec{a} + (1 - \theta) \cdot \vec{b}$, $\theta \in R$

■ 线段 $\vec{x} = \theta \cdot \vec{a} + (1 - \theta) \cdot \vec{b}$, $\theta \in [0, 1]$

□ 推广到高维：

□ 三维平面： $\vec{x} = \theta_1 \cdot \vec{a}_1 + \theta_2 \cdot \vec{a}_2 + \theta_3 \cdot \vec{a}_3$, $\theta_1, \theta_2, \theta_3 \in R$, $\theta_1 + \theta_2 + \theta_3 = 1$

□ 三角形： $\vec{x} = \theta_1 \cdot \vec{a}_1 + \theta_2 \cdot \vec{a}_2 + \theta_3 \cdot \vec{a}_3$, $\theta_1, \theta_2, \theta_3 \in [0, 1]$, $\theta_1 + \theta_2 + \theta_3 = 1$

□ 超平面：

$$\vec{x} = \theta_1 \cdot \vec{a}_1 + \theta_2 \cdot \vec{a}_2 + \cdots + \theta_k \cdot \vec{a}_k, \theta_1, \theta_2, \cdots, \theta_k \in R, \theta_1 + \theta_2 + \cdots + \theta_k = 1$$

□ (超)几何体：

$$\vec{x} = \theta_1 \cdot \vec{a}_1 + \theta_2 \cdot \vec{a}_2 + \cdots + \theta_k \cdot \vec{a}_k, \theta_1, \theta_2, \cdots, \theta_k \in [0, 1], \theta_1 + \theta_2 + \cdots + \theta_k = 1$$

仿射集(Affine set)

- 定义：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。

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

- 仿射集的例子：直线、平面、超平面

- 超平面： $Ax=b$

- $f(x)=0$ 表示定义域在 R^n 的超曲面：令 $f(x)=Ax-b$ ，则 $f(x)=0$ 表示“截距”为 b 的超平面。

- 三维空间的平面是二维的；四维空间的平面是几维的？

- n 维空间的 $n-1$ 维仿射集为 $n-1$ 维超平面。

- 后面将继续考察超平面的定义。

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。

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

□ k个点的版本：

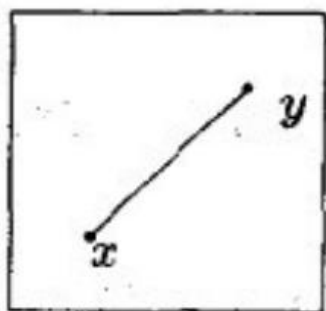
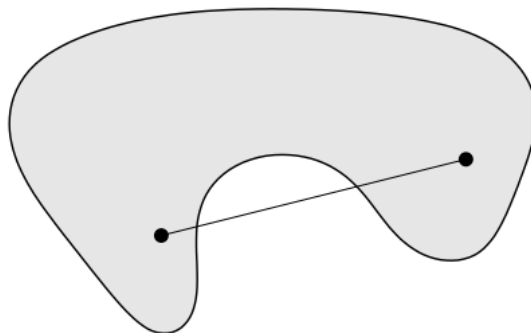
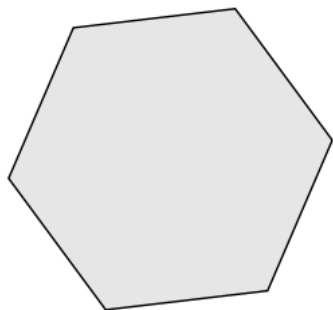
$$\forall x_1, x_2, \dots, x_k \in C, \theta_i \in [0,1] \text{ 且 } \sum_{i=1}^k \theta_i = 1, \text{ 则 } x = \sum_{i=1}^k \theta_i x_i \in C$$

■ 思考：两种表述的内涵一样吗？

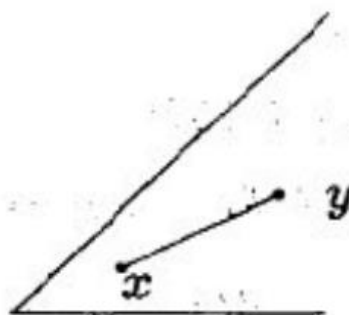
仿射集和凸集的关系

- 因为仿射集的条件比凸集的条件强，所以，仿射集必然是凸集。

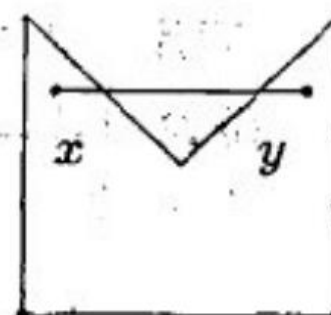
凸集



(a) 有界凸集



(b) 无界凸集



(c) 非凸集

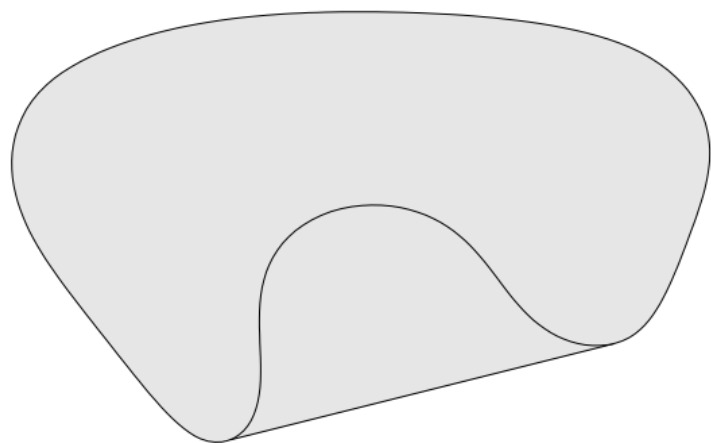
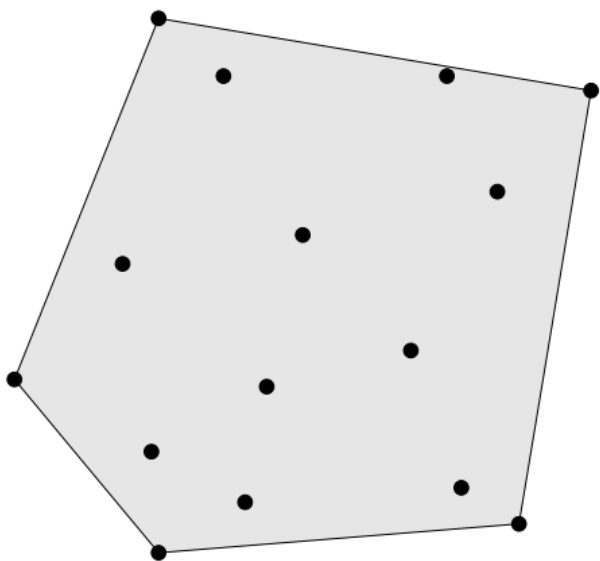
凸包

- 集合C的所有点的凸组合形成的集合，叫做集合C的凸包。

$$\text{conv } C = \left\{ \sum_{i=1}^k \theta_i x_i \mid x_i \in C, \theta_i \geq 0, \sum_{i=1}^k \theta_i = 1 \right\}$$

- 集合C的凸包是能够包含C的最小的凸集。

凸包的例子



超平面和半空间

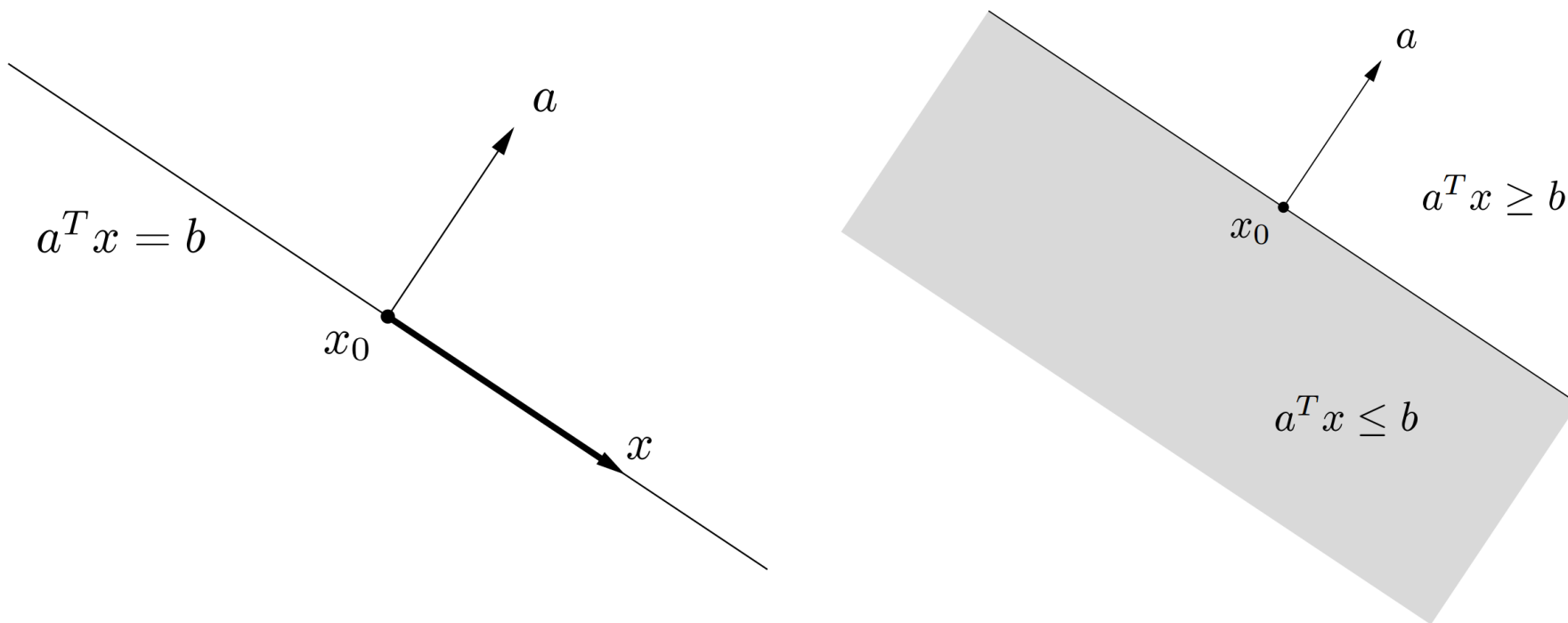
□ 超平面hyperplane

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

□ 半空间halfspace

$$\{x \mid a^T x \leq b\} \quad \{x \mid a^T x \geq b\}$$

超平面和半空间



欧式球和椭圆

□ 欧式球

$$\begin{aligned} B(x_c, r) &= \{x \mid \|x - x_c\|_2 \leq r\} \\ &= \{x \mid (x - x_c)^T (x - x_c) \leq r^2\} \end{aligned}$$

□ 椭圆

$$E = \{x \mid (x - x_c)^T P (x - x_c) \leq r^2\},$$

P 为对称正定矩阵

范数球和范数锥(欧式空间的推广)

□ 范数 $\|x\| \geq 0, \|x\| = 0$ 当且仅当 $x = 0$;

$$\|tx\| = |t| \|x\|, t \in R;$$

$$\|x + y\| \leq \|x\| + \|y\|$$

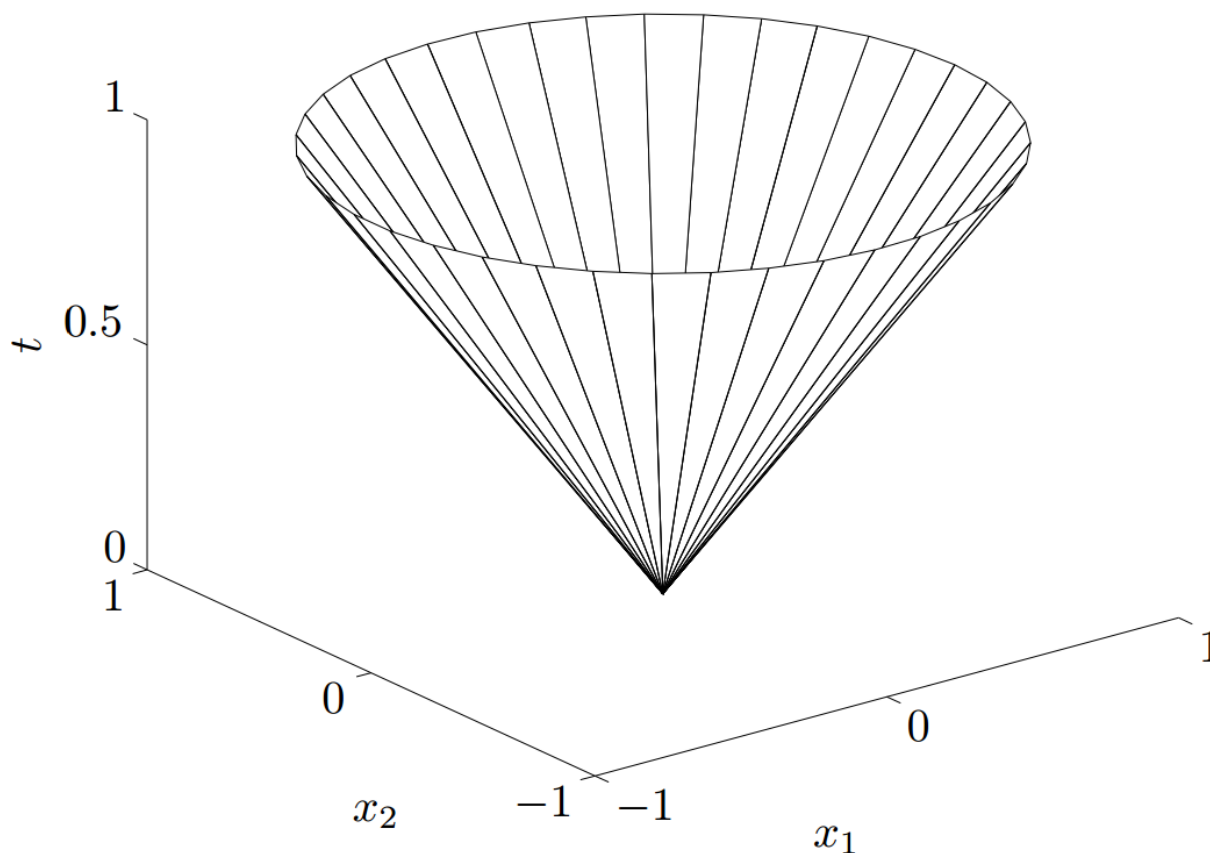
□ 范数球

$$B(x_c, r) = \{x \mid \|x - x_c\| \leq r\}$$

□ 范数锥

$$\{(x, t) \mid \|x\| \leq t\}$$

R^3 空间中的二阶锥



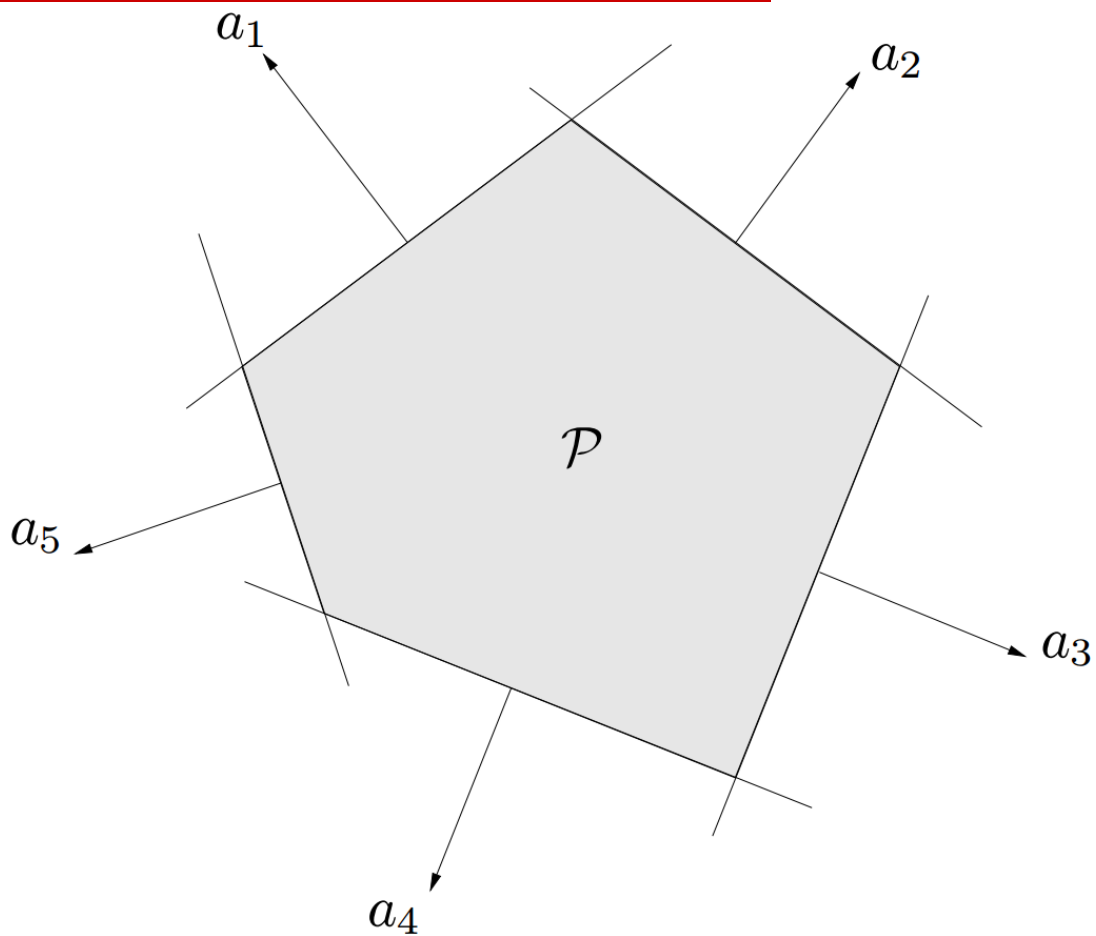
多面体

- 多面体有限个半空间和超平面的交集。

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

- 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）

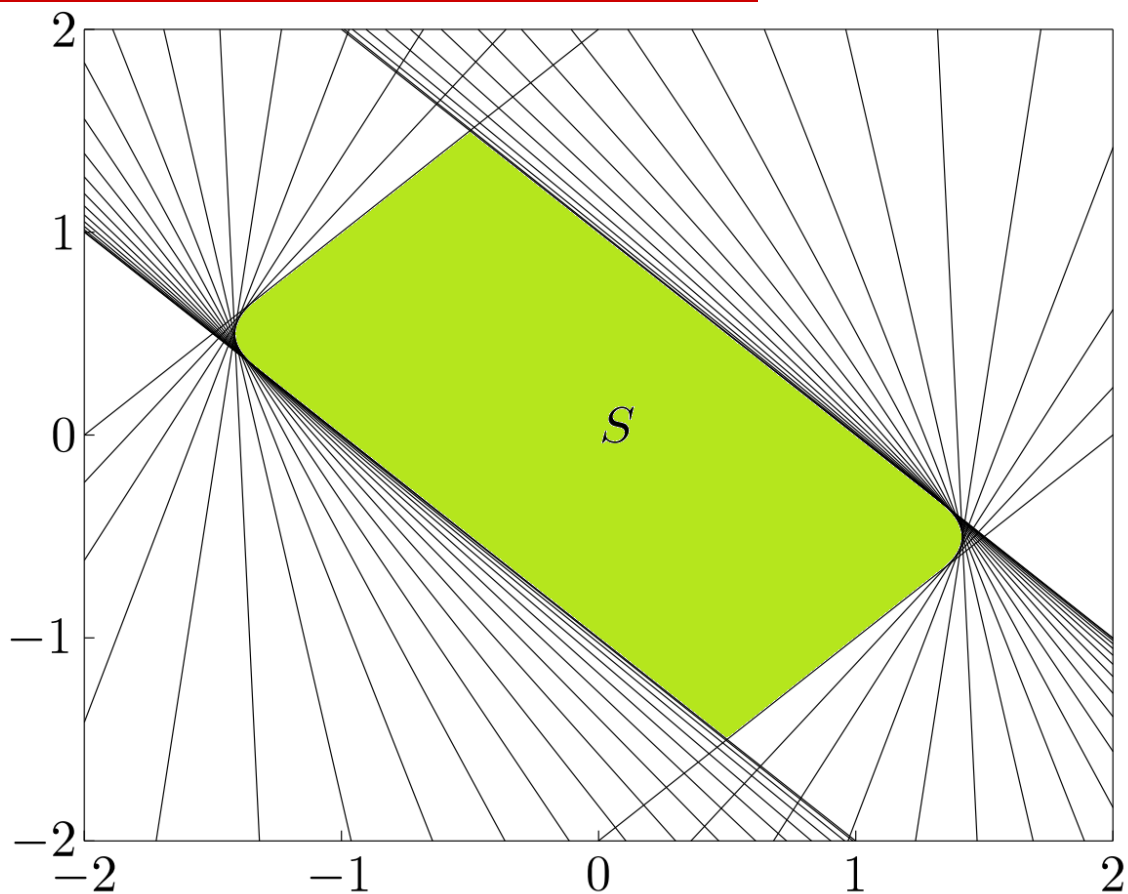
□ 仿射变换

■ 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

□ 投射变换(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, $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ $f(S) = \{f(x) \mid x \in S\}$
 - 若 S 为凸集, 则 $f(S)$ 为凸集;
 - 若 $f(S)$ 为凸集, 则 S 为凸集。

进一步分析仿射变换

□ 两个凸集的和为凸集

$$S_1 + S_2 = \{x + y \mid x \in S_1, y \in S_2\}$$

□ 两个凸集的笛卡尔积(直积)为凸集

$$S_1 \times S_2 = \{(x_1, x_2) \mid x_1 \in S_1, x_2 \in S_2\}$$

□ 两个集合的部分和为凸集(分配率)

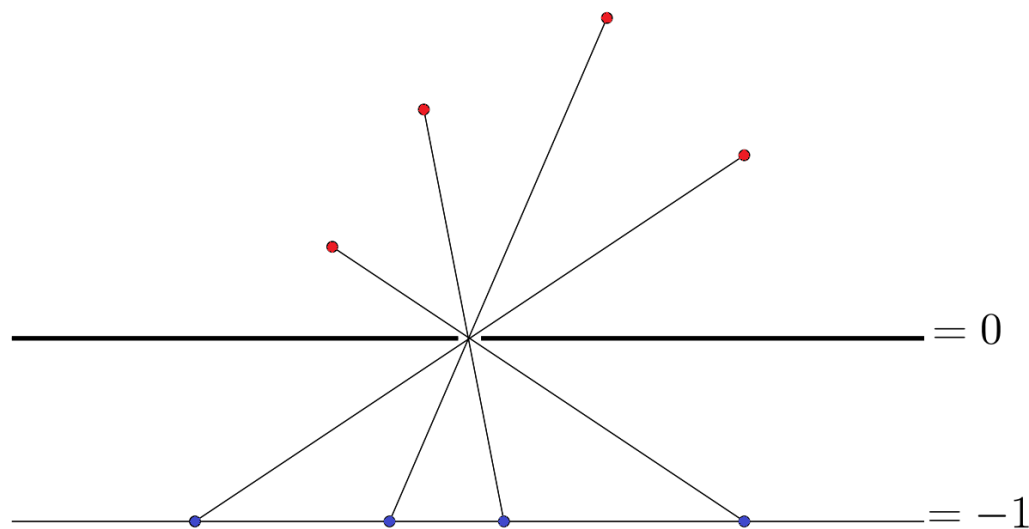
$$S = \{(x, y_1 + y_2) \mid (x, y_1) \in S_1, (x, y_2) \in S_2\}$$

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: $g: \mathbf{R}^n \rightarrow \mathbf{R}^{m+1}$

$$g(x) = \begin{bmatrix} A \\ c^T \end{bmatrix} x + \begin{bmatrix} b \\ d \end{bmatrix}$$

$$A \in \mathbf{R}^{m \times n}, b \in \mathbf{R}^m, c \in \mathbf{R}^n, d \in \mathbf{R}$$

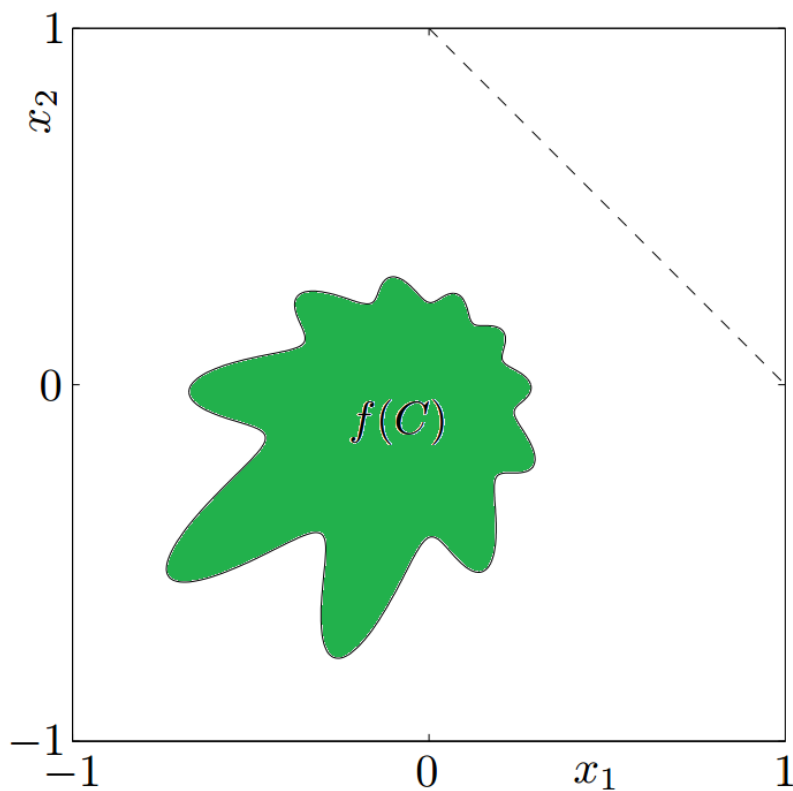
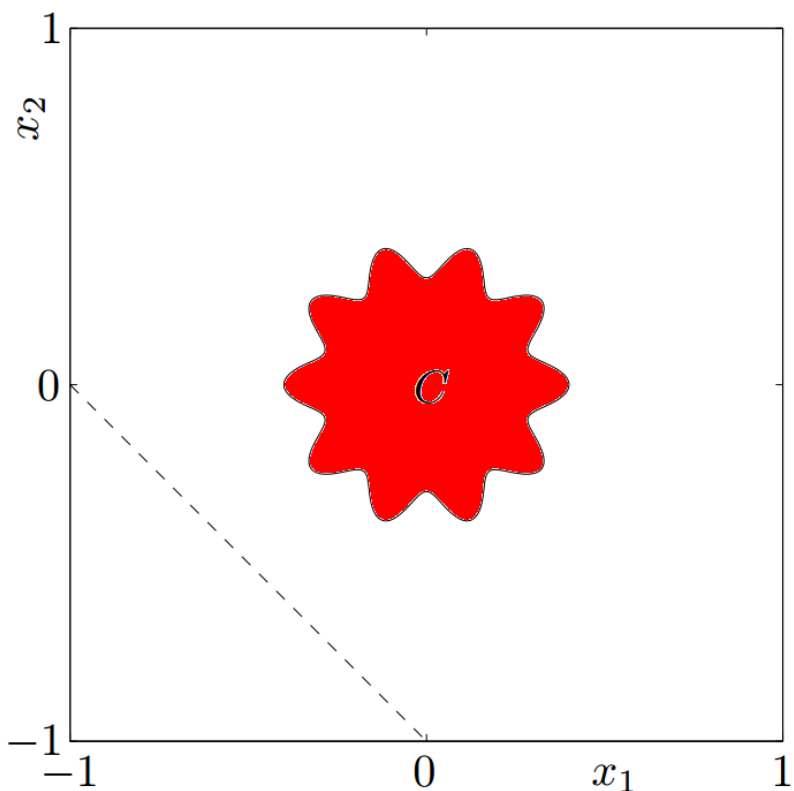
□ 定义 f 为线性分式函数

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

□ 若 $c=0, d>0$, 则 f 即为普通的仿射函数。

投射的作用

$$f(x_1, x_2) = \begin{cases} f_1(x_1) = \frac{x_1}{x_1 + x_2 + 1}, \\ f_2(x_2) = \frac{x_2}{x_1 + x_2 + 1} \end{cases}, \quad \text{dom } f = \{(x_1, x_2) \mid x_1 + x_2 + 1 > 0\}$$



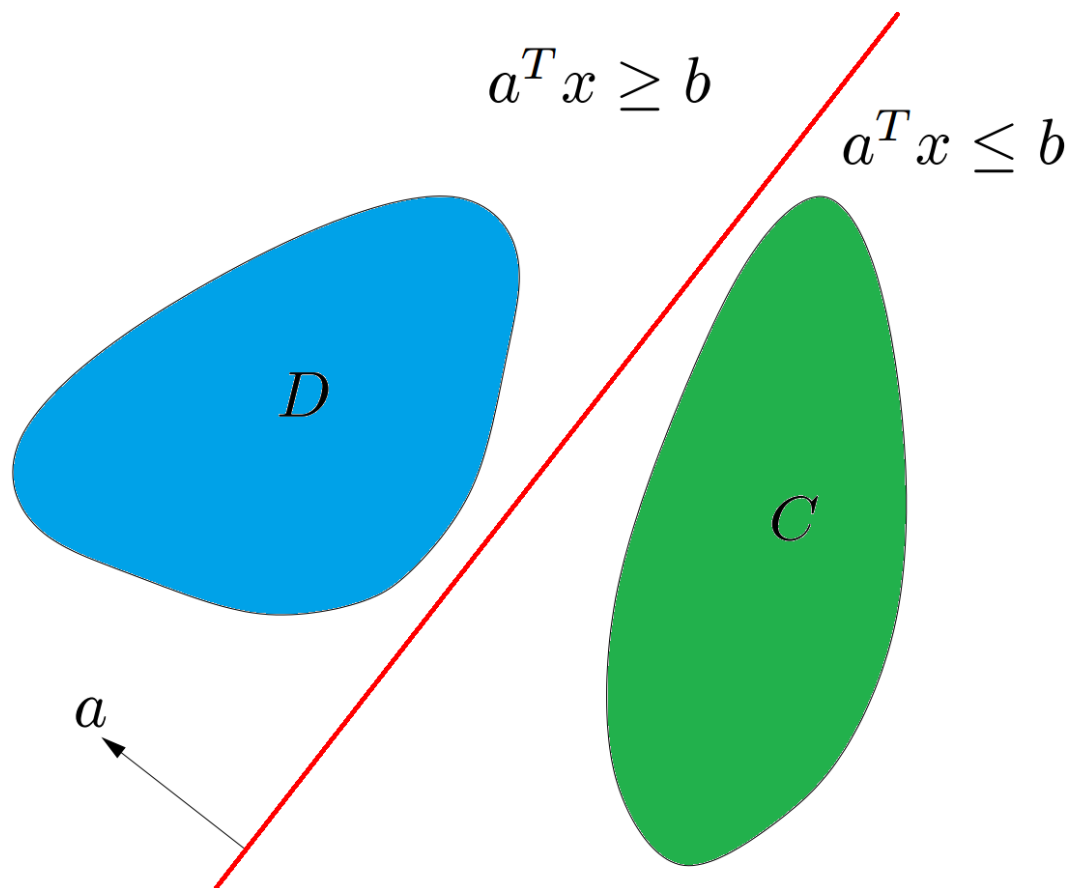
分割超平面

- 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。

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

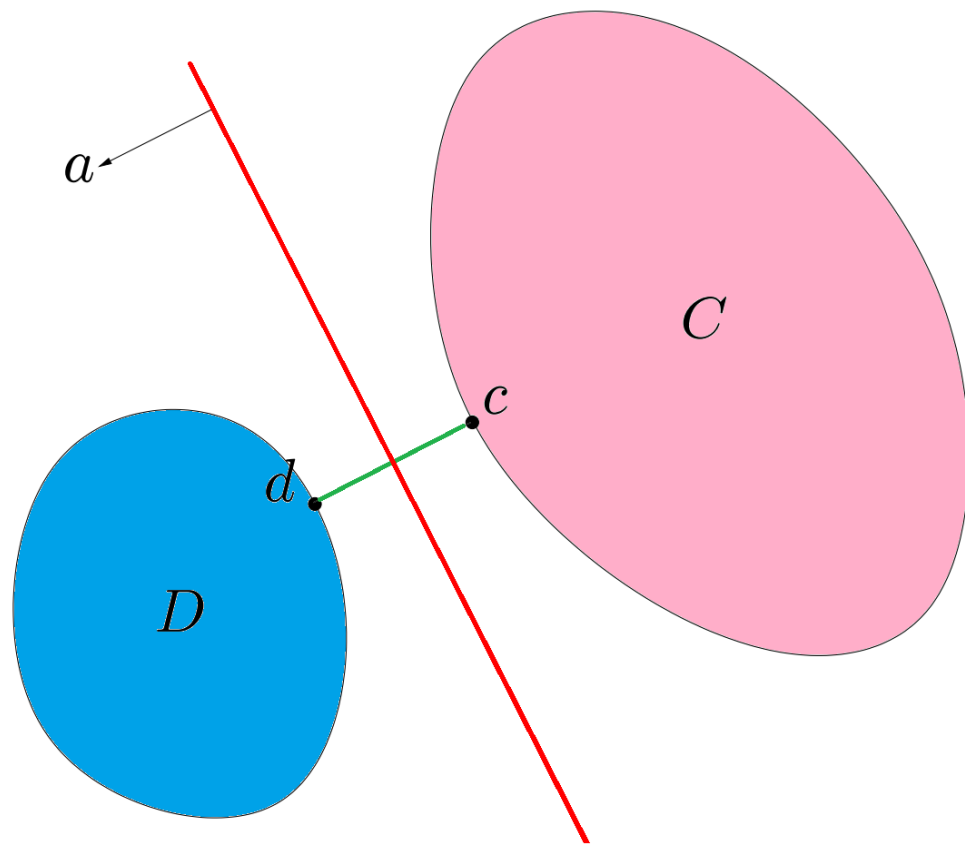
- 注意上式中可以取等号。
 - “若两个凸集C和D的分割超平面存在，C和D不相交”为假命题。
 - 加强条件：若两个凸集至少有一个是开集，那么当且仅当存在分割超平面，它们不相交。

分割超平面



分割超平面的构造

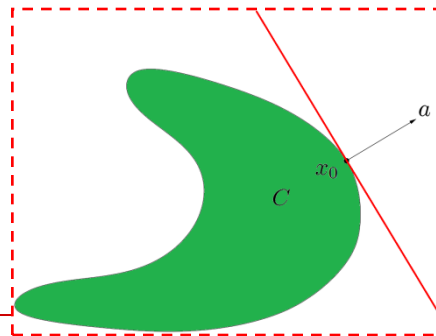
- 两个集合的距离，定义为两个集合间元素的最短距离。
- 做集合C和集合D最短线段的垂直平分线。



思考

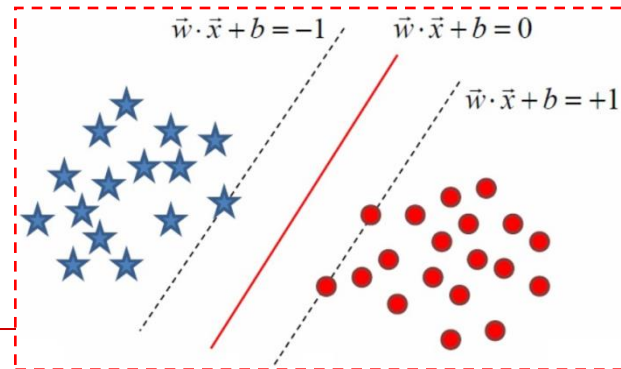
- 如何定义两个集合的“最优”分割超平面？
- 若两个集合有部分相交，如何定义超平面，使得两个集合“尽量”分开？
 - 上述“集合”的元素，可能是若干离散点。若一组集合为 $(x,1)$ ，另一组集合为 $(x,2)$ ，则为机器学习中的分类问题。

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 $a^T x \leq a^T x_0$ 成立，则称超平面 $\{x \mid a^T x = a^T x_0\}$ 为集合 C 在点 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。

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？

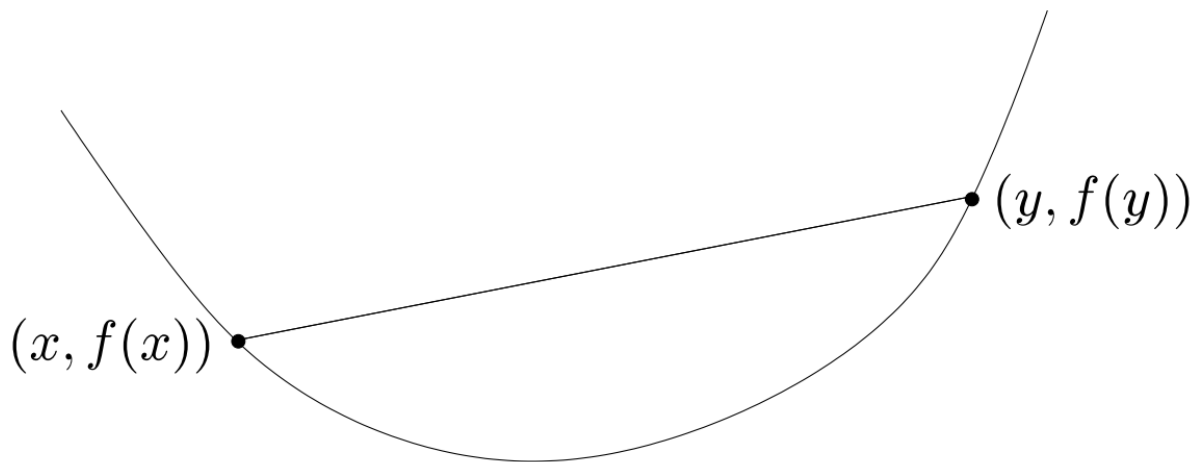
■ 注：上述“集合”不一定是凸集，可能是由若干离散点组成。若一组集合为 $(\mathbf{x}, 1)$ ，另一组集合为 $(\mathbf{x}, 2)$ ，则为机器学习中的分类问题。

凸函数

□ 若函数 f 的定义域 $\text{dom}f$ 为凸集，且满足

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

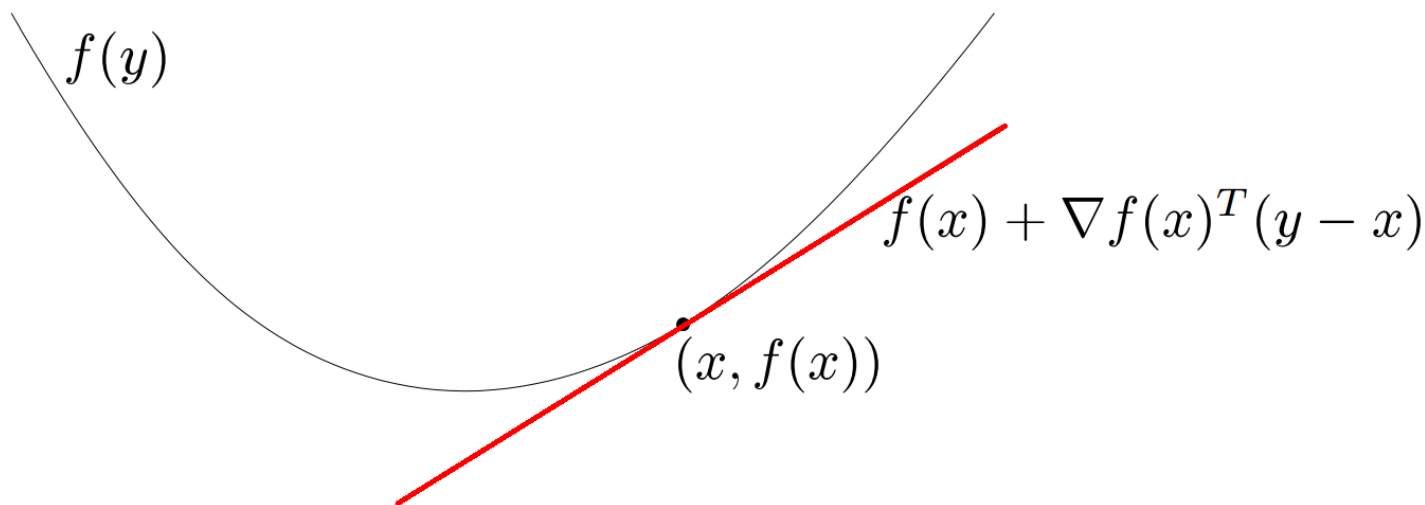
$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$$



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 $\text{dom}f$ 为凸集，且

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



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

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

- 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。

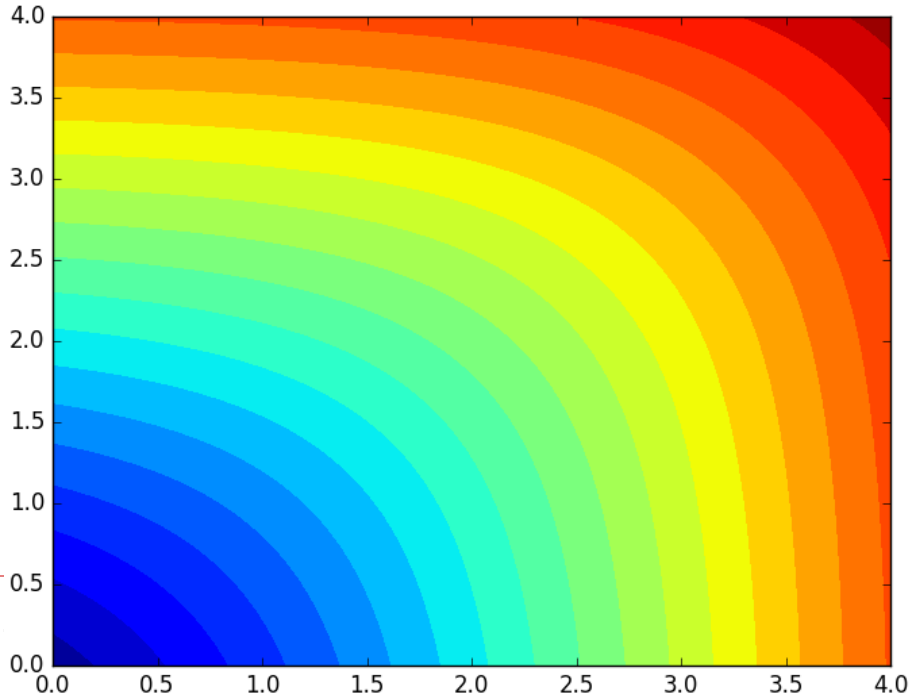
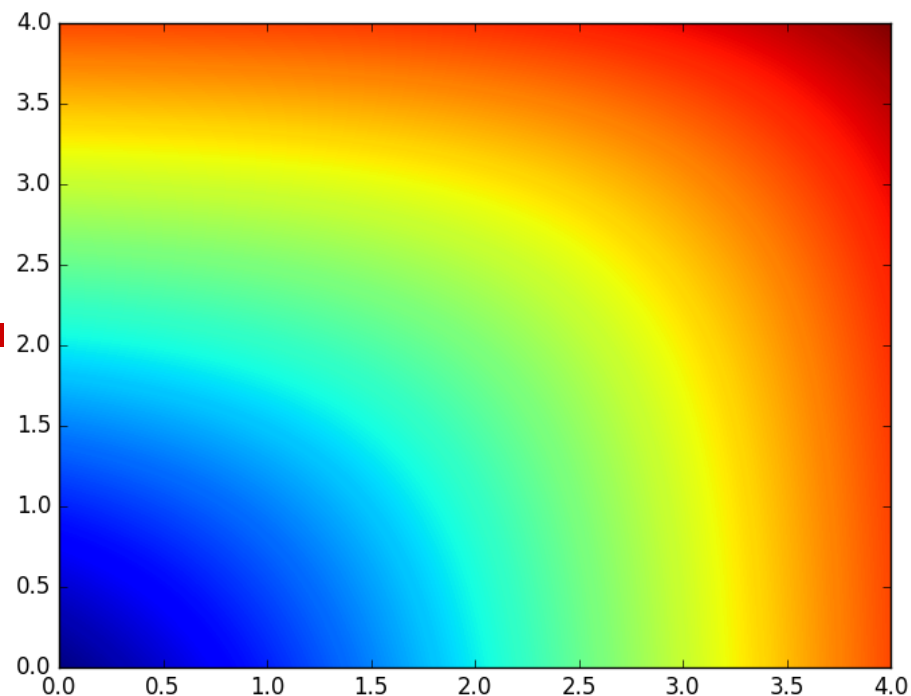
凸函数举例

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

Code

$$f(x_1, x_2) = \log(e^{x_1} + e^{x_2})$$

```
if __name__ == "__main__":  
    fig = plt.figure()  
    ax = fig.add_subplot(111)  
    u = np.linspace(0, 4, 1000)  
    x, y = np.meshgrid(u, u)  
    z = np.log(np.exp(x) + np.exp(y))  
    ax.contourf(x, y, z, 20)  
    plt.show()
```

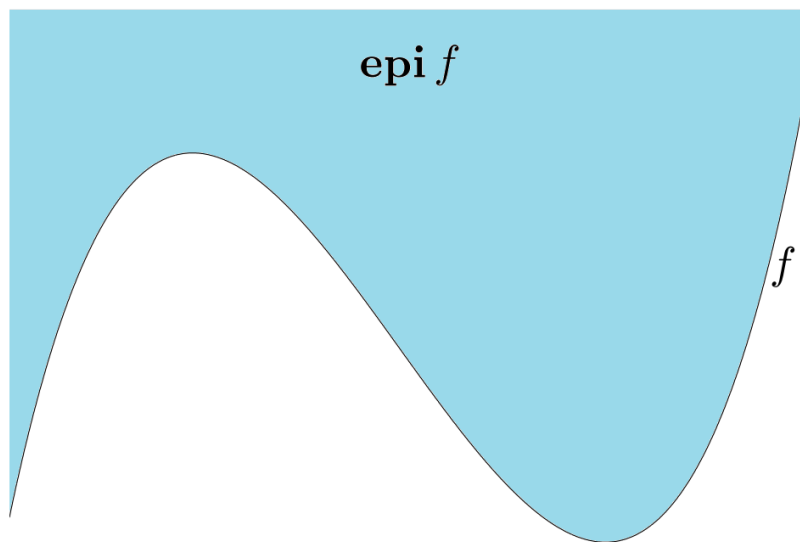


上境图

□ 函数 f 的图像定义为: $\{(x, f(x)) \mid x \in \text{dom } f\}$

□ 函数 f 的上境图(epigraph)定义为:

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



凸函数与凸集

- 某函数是凸函数，当且仅当其上图是凸集
 - 思考：如何证明？（提示：定义）
- 进一步，一个函数是凹函数，当且仅当其下图(hypograph)是凸集。

$$\text{hypo } f = \{(x, t) \mid t \leq f(x)\}$$

Jensen不等式：若f是凸函数

□ 基本Jensen不等式

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

□ 若 $\theta_1, \dots, \theta_k \geq 0, \theta_1 + \dots + \theta_k = 1$

□ 则 $f(\theta_1 x_1 + \dots + \theta_k x_k) \leq \theta_1 f(x_1) + \dots + \theta_k f(x_k)$

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

□ 则 $f\left(\int_S p(x)x dx\right) \leq \int_S f(x)p(x) dx$

$$f(\mathbf{E} x) \leq \mathbf{E} f(x)$$

Jensen不等式是几乎所有不等式的基础

□ 利用 $y=-\log x$ 是凸函数，证明：

$$\frac{a+b}{2} \geq \sqrt{ab}, \quad a > 0, b > 0$$

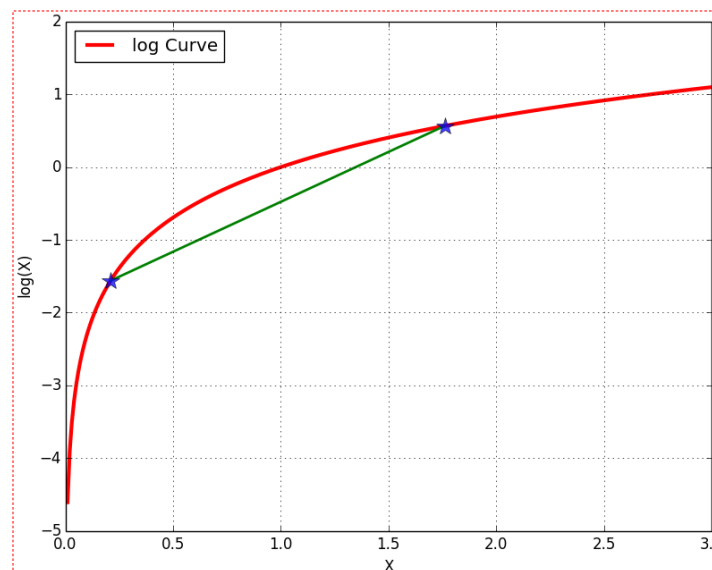
■ 提示：任取 $a, b > 0$ ， $\theta=0.5$ 带入基本Jensen不等式

□ 利用 $f(E(x)) \leq E(f(x))$ ，(f 是凸函数)，证明下式
 $D \geq 0$

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

注意到 $y=-\log x$ 在定义域上是凸函数

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



$$f(\mathbf{E} x) \leq \mathbf{E} f(x)$$

保持函数凸性的算子

□ 凸函数的非负加权和

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

□ 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

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

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

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

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

第二个不等号的证明

$$f_1(x) \leq \max\{f_1(x), f_2(x)\} \Rightarrow \\ \theta \cdot f_1(x) \leq \theta \cdot \max\{f_1(x), f_2(x)\} \dots\dots\dots \textcircled{1}$$

$$f_1(y) \leq \max\{f_1(y), f_2(y)\} \Rightarrow \\ (1-\theta) \cdot f_1(y) \leq (1-\theta) \cdot \max\{f_1(y), f_2(y)\} \dots\dots\dots \textcircled{2}$$

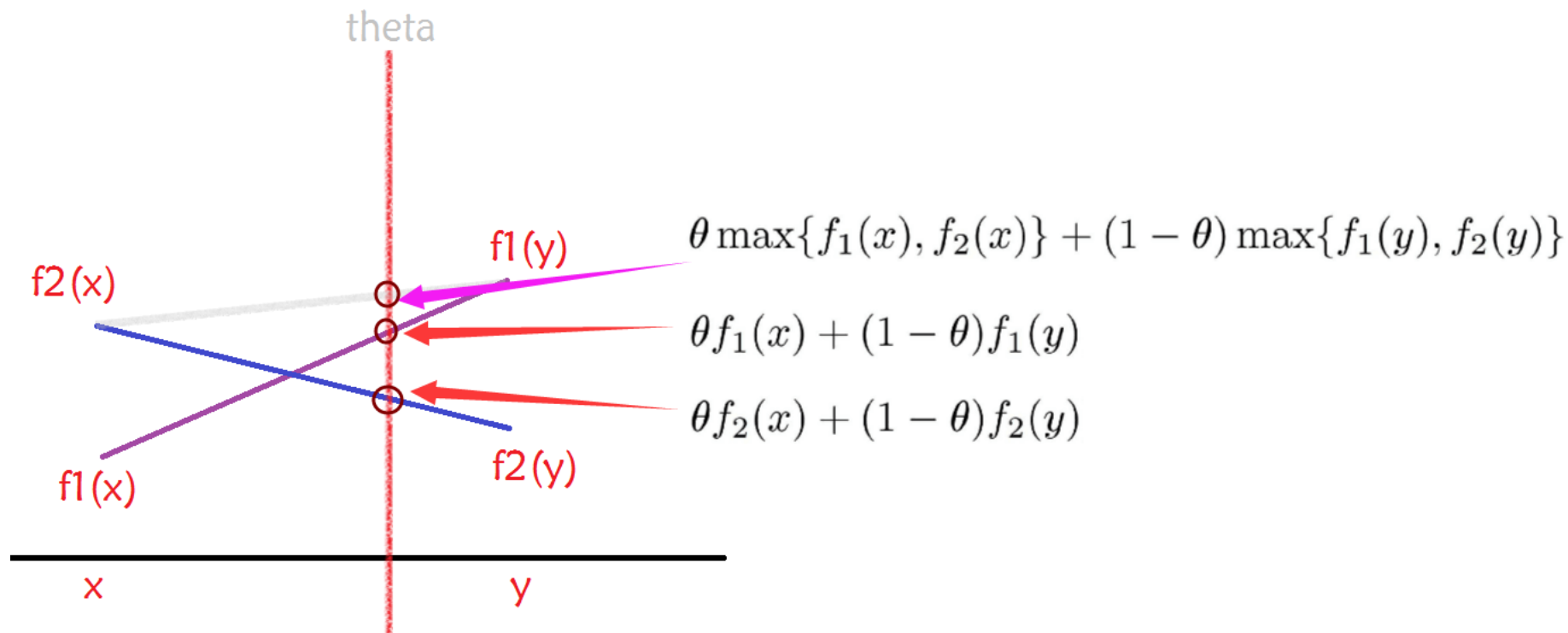
$$\textcircled{1} + \textcircled{2} \Rightarrow$$

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

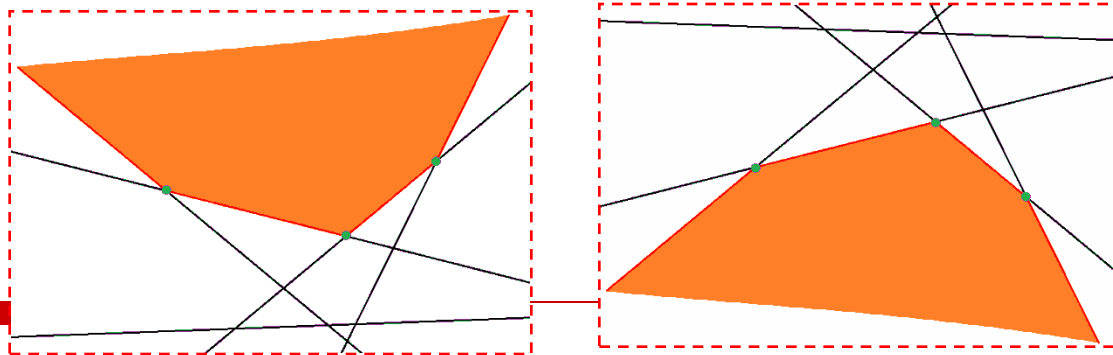
同理

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

第二个不等号的形式化表达



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。

共轭函数

□ 原函数 $f : \mathbf{R}^n \rightarrow \mathbf{R}$ 共轭函数定义：

$$f^*(y) = \sup_{x \in \text{dom } f} (y^T x - f(x))$$

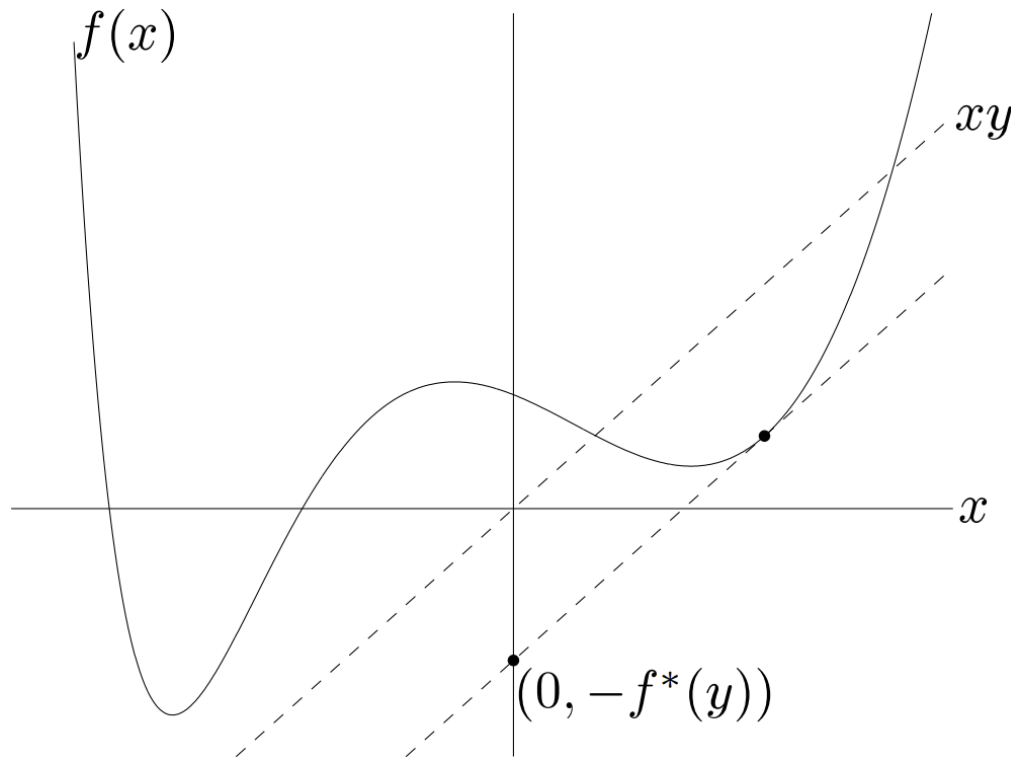
□ 显然，定义式的右端是关于 y 的仿射函数，它们逐点求上确界，得到的函数 $f^*(y)$ 一定是凸函数。

□ 该名称的原因：

■ 凸函数的共轭函数的共轭函数是其本身。

对共轭函数的理解 $f^*(y) = \sup_{x \in \text{dom } f} (y^T x - f(x))$

□ 如果函数 f 可微，在满足 $f'(x)=y$ 的点 x 处差值最大。



例：求共轭函数 $f^*(y) = \sup_{x \in \text{dom } f} (y^T x - f(x))$

- 可逆对称阵 Q ，对于任意的向量 x ，定义函数 f ： $f(x) = \frac{1}{2} x^T Q x$
- 关于 (x, y) 的函数 $y^T x - \frac{1}{2} x^T Q x$
- 在 $x = Q^{-1} y$ 时取上确界，带入，得到：

$$f^*(y) = \frac{1}{2} y^T Q^{-1} y$$

- f^* 即是 f 的共轭函数

Fenchel不等式

□ 根据定义

$$f^*(y) = \sup_{x \in \text{dom } f} (y^T x - f(x))$$

□ 立刻可以得到：

$$f(x) + f^*(y) \geq x^T y$$

Fenchel不等式的应用

□ 根据 $f(x)$ 及其共轭函数 $f^*(x)$

$$f(x) = \frac{1}{2} x^T Q x \quad f^*(y) = \frac{1}{2} y^T Q^{-1} y$$

□ 带入Fenchel不等式，得到：

$$x^T Q x + y^T Q^{-1} y \geq 2x^T y$$

凸优化

□ 优化问题的基本形式

$$\text{minimize } f_0(x), \quad x \in \mathbf{R}^n$$

$$\text{subject to } f_i(x) \leq 0, \quad i = 1, \dots, m$$

$$h_j(x) = 0, \quad j = 1, \dots, p$$

$$\text{优化变量} \quad x \in \mathbf{R}^n$$

$$\text{不等式约束} \quad f_i(x) \leq 0$$

$$\text{等式约束} \quad h_j(x) = 0.$$

$$\text{无约束优化} \quad m = p = 0$$

优化问题的基本形式

□ 优化问题的域

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

□ 可行点(解)(feasible)

■ $x \in D$, 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

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

□ 最优化解

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

凸优化问题的基本形式

minimize $f_0(x), x \in \mathbf{R}^n$

subject to $f_i(x) \leq 0, i = 1, \dots, m$

$h_j(x) = 0, j = 1, \dots, p$

□ 其中, $f_i(x)$ 为凸函数, $h_j(x)$ 为仿射函数

□ 凸优化问题的重要性质

■ 凸优化问题的可行域为凸集

■ 凸优化问题的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

$$\text{minimize } f_0(x), \quad x \in \mathbf{R}^n$$

$$\text{subject to } f_i(x) \leq 0, \quad i = 1, \dots, m$$

$$h_j(x) = 0, \quad j = 1, \dots, p$$

□ Lagrange函数

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

- 对固定的 x , Lagrange函数 $L(x, \lambda, \nu)$ 为关于 λ 和 ν 的仿射函数

Lagrange对偶函数(dual function)

□ Lagrange对偶函数

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

□ 若没有下确界，定义：

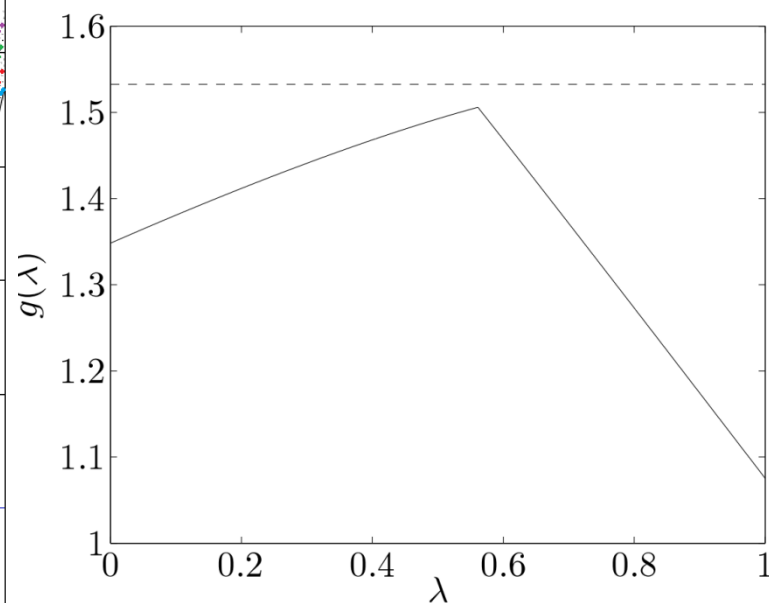
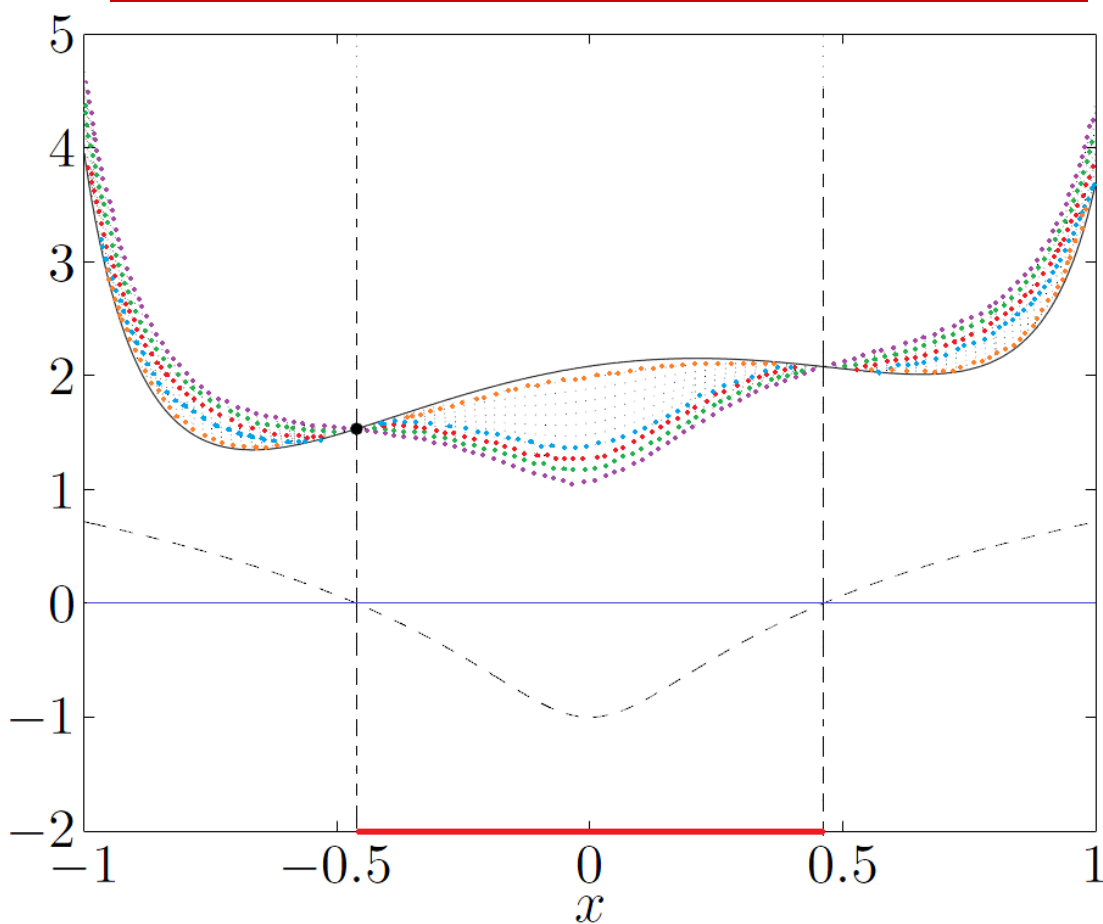
$$g(\lambda, \nu) = -\infty$$

□ 根据定义，显然有：对 $\forall \lambda > 0$ ， $\forall \nu$ ，若原优化问题有最优值 p^* ，则

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

□ 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 x_0 不可行，即存在某些 i ，使得 $f_i(x) > 0$ 。
则选择 $\lambda_i \rightarrow \infty$ ，对于其他乘子， $\lambda_j = 0, j \neq i$
- 假设 x_0 可行，则有 $f_i(x) \leq 0, (i=1, 2, \dots, m)$ ，选择

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

- 有：

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

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：

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

$$\text{证明: } \max_x \min_y f(x, y) \leq \min_y \max_x f(x, y)$$

□ 对于任意的 $(x, y) \in \text{dom} f$

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

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

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

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：

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

Karush-Kuhn-Tucker (KKT)条件

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

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

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

附：线性方程的最小二乘问题

□ 原问题 minimize $x^T x, x \in \mathbf{R}^n$

subject to $Ax = b$

□ 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 $L(x, \nu) = x^T x + \nu^T (Ax - b)$

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

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

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

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

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

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

$$\Rightarrow A^T v = -2(A^T A)^{-1} A^T b$$

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

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

极小值点 $x^* = (A^T A)^{-1} A^T b$

□ 极小值: $\min(x^T x)$

$$\begin{aligned} &= \left((A^T A)^{-1} A^T b \right)^T \left((A^T A)^{-1} A^T b \right) \\ &= b^T A (A^T A)^{-1} (A^T A)^{-1} A^T b \\ &= b^T A (A^T A)^{-2} A^T b \end{aligned}$$

□ 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。

参考文献

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

我们在这里

□ <http://wenda.ChinaHadoop.cn>

■ 视频/课程/社区

□ 微博

■ @ChinaHadoop

■ @邹博_机器学习

□ 微信公众号

■ 小象学院

■ 大数据分析挖掘



感谢大家！

恳请大家批评指正！