不止于代码

机器学习与优化方法简介

AI100学院
2017年8月

课程大纲

一．机器学习与优化

二．凸优化应用简介

三．凸优化理论初步：凸集与凸函数

课程目标与主题

- 目标：
  - 识别现实问题，并形式化为凸优化问题
  - 针对中等规模问题开发求解代码
- 主题：
  - 凸集、凸函数、优化问题
  - 示例与应用
  - 算法

▶凸优化简史

- 理论（凸分析）：1900年至1970年
- 算法：
  - 1947年：Dantzig提出单纯形算法
  - 1960年：早期的内点法
  - 1970年：椭圆法与次梯度法
  - 1980年：多项式时间内点法线性规划
  - 1980年至今：非线性凸优化的多项式时间内点法
- 应用：
  - 1990年前：主要用于运筹学，工程中很少用
  - 1990年后：工程领域应用涌现（控制、信息处理、通讯、电路设计）以及新的问题类型（半正定、二阶锥规划、鲁棒优化等）

▶机器学习中的优化问题

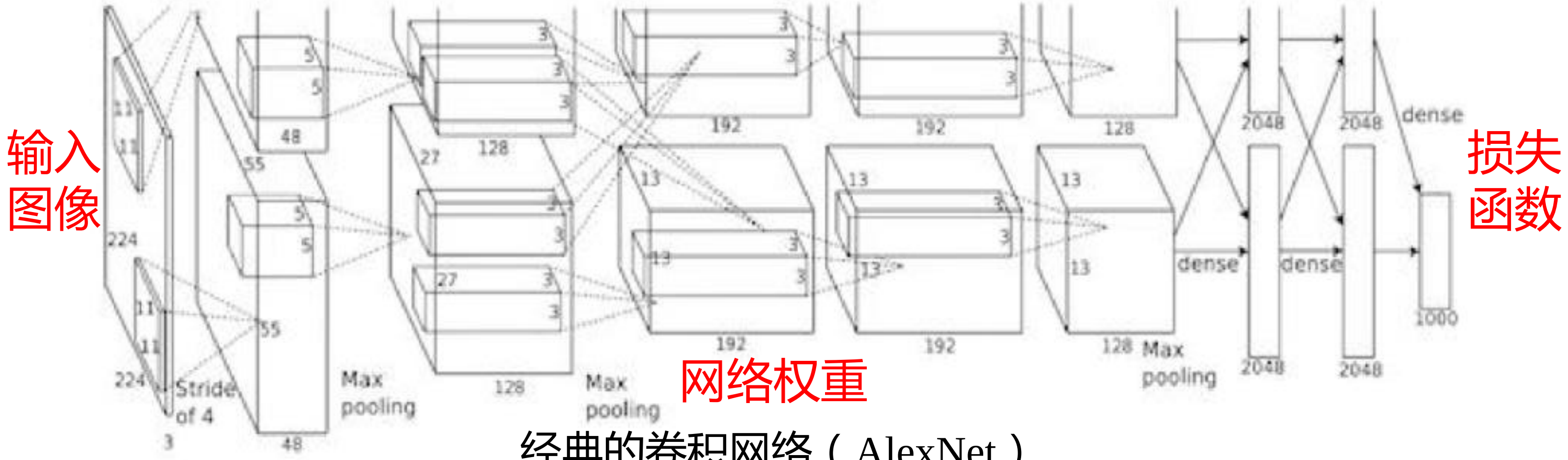
不止于代码

- 机器学习的很多问题都可以写为优化问题：

$$\min_{x\in\mathcal{X}} \sum_{i=1}^{N} \ell_i(x) + \lambda r(x)$$

- $\ell_i(x)$称为损失函数，度量模型与数据拟合程度
- $r(x)$称为正则函数，度量模型的复杂度，避免过拟合
- 实例：深度神经网络、支持向量机、压缩感知等

▶ 实例：深度神经网络



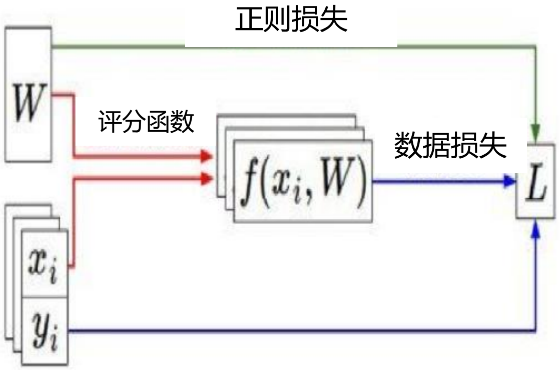
经典的卷积网络（AlexNet）

- **损失函数及其计算：**

Softmax

$$L_i = -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

$$L = \frac{1}{N}\sum_{i=1}^{N} L_i + R(W)$$



数学优化的形式化

- 一般的数学优化问题形式化如下：

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le b_i, \quad i = 1, \ldots, m \end{aligned}$$

- 其中：

$x = (x_1, x_2, \ldots, x_n)$ 为优化变量

$f_0 : \mathbb{R}^n \to \mathbb{R}$ 为目标函数

$f_i : \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots, m$ 为约束函数

- 最优解 x^* 达到目标函数最小值，并满足所有约束函数

优化的应用示例

- 投资组合优化：
  - 变量：不同资产的投资数额
  - 约束：预算、每个资产投资的上下限、最小回报等
  - 目标函数：整体风险或回报的方差
- 电子电路中元件尺寸
  - 变量：元件宽度与长度
  - 约束：制造极限、时序要求、最大面积等
  - 目标函数：功耗
- 数据拟合
  - 变量：模型参数
  - 约束：先验信息、参数范围
  - 目标函数：拟合误差或预测精度

优化问题求解

- 一般优化问题：
  - 非常难以求解
  - 求解方法涉及妥协：例如非常长的计算时间、或者并不总能找到最优解
- 特例：特定类型的问题可以高效可靠求解
  - 最小二乘问题
  - 线性规划问题
  - 凸优化问题

最小二乘

- 目标函数： $\min \quad \|Ax-b\|_2^2$
- 求解方法：
  - 数学解析解 $x^* = (A^TA)^{-1}A^Tb$
  - 可靠高效算法与软件，成熟技术
  - 计算复杂度正比与 n^2k $(A \in \mathbb{R}^{k\times n})$，稀疏更低
- 实际应用
  - 现实问题容易被识别为最小二乘，增加权重与正则化项提高灵活性

线性规划

- 目标函数：
$$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & a_i^T x \leq b_i, \quad i = 1, \ldots, m \end{aligned}$$
- 求解：
  - 没有数学解析解，具备可靠高效的算法与软件
  - 成熟的技术，计算复杂度正比与 $n^2 m$ (m > n)
- 实际应用：
  - 从现实问题转化为线性规划不是很直观
  - 可以使用一些标准技巧将问题转化为线性规划

凸优化问题

- 目标函数：

$$\begin{array}{ll} \min & f_0(x) \\ \text{s.t.} & f_i(x) \leq b_i, \quad i = 1, \dots, m \end{array}$$

- 其中目标函数与约束函数均为凸函数

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y)$$

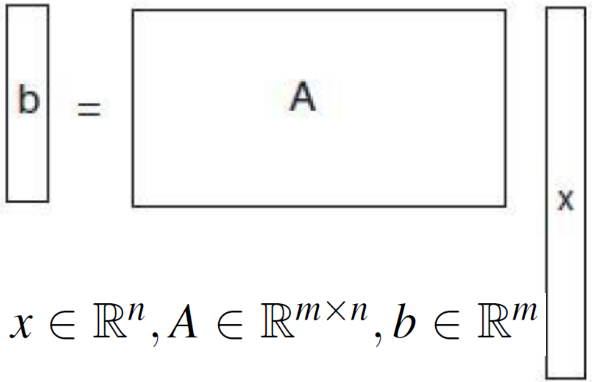
$$\alpha + \beta = 1, \alpha \geq 0, \beta \geq 0$$

- 最小二乘与线性规划均为凸优化的特例

凸优化的求解与应用

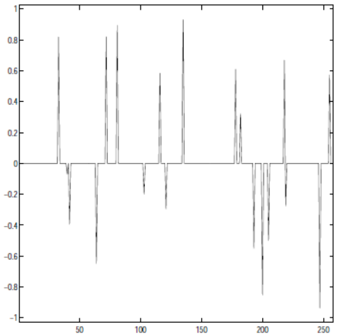
- 求解：
  - 没有数学解析解，具备可靠高效的算法，接近成熟的技术
  - 时间复杂度粗略近似于$\max\{n^3, n^2m, F\}$，其中F为求解目标函数与约束函数一、二阶导数的计算代价
- 应用：
  - 经常难以从现实问题中识别出凸优化，需要很多技巧转化为凸形式
  - 现实中的很多问题可以通过凸优化求解

►示例：解欠定线性方程组



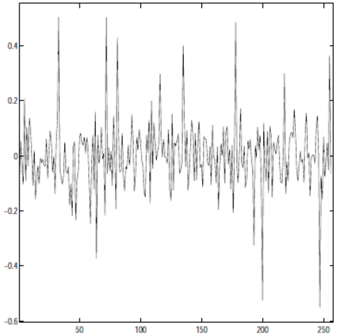
$$x \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$$

- 当 m << n时，线性方程Ax=b 没有唯一解。在指定目标函数后，可以恢复较准确的x, 类似问题在很多科学与工程领域出现



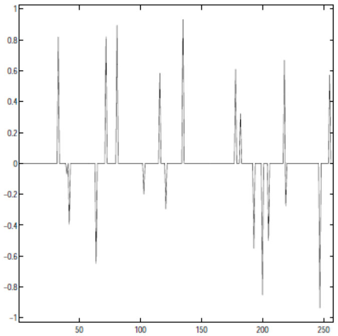
$$\begin{cases} \min\limits_{x} \|x\|_0 \\ \text{s.t. } Ax = b \end{cases}$$

(a) ℓ_0-minimization



$$\begin{cases} \min\limits_{x} \|x\|_2 \\ \text{s.t. } Ax = b \end{cases}$$

(b) ℓ_2-minimization



$$\begin{cases} \min\limits_{x} \|x\|_1 \\ \text{s.t. } Ax = b \end{cases}$$

(c) ℓ_1-minimization

设定目标函数为不同的范数，恢复最稀疏解（10%非零）

```
n=256, m=128;
A = randn(m,n);
u = sprandn(n, 1, 0.1);
b = A*u;
```

课程大纲

一．机器学习与优化

二．凸优化应用简介

三．凸优化理论初步：凸集与凸函数

▶压缩感知

- 给定(A, b, Ψ)，寻找最稀疏解：

$$x^* = \arg\min\{\|\Psi x\|_0 : Ax = b\}$$

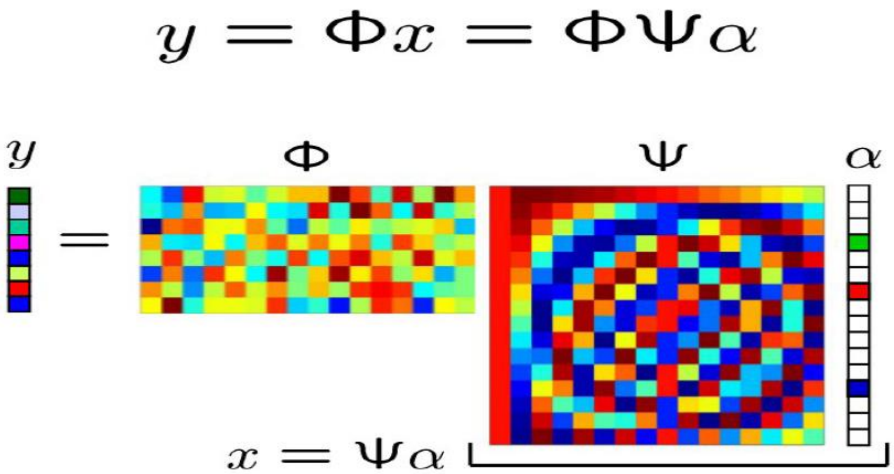
- 从组合优化松弛为凸优化：

$$\bar{x} = \arg\min\{\|\Psi x\|_1 : Ax = b\}$$

- 一范数提升稀疏性，Donoho在1998年提出基追踪的求解方法，多种变形针对b含有噪音
- 理论证明：零范数与一范数等价性（Candes与Tao 2005）

压缩感知

- 从字典Ψ中稀疏元素合成向量x，因此向量α具有稀疏性
- 字典可以由DCT、小波、Gabor及其组合构成，也可以从训练数据或部分信号中学习得到



▶小波近似

- 2.5%系数近似恢复原图像

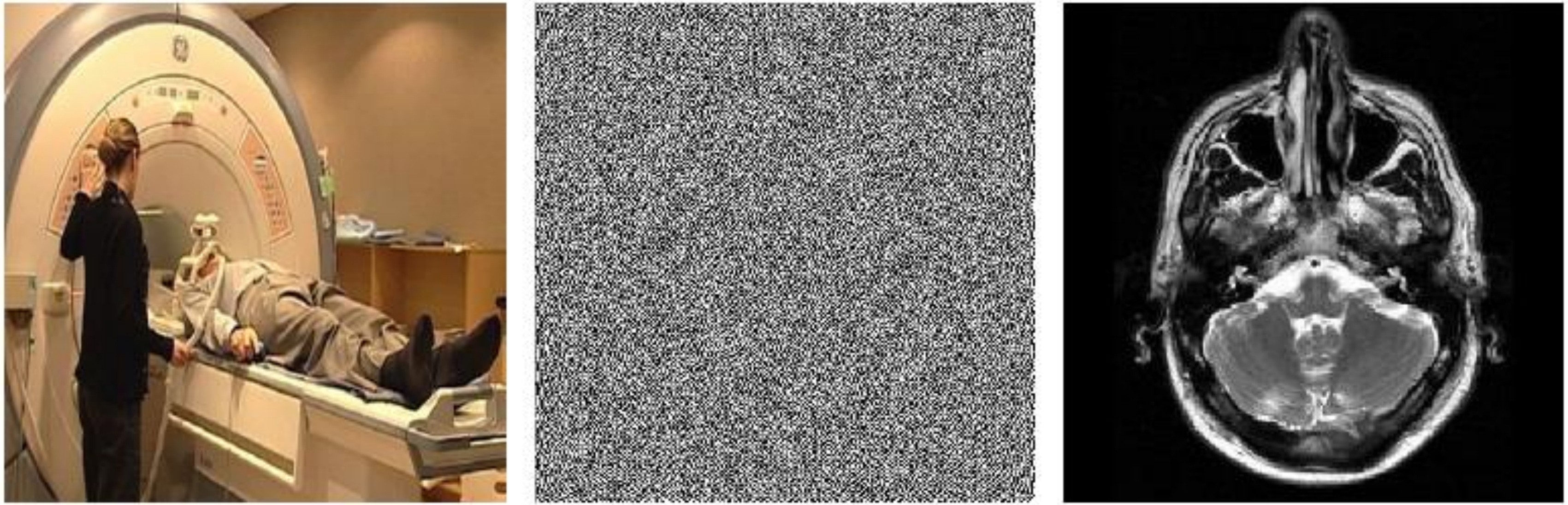
1 megapixel image

25k term approx

$$\|f - f_K\|_2 \approx .01 \cdot \|f\|_2$$

其中f为原始图像，f_K为近似图像

▶ 核磁共振成像



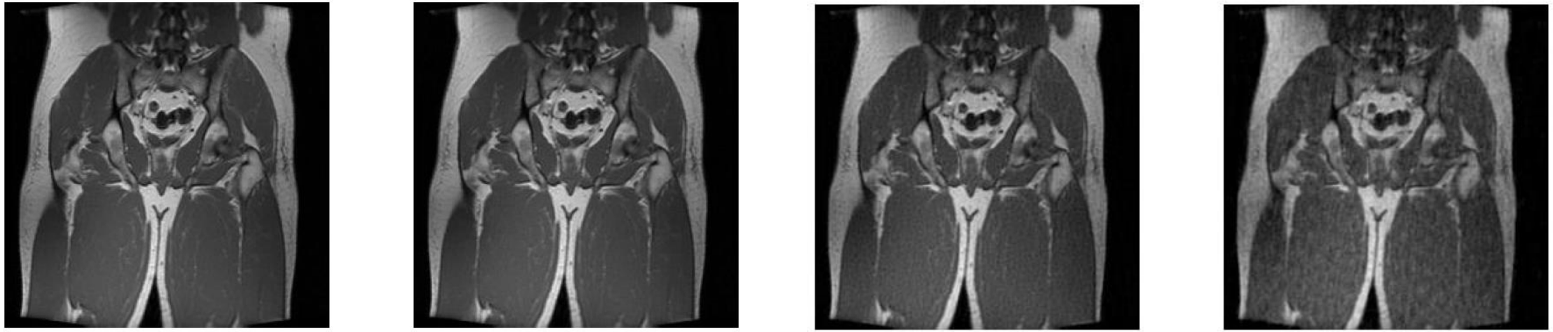
(a) MRI Scan　(b) Fourier Coefficients　(c) Image

能否通过凸优化方法减少一半扫描时间？

核磁共振成像

- MRI图像在一些小波变换Φ下，经常具备稀疏表示的特性，求解如下公式重建MRI图像：

$$\min_{u} \|\Phi u\|_1 + \frac{\mu}{2}\|Ru - b\|^2$$



(a) full sampling

(b) 39% sampling, SNR=32.2

(c) 22% sampling, SNR=21.4

(d) 14% sampling, SNR=15.8

信噪比SNR越高，图像质量越好

▶ Netflix电影推荐竞赛

- Netflix数据库：百万用户，25000部电影
- 用户提供电影评分，矩阵非常稀疏
- 挑战：补齐Netflix矩阵（百万美元奖励）

▶矩阵秩最小化

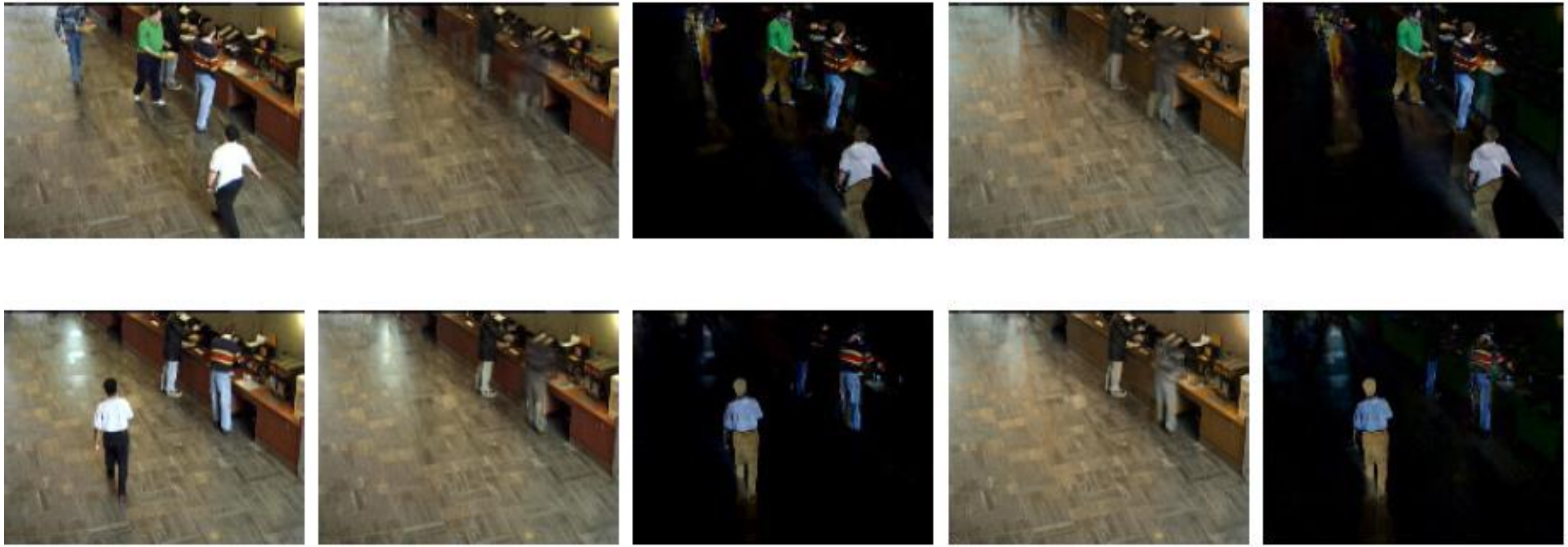
不 止 于 代 码

- 给定 $X \in \mathbb{R}^{m\times n}, \mathcal{A}: \mathbb{R}^{m\times n} \to \mathbb{R}^p, b \in \mathbb{R}^p$，我们考虑：
  - 矩阵补齐问题： $\min\ \operatorname{rank}(X),\ \text{s.t. } X_{ij} = M_{ij}, (i,j) \in \Omega$
  - 矩阵秩最小化问题： $\min\ \operatorname{rank}(X),\ \text{s.t. } \mathcal{A}(X) = b$
  - 核范数最小： $\min\ \|X\|_*\ \text{ s.t. } \mathcal{A}(X) = b$
- 其中核范数为矩阵X奇异值σ_i的和，

$$\|X\|_* = \sum_i \sigma_i$$

视频分割

- 将视频分割为运动（前景）与静止（背景）部分



给定矩阵M，找到低秩矩阵W与稀疏矩阵E，使得M = W + E。
凸优化形式（鲁棒主元分析）：

$$\min_{W,E} \|W\|_* + \mu\|E\|_1, \text{ s.t. } W + E = M$$

投资组合优化

- r_i，随机变量表示股票 i 的回报率
- x_i, 投资股票 i 的相对份额
- 回报： $r = r_1x_1 + r_2x_2 + \ldots + r_nx_n$
- 风险： $V = Var(r) = \sum_{i,j}\sigma_{ij}x_ix_j = x^\top\Sigma x$
- 凸优化思路：确保最小回报并最小化风险

$$\min \frac{1}{2}x^\top\Sigma x \quad \text{s.t.} \quad \sum \mu_i x_i \geq r_0, \quad \sum x_i = 1, \quad x_i \geq 0$$

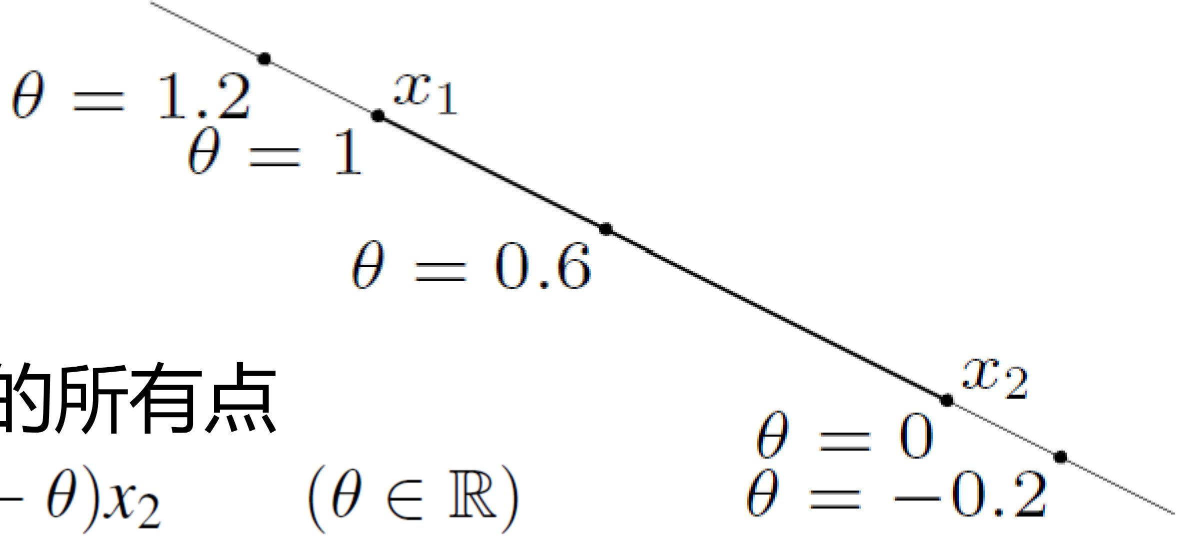
课程大纲

一．机器学习与优化

二．凸优化应用简介

三．凸优化理论初步：凸集与凸函数

▶ 凸集



- 仿射集：
  - 通过x_1,x_2两点的所有点

    $$x = \theta x_1 + (1-\theta)x_2 \qquad (\theta \in \mathbb{R})$$

  - 例子：

    线性方程组的解集：$\{x \mid Ax = b\}$

    （每个仿射集都可以表示为线性方程组的解集）

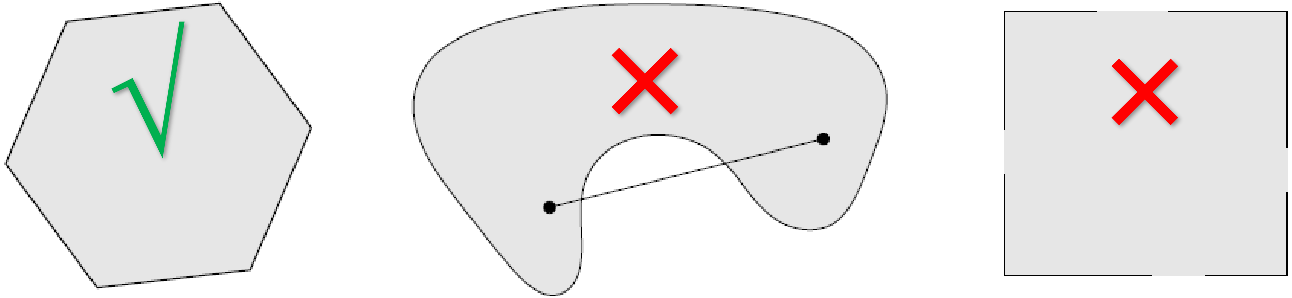
▶ 凸集及示例

- 两点x_1,x_2之间的线段上所有点：

$$x = \theta x_1 + (1-\theta)x_2 \qquad 0 \le \theta \le 1$$

- 凸集：
  – 包含集合中任意两点构成的线段

$$x_1, x_2 \in C, \quad 0 \le \theta \le 1 \quad \Longrightarrow \quad \theta x_1 + (1-\theta)x_2 \in C$$



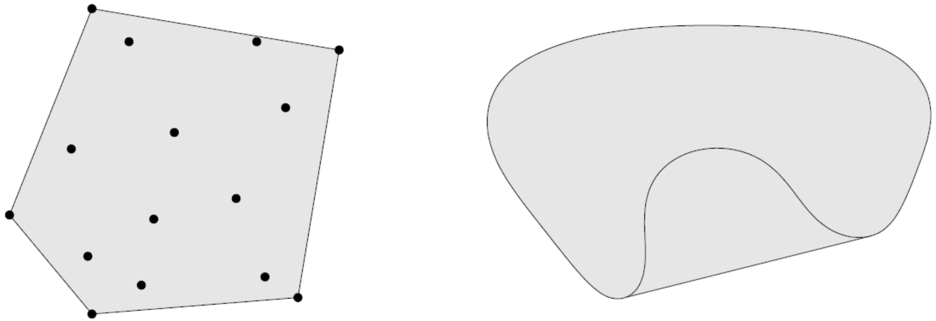
凸组合与凸包

- 凸组合：$x_1,\ldots,x_k$的凸组合表示为：

$$x = \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_k x_k$$

其中 $\theta_1 + \ldots + \theta_k = 1, \theta_i \geq 0$

- 凸包：所有点的凸组合集合



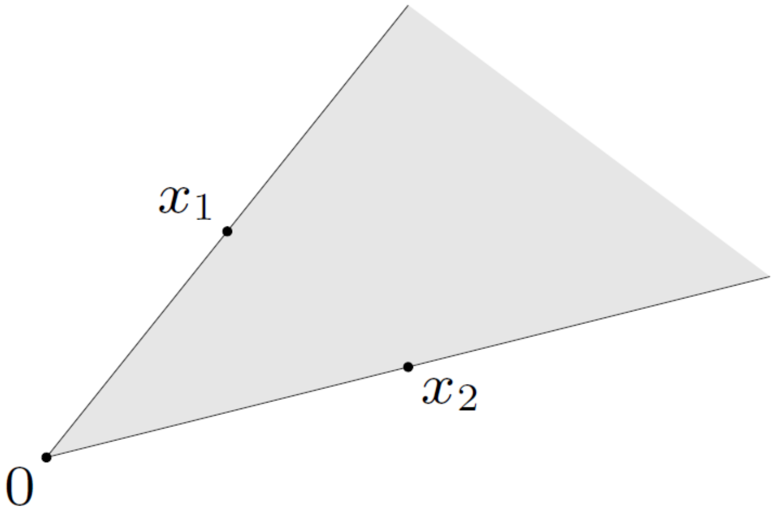
▶凸锥

- 两点x_1,x_2的非负组合：

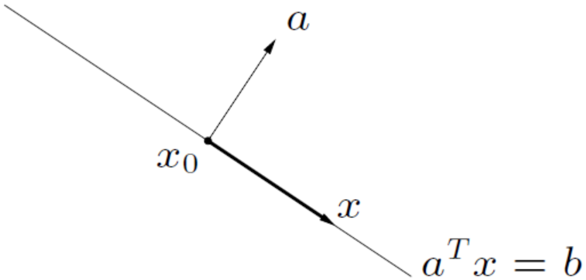
$$x = \theta_1 x_1 + \theta_2 x_2$$

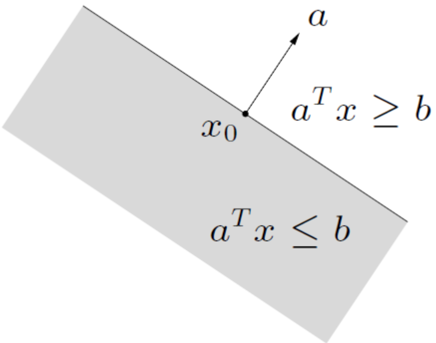
其中 $\theta_1 \geq 0,\ \theta_2 \geq 0$

- 凸锥：
  - 两点所有非负组合的集合



▶超平面与半空间





超平面：形如 $\{x|a^Tx = b\}(a \neq 0)$的集合　**半空间**：形如 $\{x|a^Tx \leq b\}(a \neq 0)$ 的集合

- a 为法线向量
- 超平面为仿射集及凸集
- 半空间为凸集

欧式球与椭圆

- 中心在 x_c ，半径为 r 的欧式球表示为：

$$B(x_c, r) = \{x \mid \|x - x_c\|_2 \le r\} = \{x_c + ru \mid \|u\|_2 \le 1\}$$

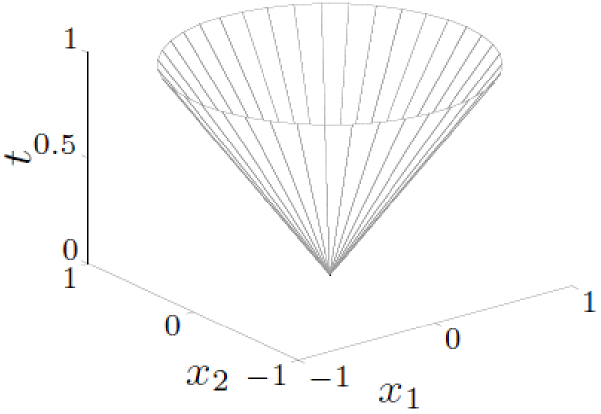
- 椭圆可以表示为：

$$\{x \mid (x - x_c)^T P^{-1} (x - x_c) \le 1\}$$

其中P为正定矩阵

▶ 范数球与范数锥

- 范数的定义：
  1. $\|x\| \geq 0$；$\|x\|=0$ 仅当 $x = 0$（非负）
  2. $\|t\,x\| = |t|\,\|x\|$ $t \in R$（线性）
  3. $\|x + y\| \leq \|x\| + \|y\|$（三角不等式）



范数球的中心为 x_c，半径为 r

$$\{x \mid \|x - x_c\| \leq r\}$$

范数锥：$\{(x, t) \mid \|x\| \leq t\}$

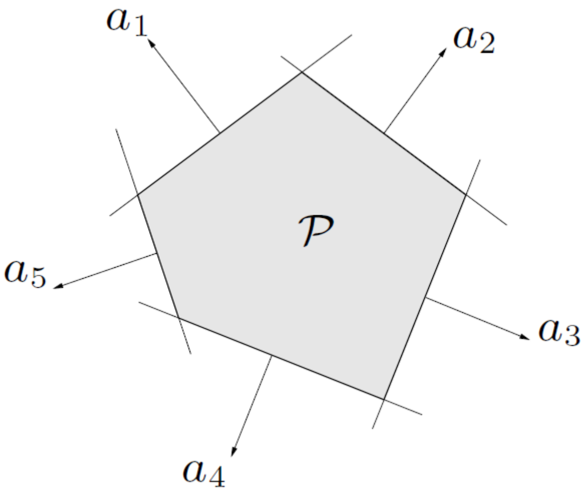
注：欧氏范数锥被称为二阶锥，范数球与范数锥都是凸集

▶ 多面体

- 有限多个线性不等式与等式的解

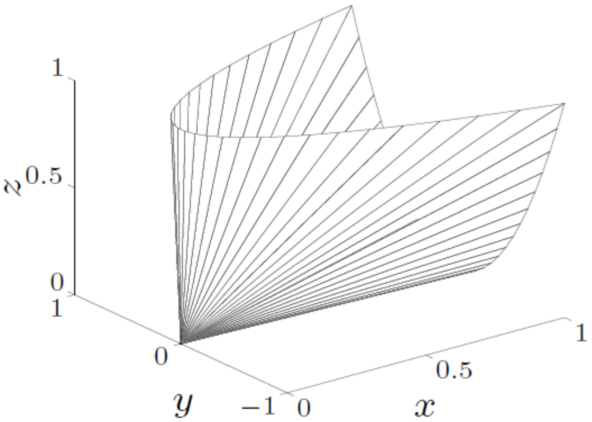
$$Ax \preceq b, \qquad Cx = d$$

- 其中≤为逐元素不等 $A \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{p\times n}$
- 多面体为有限个半空间和超平面的交集



半正定锥

- 对称、半正定、正定矩阵定义：
  - $\mathbb{S}^n$为对称 $n \times n$矩阵的集合
  - $\mathbb{S}_+^n = \{X \in \mathbb{S}^n | X \succeq 0\}$ 表示半正定$n \times n$ 矩阵集合
  - 等价定义，对于所有z, $X \in \mathbb{S}_+^n \iff z^T X z \geq 0$
  - $\mathbb{S}_{++}^n = \{X \in \mathbb{S}^n | X \succ 0\}$ 表示 $n \times n$ 正定矩阵集合



$$\begin{bmatrix} x & y \\ y & z \end{bmatrix} \in \mathbb{S}_+^2$$

▶ 保持凸性的运算

- 建立凸集C的两类实用方法：
  - 定义法：

$$x_1, x_2 \in C, \quad 0 \le \theta \le 1 \quad \Longrightarrow \quad \theta x_1 + (1-\theta) x_2 \in C$$

  - 构造法：在简单凸集上，通过保凸性运算获得
    - 交集
    - 仿射函数
    - 透视函数
    - 线性—分数函数

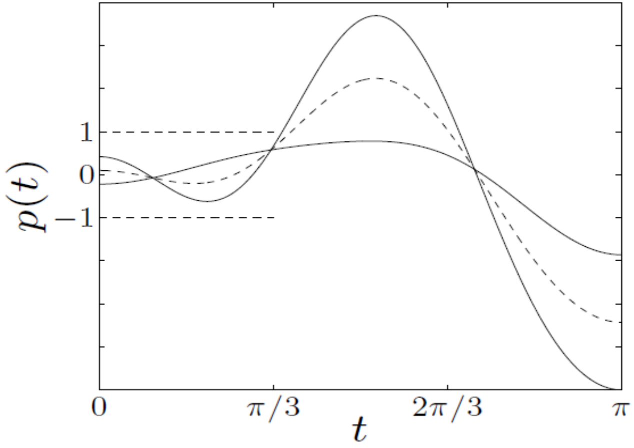
▶ 交集

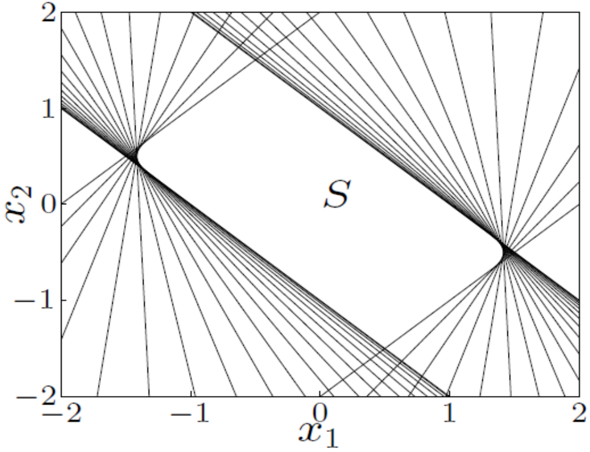
- 任意数量凸集的交集仍为凸集

示例：$S = \{x \in \mathbb{R}^m \mid |p(t)| \leq 1 \text{ for } |t| \leq \pi/3\}$

其中 $p(t) = x_1 \cos t + x_2 \cos 2t + \ldots + x_m \cos mt$

当m = 2时，如下图所示：





▶ 仿射函数

- 仿射函数定义：
  - 函数 $f:\mathbb{R}^n \rightarrow \mathbb{R}^m$ 为仿射，即
    $f(x) = Ax + b$ 并且 $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$
- 性质：
  - 凸集在仿射函数及其反函数的投影仍为凸集
- 示例：
  - 缩放、平移、投影操作保持凸性
  - 线性矩阵不等式的解集为凸集

$$\{x | x_1 A_1 + \ldots + x_m A_m \preceq B\}$$

透视与线性—分数函数

- 透视函数定义 P : $\mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$

$$P(x,t) = x/t, \quad \mathrm{dom}\, P = \{(x,t) | t > 0\}$$

- 性质 ：
  - 凸集在仿射变换下的正逆变换均为凸集
- 线性分数函数 f : $\mathbb{R}^n \rightarrow \mathbb{R}^m$

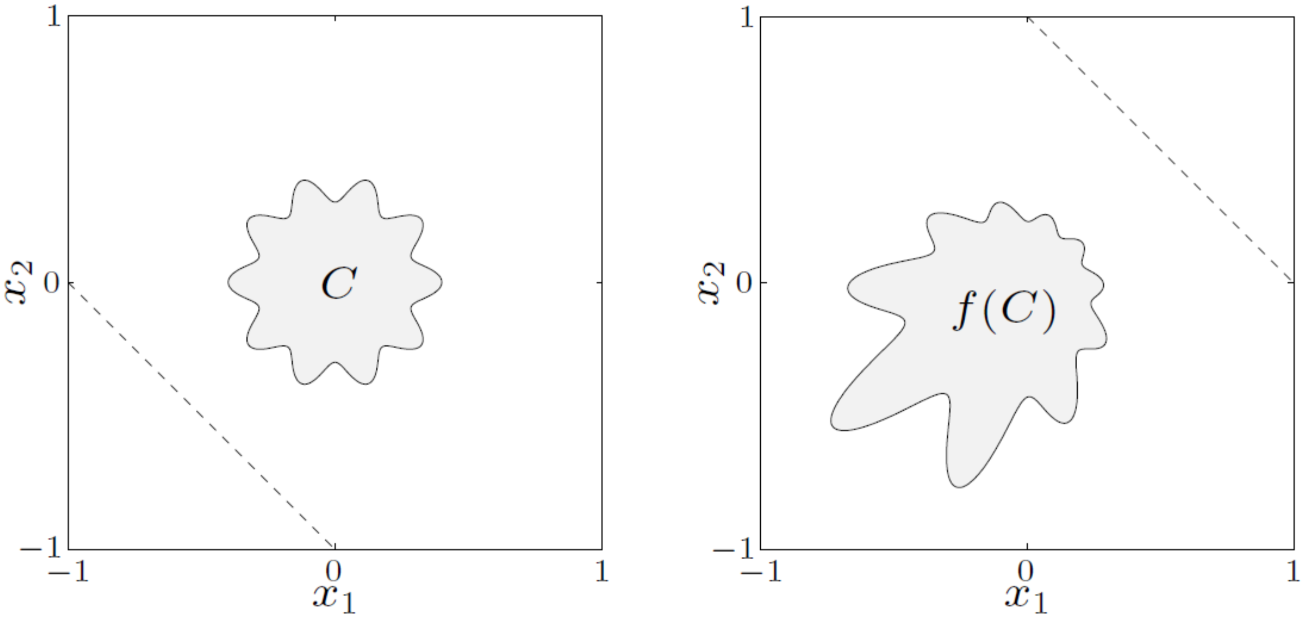
$$f(x) = \frac{Ax+b}{c^T x + d}, \quad \mathrm{dom}\, f = \{x | c^T x + d > 0\}$$

- 性质：凸集在线性-分数变换下的正逆变换仍为凸

▶示例：线性-分数函数

不 止 于 代 码

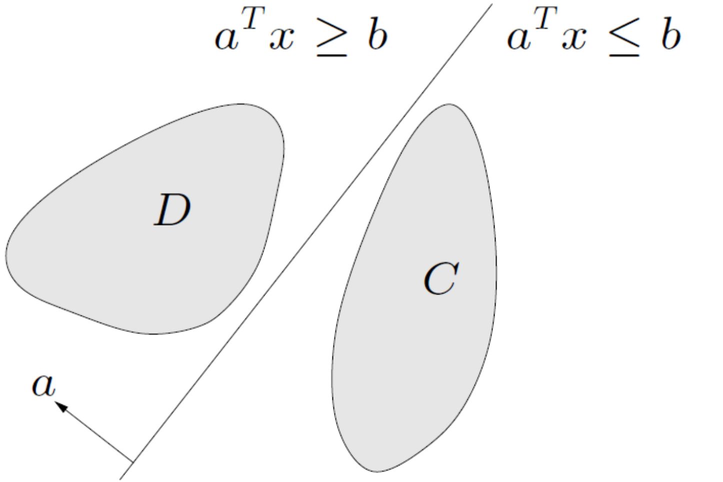
$$f(x) = \frac{1}{x_1 + x_2 + 1} x$$



分割超平面定理

- 如果C与D为不相交的凸集，则存在a $\neq 0$, b 满足：

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



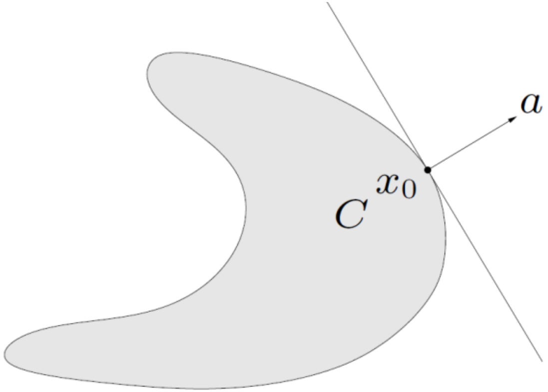
- 超平面 $\{x|a^T x = b\}$ 将凸集C和D分离

支持超平面定理

- 定理：若C为凸集，则每一个边界点处都存在一个支持超平面
- 凸集C边界点x_0处支持超平面表示为：

$$\{x|a^T x = a^T x_0\}$$

其中 $a \neq 0$ 且 $a^T x \leq a^T x_0$ 所有 $x \in C$

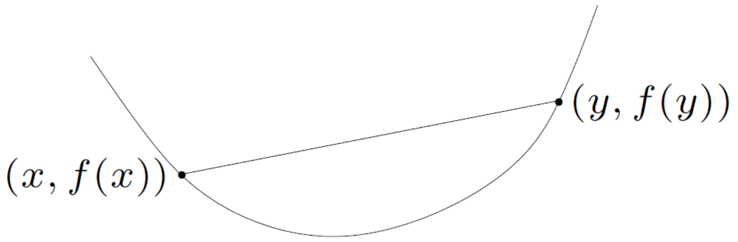


凸函数

- 定义：
  - 函数 $f : \mathbb{R}^n \to \mathbb{R}$ 为凸，要求其定义域为凸集且

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

  - 满足 $x, y \in \mathrm{dom}\, f$，$0 \leq \theta \leq 1$



  - 如果 –f 为凸，则 f 为凹函数
  - 严格凸函数则将上述 $\leq$ 替换为 $<$

▶示例：实数域凹凸函数

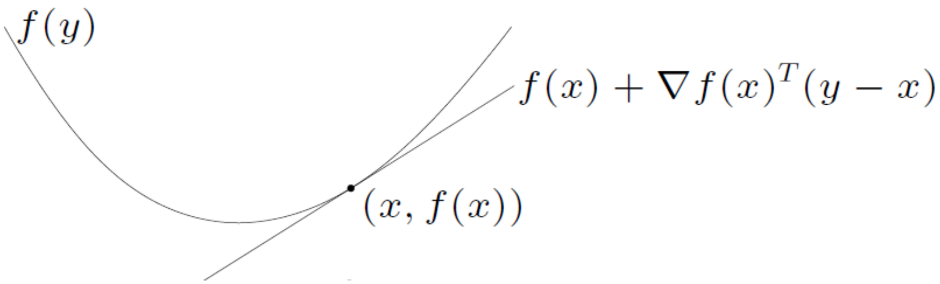
- 凸函数：
  - 仿射函数：$ax + b$ on $\mathbb{R}$, for any $a, b \in \mathbb{R}$
  - 指数函数：e^{ax}, for any $a \in \mathbb{R}$
  - 幂函数： x^{α} on $\mathbb{R}_{++}$, for $\alpha \geq 1$ or $\alpha \leq 0$
  - 负熵函数：$x \log x$ on $\mathbb{R}_{++}$
- 凹函数：
  - 对数函数：$\log x$ on $\mathbb{R}_{++}$

▶ 凸函数的一阶条件

- 函数 f 可微分且定义域为开集，所有点的梯度为：

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)$$

- 一阶条件：
  - 当且仅当 $f(y) \geq f(x) + \nabla f(x)^T (y - x)$ for all $x, y \in \mathbf{dom}\, f$
  - 凸集上可微分的函数 f 为凸函数
  - 函数一阶近似为全局低估



二阶条件

- 函数 f 二阶可微，若其定义域为开集，则Hessian矩阵 $\nabla^2 f(x) \in \mathbb{S}^n$处处存在：

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i,j = 1, \ldots, n,$$

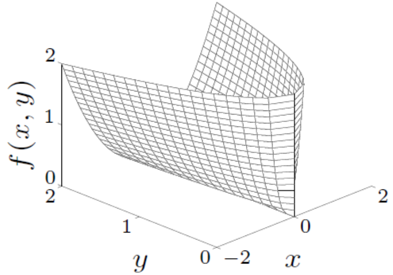
- 二阶条件：
  - 凸定义域二次可微函数 f 为凸函数的充分必要条件

$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x \in \operatorname{dom} f$$

▶示例：凸函数的二阶条件

- 二次函数 $f(x) = (1/2)x^T Px + q^T x + r \ (P \in \mathbb{S}^n)$
  - 一阶、二阶导数分别为 $\nabla f(x) = Px + q, \qquad \nabla^2 f(x) = P$
  - 当P为正定矩阵时为凸函数
- 最小二乘目标函数 $\quad f(x) = \|Ax - b\|_2^2$
  - ：一阶、二阶导数分别为

    $$\nabla f(x) = 2A^T(Ax - b), \quad \nabla^2 f(x) = 2A^T A$$
  - 对于任意矩阵A，目标函数均为凸函数



quadratic-over-linear: $f(x, y) = x^2/y$

$$\nabla^2 f(x, y) = \frac{2}{y^3} \begin{bmatrix} y \\ -x \end{bmatrix} \begin{bmatrix} y \\ -x \end{bmatrix}^T \succeq 0$$

convex for $y > 0$

▶ 上境图与水平子集

- 函数 $f:\mathbb{R}^n\to\mathbb{R}$ 的 α 水平子集(sublevel set)定义为：

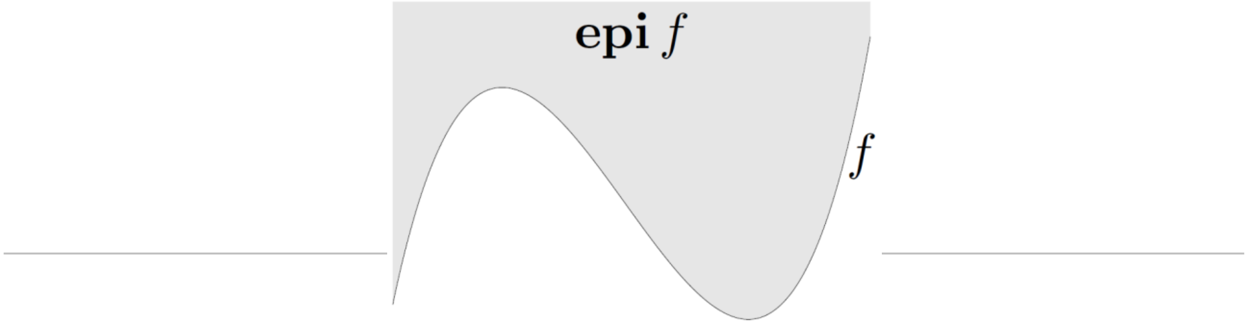
$$C_\alpha=\{x\in \mathrm{dom}\, f | f(x)\le\alpha\}$$

凸函数的水平子集仍为凸函数（反之不成立）

- 函数 $f:\mathbb{R}^n\to\mathbb{R}$ 的上境图（epigraph）定义为：

$$\mathrm{epi}\, f=\{(x,t)\in\mathbb{R}^{n+1} | x\in \mathrm{dom}\, f, f(x)\le t\}$$

函数为凸的充分必要条件是其上境图为凸集



Jensen不等式

- 基本不等式：
  – 如果函数 f 为凸，若 $0 \le \theta \le 1$则

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

- 扩展形式：
  – 如果函数 f 为凸函数，所有随机变量 z 的期望满足：

$$f(\mathbf{E}z) \le \mathbf{E}f(z)$$

▶ 保持函数凸性的运算

- 定义验证：二阶可微函数导数为正定矩阵
- 保持函数凸性的运算：
  - 仿射函数的组合
  - 逐点最大值与上确界
  - 最小化

▶ 逐点最大值

- 若 $f_1, \ldots, f_m$ 均为凸函数，则如下函数亦为凸函数

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

- 示例：分段线性函数为凸函数

$$f(x) = \max_{i=1,\ldots,m}(a_i^T x + b_i)$$

逐点上确界

- 若函数 f(x,y)为 x 的凸函数，则对于任意 y ∈A，如下逐点下确界函数为凸：

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

- 示例：
  - 集合 C 的支持函数为凸：$S_C(x) = \sup_{y \in C} y^T x$
  - 集合C中到最远点的距离：$f(x) = \sup_{y \in C} \|x - y\|$
  - 对称矩阵的最大特征值：$\lambda_{\max}(X) = \sup_{\|y\|_2 = 1} y^T X y$

最小化

- 若 f(x,y) 为凸函数且C为凸集，则最小化为凸函数

$$g(x) = \inf_{y \in C} f(x, y)$$

- 示例：

函数 $f(x,y) = x^T Ax + 2x^T By + y^T Cy$ 满足 $\begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \succeq 0, \quad C \succ 0$

最小化 y 得到凸函数

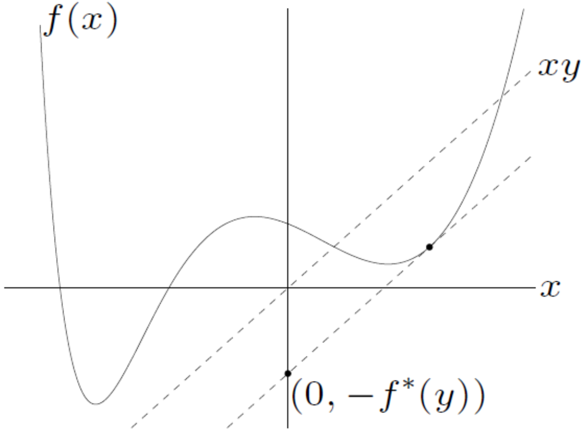
$$g(x) = \inf_y f(x,y) = x^T (A - BC^{-1}B^T)x$$

▶ 共轭函数

- 函数 f 的共轭定义为：

$$f^*(y) = \sup_{x \in \mathrm{dom}\, f} \left(y^T x - f(x) \right)$$

- 无论原函数 f 是否为凸，共轭函数f* 恒为凸函数
- 上述特性在凸优化求解中非常有用



negative logarithm $f(x) = -\log x$

$$f^*(y) = \sup_{x>0}(xy + \log x)$$
$$= \begin{cases} -1 - \log(-y) & y < 0 \\ \infty & \text{otherwise} \end{cases}$$

convex quadratic $f(x) = (1/2)x^T Q x$ with $Q \in \mathbb{S}^n_{++}$

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

不 止 于 代 码

THANK YOU

AI100