第一章 绪论

修贤超

https://xianchaoxiu.github.io

目录

最优化问题的一般形式

- 最优化问题一般可以描述为

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in \mathcal{X} \end{aligned} \tag{1}$$

  - $x = (x_1, x_2, \cdots, x_n)^\top \in \mathbb{R}^n$ 是**决策变量**
  - $f: \mathbb{R}^n \to \mathbb{R}$ 是**目标函数**
  - $\mathcal{X} \subseteq \mathbb{R}^n$ 是**约束集合**或**可行域**, 可行域包含的点称为**可行解**或**可行点**
  - 当 $\mathcal{X} = \mathbb{R}^n$ 时, 问题 (1) 称为**无约束优化问题**
  - 集合 $\mathcal{X}$ 通常可以由约束函数 $c_i(x): \mathbb{R}^n \to \mathbb{R}, i = 1, 2, \cdots, m + l$ 表达为

$$\begin{aligned} \mathcal{X} = \{x \in \mathbb{R}^n \mid\ & c_i(x) \le 0, \quad i = 1, 2, \cdots, m, \\ & c_i(x) = 0, \quad i = m+1, m+2, \cdots, m+l\} \end{aligned}$$

最优化问题的一般形式

- 在所有满足约束条件的决策变量中，使目标函数取最小值的变量 x^* 称为优化问题 (1) 的**最优解**，即对任意 $x \in \mathcal{X}$ 都有

$$f(x) \geq f(x^*)$$

  - 如果求解目标函数 f 的最大值，则 "min" 应替换为 "max"
  - 函数 f 的最小（最大）值不一定存在，但其下（上）确界总是存在的
  - x 可以是矩阵、多维数组或张量等

最优化问题的类型

- 线性规划 目标函数和约束函数均为线性函数的问题
- 整数规划 变量只能取整数的问题
- 非线性规划 目标函数和约束函数中至少有一个为非线性函数的问题
- 二次规划 目标函数是二次函数而约束函数是线性函数的问题
- 半定规划 极小化关于半正定矩阵的线性函数的问题
- 稀疏优化 最优解只有少量非零元素的问题
- 非光滑优化 包含非光滑函数的问题
- 低秩矩阵优化 最优解是低秩矩阵的问题
- 鲁棒优化、组合优化、随机优化、零阶优化、流形优化、分布式优化等

目录

实例

- 给定 $b \in \mathbb{R}^m$, 矩阵 $A \in \mathbb{R}^{m\times n}$ ，且向量 b 的维数远小于向量 x 的维数, 即 $m \ll n$. 考虑线性方程组求解问题

$$Ax = b$$

  - 方程组欠定, 存在无穷多个解
  - 原始信号中有较多的零元素, 即稀疏解

$$(\ell_0)\begin{cases}\min\limits_{x\in\mathbb{R}^n} & \|x\|_0 \\ \text{s.t.} & Ax=b\end{cases} \qquad (\ell_2)\begin{cases}\min\limits_{x\in\mathbb{R}^n} & \|x\|_2 \\ \text{s.t.} & Ax=b\end{cases} \qquad (\ell_1)\begin{cases}\min\limits_{x\in\mathbb{R}^n} & \|x\|_1 \\ \text{s.t.} & Ax=b\end{cases}$$

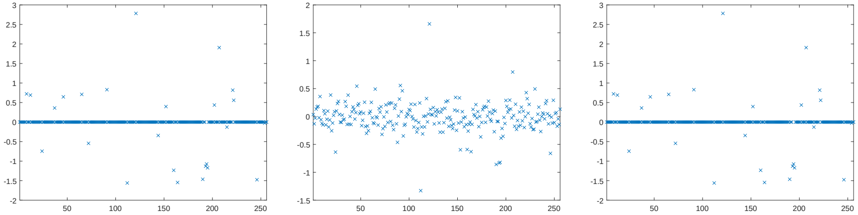
- 压缩感知（compressive sensing), 即通过部分信息恢复全部信息的解决方案

稀疏优化

- MATLAB 仿真

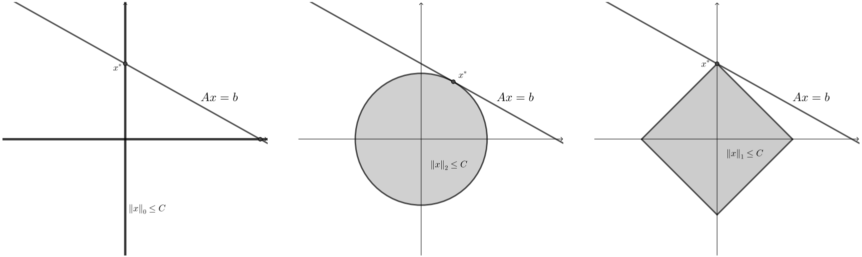
```
m = 128; n = 256;
A = randn(m, n); u = sprandn(n, 1, 0.1);
b = A * u;
```

- 若 A, b 满足一定的条件，向量 u 也是 ℓ_1 范数优化问题的**唯一最优解**



稀疏优化

- 原点到仿射集 $Ax = b$ 的投影



- 绝对值函数在零点处不可微，即**非光滑**
- A 通常是稠密矩阵，甚至元素未知或者不能直接存储

LASSO 问题

- 考虑带 ℓ_1 范数正则项的优化问题

$$\min_{x\in\mathbb{R}^n} \quad \mu\|x\|_1 \quad \text{s.t.} \quad Ax = b \tag{2}$$

$$\Downarrow$$

$$\min_{x\in\mathbb{R}^n} \quad \mu\|x\|_1 + \frac{1}{2}\|Ax-b\|_2^2 \tag{3}$$

  - $\mu > 0$ 是给定的正则化参数
  - 称为 LASSO（least absolute shrinkage and selection operator）
- 本课程大部分算法都将针对 (2) 和 (3) 给出

深度学习

- 深度学习（deep learning）是机器学习的一个子领域
- 常见的激活函数类型
  - Sigmoid 函数

$$t(z) = \frac{1}{1 + \exp(-z)}$$

  - Heaviside 函数

$$t(z) = \begin{cases} 1, & z \geq 0 \\ 0, & z < 0 \end{cases}$$

  - ReLU 函数

$$t(z) = \max\{0, z\}$$

深度学习中的优化算法

- 典型的数学模型

$$\min_{x \in \mathcal{W}} \quad \frac{1}{N} \sum_{i=1}^{N} \ell(f(a_i, x), b_i) + \mu \varphi(x)$$

- 随机梯度类算法
  - pytorch/caffe2: adadelta, adagrad, adam, nesterov, rmsprop, YellowFin
    https://github.com/pytorch/pytorch/tree/master/caffe2/sgd
  - pytorch/torch: sgd, asgd, adagrad, rmsprop, adadelta, adam, adamax
    https://github.com/pytorch/pytorch/tree/master/torch/optim
  - tensorflow: Adadelta, AdagradDA, Adagrad, ProximalAdagrad, Ftrl, Momentum, adam, Momentum, CenteredRMSProp
    https://github.com/tensorflow/tensorflow/blob/master/tensorflow/core/kernels/training_ops.cc

目录

向量范数的定义

- **定义** 令记号 $\|\cdot\|:\ \mathbb{R}^n \to \mathbb{R}^+$ 是一种非负函数, 如果满足
  - **正定性** 对于 $\forall v \in \mathbb{R}^n$, 有 $\|v\| \geqslant 0$, 且 $\|v\| = 0$ 当且仅当 $v = 0_{n\times 1}$
  - **齐次性** 对于 $\forall v \in \mathbb{R}^n$ 和 $\alpha \in \mathbb{R}$, 有 $\|\alpha v\| = |\alpha|\|v\|$
  - **三角不等式** 对于 $\forall v, w \in \mathbb{R}^n$, 均成立 $\|v+w\| \leqslant \|v\| + \|w\|$

  则称 $\|\cdot\|$ 是定义在向量空间 $\mathbb{R}^n$ 上的**向量范数**

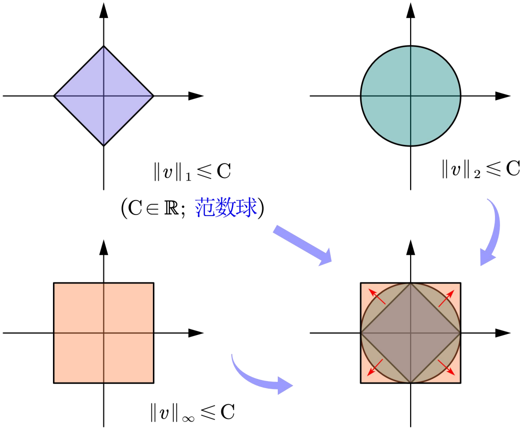
- 最常用的向量范数

$$\|v\|_p = \left(|v_1|^p + |v_2|^p + \cdots + |v_n|^p\right)^{\frac{1}{p}} \ (p \geq 1)$$

$$\|v\|_\infty = \max_{1\leqslant j\leqslant n} |v_j|$$

向量范数的定义

- **不同范数所度量的距离分别具有怎样的特征呢?**



矩阵范数

- ℓ_1 范数 $\|A\|_1 = \sum_{i,j} |A_{ij}|$
- Frobenius 范数 $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2} = \sqrt{\mathrm{Tr}(AA^\top)}$
- 算子范数是一类特殊的矩阵范数，由向量范数诱导得到

$$\|A\|_{(m,n)} = \max_{x\in\mathbb{R}^n, \|x\|_{(n)}=1} \|Ax\|_{(m)}$$

  - $p=1$ 时，$\|A\|_{p=1} = \max_{\|x\|_1=1} \|Ax\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^m |a_{ij}|$
  - $p=2$ 时，$\|A\|_{p=2} = \max_{\|x\|_2=1} \|Ax\|_2 = \sqrt{\lambda_{\max}(A^\top A)}$，又称为 A 的谱范数
  - $p=\infty$ 时，$\|A\|_{p=\infty} = \max_{\|x\|_\infty=1} \|Ax\|_\infty = \max_{1\leqslant i\leqslant m} \sum_{j=1}^n |a_{ij}|$

矩阵范数

- **核范数**

$$\|A\|_* = \sum_{i=1}^{r} \sigma_i$$

- **矩阵内积**

$$\langle A, B \rangle = \mathrm{Tr}(AB^\top) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ij}$$

- **命题** 设 $A, B \in \mathbb{R}^{m \times n}$, 则

$$|\langle A, B \rangle| \leqslant \|A\|_F \|B\|_F$$

等号成立当且仅当 A 和 B 线性相关, 即柯西不等式

- **性质** 同一矩阵空间内, 矩阵范数彼此之间是相互等价的

目录

梯度

- **定义** 给定函数 $f: \mathbb{R}^n \to \mathbb{R}$，且 f 在点 x 的一个邻域内有意义，若存在向量 $g \in \mathbb{R}^n$ 满足

$$\lim_{p \to 0} \frac{f(x+p) - f(x) - \langle g, p \rangle}{\|p\|} = 0$$

  其中 $\|\cdot\|$ 是任意的向量范数，称 f 在点 x 处可微（或 Fréchet 可微），g 为 f 在点 x 处的梯度，记作

$$\nabla f(x) = [\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \cdots, \frac{\partial f(x)}{\partial x_n}]^\top$$

- 如果对区域 D 上的每一个点 x 都有 $\nabla f(x)$ 存在，则称 f 在 D 上可微

海瑟矩阵

- 定义 如果函数 $f(x): \mathbb{R}^n \to \mathbb{R}$ 在点 x 处的二阶偏导数 $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ 都存在，则 f 在点 x 处的海瑟矩阵为

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_3} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_3} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \frac{\partial^2 f(x)}{\partial x_n \partial x_3} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

- 当 $\nabla^2 f(x)$ 在区域 D 上的每个点 x 处都存在时，称 f 在 D 上二阶可微
- 若 $\nabla^2 f(x)$ 在 D 上还连续，则称 f 在 D 上二阶连续可微

梯度利普希茨连续

- **定义** 给定可微函数 f, 若存在 $L > 0$, 对任意的 $x, y \in \mathrm{dom}\ f$ 有

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$$

  则称 f 是**梯度利普希茨连续的**, 相应利普希茨常数为 L

- **引理** 设可微函数 $f(x)$ 的定义域为 $\mathbb{R}^n$ 且为梯度 L -利普希茨连续的, 则函数 $f(x)$ 有**二次上界**

$$f(y) \le f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2, \quad \forall\ x, y \in \mathrm{dom}\ f$$

- $f(x)$ 定义域的要求可减弱为凸集

梯度利普希茨连续

- **推论** 设可微函数 $f(x)$ 的定义域为 $\mathbb{R}^n$ 且存在一个全局极小点 x^*, 若 $f(x)$ 为梯度 L-利普希茨连续的, 则对任意的 x 有

$$\frac{1}{2L}\|\nabla f(x)\|^2 \leq f(x) - f(x^*)$$

证明 由于 x^* 是全局极小点, 有

$$f(x^*) \leq f(y) \leq f(x) + \nabla f(x)^\top (y-x) + \frac{L}{2}\|y-x\|^2$$

上式对任意的 y 均成立, 因此可对不等号右边取下确界

$$\begin{aligned} f(x^*) &\leq \inf_{y\in\mathbb{R}^n} \{f(x) + \nabla f(x)^\top (y-x) + \frac{L}{2}\|y-x\|^2\} \\ &= f(x) - \frac{1}{2L}\|\nabla f(x)\|^2 \end{aligned}$$

矩阵变量函数的导数

- 对于函数 $f(X)$, 若存在矩阵 $G \in \mathbb{R}^{m\times n}$ 满足

$$\lim_{V\to 0} \frac{f(X+V) - f(X) - \langle G, V\rangle}{\|V\|} = 0$$

其中 $\|\cdot\|$ 是任意矩阵范数, 称矩阵变量函数 f 在 X 处 Fréchet 可微, G 为 f 在 Fréchet 可微意义下的梯度, 记为

$$\nabla f(X) = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{2n}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \frac{\partial f}{\partial x_{m2}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix}$$

- **定义** 如果对任意方向 $V \in \mathbb{R}^{m\times n}$, 存在矩阵 $G \in \mathbb{R}^{m\times n}$ 满足

$$\lim_{V\to 0} \frac{f(X+V) - f(X) - \langle G, V\rangle}{\|V\|} = 0$$

$$\Downarrow$$

$$\lim_{t\to 0} \frac{f(X+tV) - f(X) - t\langle G, V\rangle}{t} = 0$$

  则称 f 关于 X *Gâteaux* **可微**, G 为 f 在 X 处 *Gâteaux* 可微意义下的梯度

- 当 f 是 Fréchet 可微函数时, f 也是 Gâteaux 可微的, 且梯度相等

例

- **线性函数** $f(X) = \mathrm{Tr}(AX^\top B)$

$$\begin{aligned}\lim_{t\to 0}\frac{f(X+tV)-f(X)}{t} &= \lim_{t\to 0}\frac{\mathrm{Tr}(A(X+tV)^\top B)-\mathrm{Tr}(AX^\top B)}{t}\\ &= \mathrm{Tr}(AV^\top B) = \langle BA, V\rangle\end{aligned}$$

$$\Rightarrow \quad \nabla f(X) = BA$$

- **二次函数** $f(X,Y) = \|XY - A\|_F^2$

$$\begin{aligned}f(X,Y+tV)-f(X,Y) &= \|X(Y+tV)-A\|_F^2 - \|XY-A\|_F^2\\ &= 2\langle tXV, XY-A\rangle + t^2\|XV\|_F^2\\ &= 2t\langle V, X^\top(XY-A)\rangle + \mathcal{O}(t^2)\end{aligned}$$

$$\Rightarrow \quad \frac{\partial f}{\partial Y} = 2X^\top(XY-A)$$

目录

广义实值函数

- 在最优化领域，经常涉及取 $\inf$ （$\sup$） 操作，可能为无穷
- 定义 令 $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ 为广义实数空间，则映射

$$f: \ \mathbb{R}^n \to \overline{\mathbb{R}}$$

  称为广义实值函数
- 规定
  - $-\infty < \alpha < \infty, \forall \alpha \in \mathbb{R}$
  - $(+\infty) + (+\infty) = +\infty, \quad (+\infty) + \alpha = +\infty, \forall \alpha \in \mathbb{R}$

适当函数

- 定义 给定广义实值函数 f 和非空集合 $\mathcal{X}$, 若存在 $x \in \mathcal{X}$ 使 $f(x) < +\infty$, 并且对任意的 $x \in \mathcal{X}$ 都有 $f(x) > -\infty$, 则称函数 f 是关于集合 $\mathcal{X}$ 的适当函数
- 具体含义
  - 至少有一处取值不为正无穷
  - 处处取值不为负无穷
- 对于适当函数 f, 规定其定义域

$$\mathrm{dom}\, f = \{x \mid f(x) < +\infty\}$$

- 若无特殊说明, 定理中所讨论的函数均为适当函数

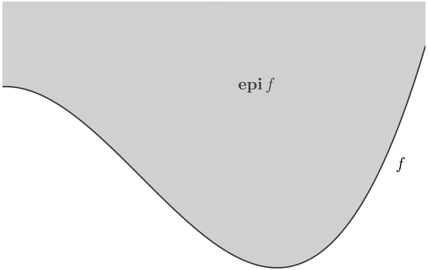
闭函数

- **定义** 设 f 为广义实值函数，α -下水平集定义为

$$C_\alpha = \{x \mid f(x) \le \alpha\}$$

- **定义** 设 f 为广义实值函数，上方图定义为

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

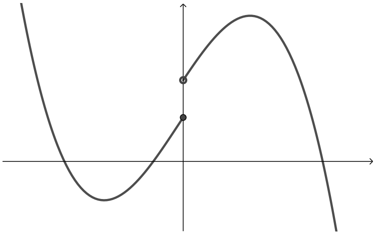


下半连续函数

- **定义** 设 f 为广义实值函数，若 $\operatorname{epi} f$ 为闭集，则称 f 为**闭函数**
- **定义** 设 f 为广义实值函数，若对任意的 $x \in \mathbb{R}^n$，有

$$\liminf_{y \to x} f(y) \geq f(x)$$

则 $f(x)$ 为**下半连续函数**



闭函数与下半连续函数

- **定理** 设广义实值函数 $f: \mathbb{R}^n \to \overline{\mathbb{R}}$, 则以下命题等价
  - $f(x)$ 的任意 α -下水平集都是闭集
  - $f(x)$ 是下半连续的
  - $f(x)$ 是闭函数
- **性质**
  - 若 f 与 g 均为适当的闭（下半连续）函数, 并且 $\mathrm{dom}\ f \cap \mathrm{dom}\ g \neq \emptyset$, 则 $f+g$ 也是闭（下半连续）函数
  - 若 f 为闭（下半连续）函数, 则 $f(Ax+b)$ 也为闭（下半连续）函数

目录

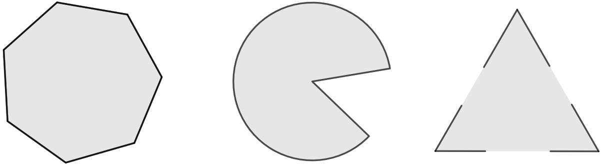
凸集的几何定义

- **定义** 若过集合 $\mathcal{C}$ 中的任意两点的直线都在 $\mathcal{C}$ 内, 则称 $\mathcal{C}$ 为**仿射集**, 即

$$x_1, x_2 \in \mathcal{C} \quad \Rightarrow \quad \theta x_1 + (1-\theta) x_2 \in \mathcal{C}, \forall \theta \in \mathbb{R}$$

- **定义** 若连接集合 $\mathcal{C}$ 中的任意两点的线段都在 $\mathcal{C}$ 内, 则称 $\mathcal{C}$ 为**凸集**, 即

$$x_1, x_2 \in \mathcal{C} \quad \Rightarrow \quad \theta x_1 + (1-\theta) x_2 \in \mathcal{C}, \forall 0 \leqslant \theta \leqslant 1$$



(a) (b) (c)

凸集的性质

- 若 $\mathcal{S}$ 是凸集, 则 $k\mathcal{S} = \{ks \mid k \in \mathbb{R}, s \in \mathcal{S}\}$ 是凸集
- 若 $\mathcal{S}$ 和 $\mathcal{T}$ 均是凸集, 则 $\mathcal{S} + \mathcal{T} = \{s + t \mid s \in \mathcal{S}, t \in \mathcal{T}\}$ 是凸集
- 若 $\mathcal{S}$ 和 $\mathcal{T}$ 均是凸集, 则 $\mathcal{S} \cap \mathcal{T}$ 是凸集

  证明 设 $x, y \in \mathcal{S} \cap \mathcal{T}$ 且 $\theta \in [0, 1]$. 由于 $\mathcal{S}$ 和 $\mathcal{T}$ 均为凸集, 则

$$\theta x + (1 - \theta) y \in \mathcal{S} \cap \mathcal{T}$$

- 凸集的内部和闭包都是凸集
- 任意多凸集的交都是凸集

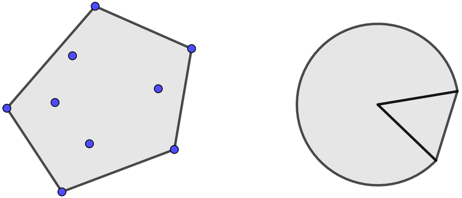
凸组合和凸包

- 形如

$$x = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k$$
$$\theta_1 + \cdots + \theta_k = 1, \theta_i \geqslant 0, i = 1, \cdots, k$$

  的点称为 $x_1, \cdots, x_k$ 的凸组合

- 集合 $\mathcal{S}$ 的所有点的凸组合构成的点集为 $\mathcal{S}$ 的凸包, 记为 $\mathrm{conv}\ \mathcal{S}$

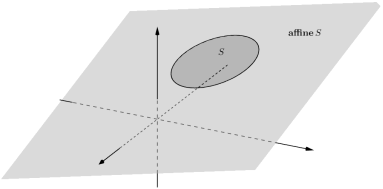


- $\mathrm{conv}\ \mathcal{S}$ 是包含 $\mathcal{S}$ 的最小凸集

仿射组合和仿射包

- **定义** 形如
$$x = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k$$
$$\theta_1 + \cdots + \theta_k = 1, \theta_i \in \mathbb{R}, i = 1, \cdots, k$$
的点称为 $x_1, \cdots, x_k$ 的**仿射组合**

- 集合 $\mathcal{S}$ 的所有点的仿射组合构成的点集为 $\mathcal{S}$ 的**仿射包**, 记为 affine $\mathcal{S}$



- affine $\mathcal{S}$ 是包含 $\mathcal{S}$ 的最小仿射集

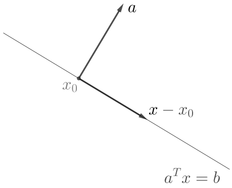
锥组合和凸锥

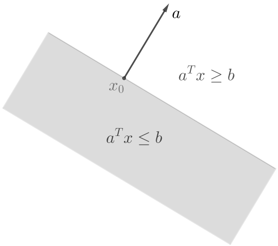
- 形如
$$x = \theta_1 x_1 + \cdots + \theta_k x_k, \theta_i > 0, i = 1, \cdots, k$$
的点称为 $x_1, \cdots, x_k$ 的**锥组合**
- 若集合 $\mathcal{S}$ 中任意点的锥组合都在 $\mathcal{S}$ 中, 则称 $\mathcal{S}$ 为**凸锥**

- 锥组合不要求系数的和为 1

超平面和半空间

- 任取非零向量 $a \in \mathbb{R}^n$, 称 $\{x \mid a^\top x = b\}$ 为超平面, $\{x \mid a^\top x \leqslant b\}$ 为半空间
- 满足线性等式和不等式组点的集合 $\{x \mid Ax \leqslant b, Cx = d\}$ 称为多面体





- 超平面是仿射集和凸集, 半空间是凸集但不是仿射集
- 多面体是有限个半空间和超平面的交

分离超平面定理

- **定理** 如果 $\mathcal{C}$ 和 $\mathcal{D}$ 是不相交的凸集, 则存在非零向量 a 和常数 b, 使得

$$a^\top x \leqslant b, \forall x \in \mathcal{C} \quad \text{且} \quad a^\top x \geqslant b, \forall x \in \mathcal{D}$$

即超平面 $\{x \mid a^\top x = b\}$ 分离了 $\mathcal{C}$ 和 $\mathcal{D}$

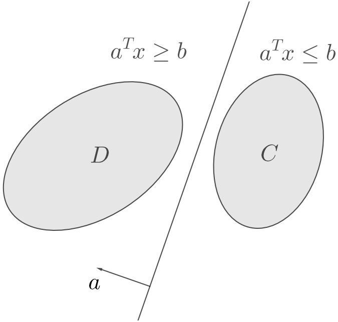
- **定理** 如果存在非零向量 a 和常数 b, 使得

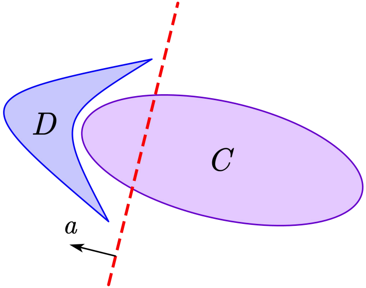
$$a^\top x < b, \forall x \in \mathcal{C} \quad \text{且} \quad a^\top x > b, \forall x \in \mathcal{D}$$

即超平面 $\{x \mid a^\top x = b\}$ **严格**分离了 $\mathcal{C}$ 和 $\mathcal{D}$

分离超平面的示意

- 在 $\mathbb{R}^2$ 中的 2 个凸集使用超平面即可轻松划分，但遇到非凸集合就必须使用更加复杂的平面

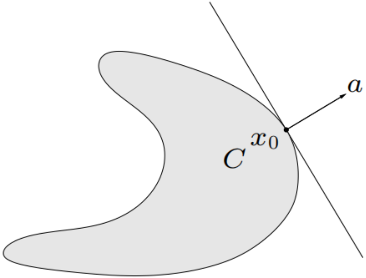




支撑超平面

- **定义** 给定集合 $\mathcal{C}$ 及其边界点 x_0, 如果 $a \neq 0$ 满足 $a^\top x \leqslant a^\top x_0, \forall x \in \mathcal{C}$, 则称集合
$$\{x \mid a^\top x = a^\top x_0\}$$
为 $\mathcal{C}$ 在边界点 x_0 处的**支撑超平面**

- **定理** 若 $\mathcal{C}$ 是凸集, 则 $\mathcal{C}$ 的任意边界点处都存在支撑超平面



球和椭球

- 称空间中到点 x_c 的距离小于等于定值 r 的集合为欧几里得球, 即

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

- 设形如

$$\{x \mid (x - x_c)^\top P^{-1}(x - x_c) \leqslant 1\} = \{x_c + Au \mid \|u\|_2 \leqslant 1\}$$

的集合为椭球, 其中 x_c 为椭球中心, P 为对称正定, 且 A 非奇异

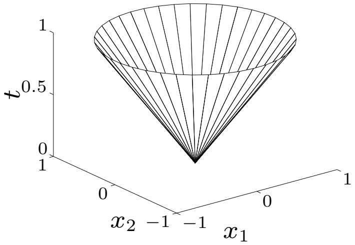
- 令 $\|\cdot\|$ 是任意一个范数, 称

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

为中心为 x_c 半径为 r 的范数球

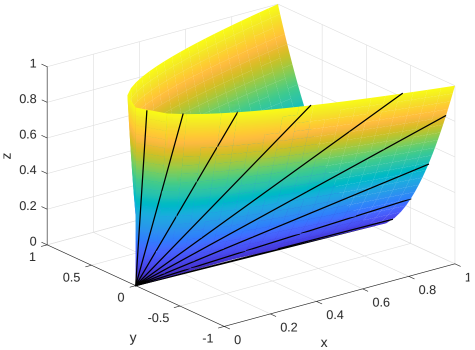
范数锥

- 形如
$$\{(x,t) \mid \|x\| \leqslant t\}$$
的集合为范数锥
- 使用 $\|\cdot\|_2$ 度量距离的锥为二次锥, 也称冰淇淋锥



- 范数球和范数锥都是凸集

- 记 $\mathcal{S}^n$ 为对称矩阵的集合，即 $\mathcal{S}^n = \{X \in \mathbb{R}^{n\times n} \mid X^\top = X\}$
- 记 $\mathcal{S}^n_+$ 为半正定矩阵的集合，即 $\mathcal{S}^n_+ = \{X \in \mathcal{S}^n \mid X \succeq 0\}$
- 记 $\mathcal{S}^n_{++}$ 为正定矩阵的集合，即 $\mathcal{S}^n_{++} = \{X \in \mathcal{S}^n \mid X \succ 0\}$



对于矩阵 $\begin{pmatrix} x & y \\ y & z \end{pmatrix}$，其特征值应全部大于等于 0

$$\Downarrow$$

$$\{(x,y,z) \mid x \geqslant 0, z \geqslant 0, xz \geqslant y^2\}$$

目录

凸函数的定义

- **定义** 设 $f:\ \mathbb{R}^n \to \mathbb{R}$ 为适当函数, 如果 $\mathrm{dom}\ f$ 是凸集, 且

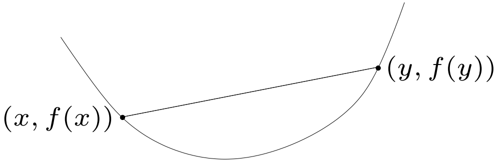
$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$$

  对所有 $x, y \in \mathrm{dom}\ f$, $0 \leq \theta \leq 1$ 都成立, 则称 f 是凸函数

- 若对所有 $x, y \in \mathrm{dom}\ f$, $x \neq y$, $0 < \theta < 1$, 有

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

  则称 f 是严格凸函数



一元凸函数的例

- 仿射函数　对任意 $a, b \in \mathbb{R}$, $ax + b$ 是 $\mathbb{R}$ 上的凸 (凹)函数
- 指数函数　对任意 $a \in \mathbb{R}$, e^{ax} 是 $\mathbb{R}$ 上的凸函数
- 绝对值的幂　对 $p \geq 1$, $|x|^p$ 是 $\mathbb{R}$ 上的凸函数
- 幂函数　对 $\alpha \geq 1$ 或 $\alpha \leq 0$, x^α 是 $\mathbb{R}_{++}$ 上的凸函数
- 幂函数　对 $0 \leq \alpha \leq 1$, x^α 是 $\mathbb{R}_{++}$ 上的凹函数
- 对数函数　$\log x$ 是 $\mathbb{R}_{++}$ 上的凹函数
- Sigmoid 函数、Heaviside 函数、ReLU 函数 . . .

多元凸函数的例

- 所有的仿射函数既是凸函数, 又是凹函数

$$f(x) = a^\top x + b$$

$$f(X) = \mathrm{Tr}(A^\top X) + b = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}X_{ij} + b$$

- 所有的范数都是凸函数

$$f(x) = \|x\|_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p} \ (p \geq 1)$$

$$f(X) = \|X\|_2 = \sigma_{\max}(X) = (\lambda_{\max}(X^\top X))^{1/2}$$

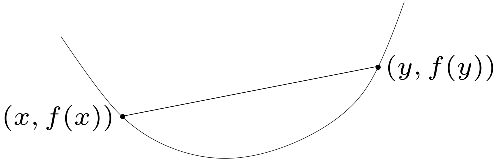
强凸函数

- **定义** 若存在常数 $m > 0$, 使得

$$g(x) = f(x) - \frac{m}{2}\|x\|^2$$

为凸函数, 则称 $f(x)$ 为**强凸函数**

- 为了方便也称 $f(x)$ 为 m-强凸函数
- **命题** 设 f 为强凸函数且存在最小值, 则 f 的最小值点唯一



凸函数判定定理

- **定理** $f(x)$ 是凸函数当且仅当对每个 $x \in \operatorname{dom} f$, $v \in \mathbb{R}^n$, 函数 $g: \mathbb{R} \to \mathbb{R}$ 是关于 t 的凸函数

$$g(t) = f(x + tv), \quad \operatorname{dom} g = \{t \mid x + tv \in \operatorname{dom} f\}$$

- **例** $f(X) = -\log\det X$ 是凸函数, 其中 $\operatorname{dom} f = \mathcal{S}_{++}^n$

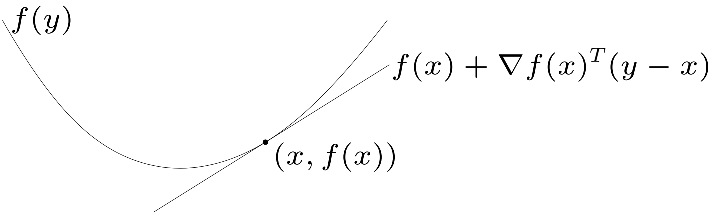
证明 任取 $X \succ 0$ 以及方向 $V \in \mathcal{S}^n$, 将 f 限制在直线 $X + tV$ 上, 则

$$\begin{aligned} g(t) &= -\log\det(X + tV) \\ &= -\log\det X - \log\det(I + tX^{-1/2}VX^{-1/2}) \\ &= -\log\det X - \sum_{i=1}^{n} \log(1 + t\lambda_i) \end{aligned}$$

一阶条件

- **定理** 对于定义在凸集上的可微函数 f, 则 f 是凸函数当且仅当

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) \quad \forall x, y \in \operatorname{dom} f$$



- **定理** 设 f 为可微函数, 则 f 为凸函数当且仅当 $\operatorname{dom} f$ 为凸集且 ∇f 为**单调映射**

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq 0, \quad \forall\ x, y \in \operatorname{dom} f$$

二阶条件

- **定理** 设 f 为定义在凸集上的二阶连续可微函数, 则 f 是凸函数当且仅当

$$\nabla^2 f(x) \succeq 0 \quad \forall x \in \mathrm{dom}\ f$$

- 如果

$$\nabla^2 f(x) \succ 0 \ \forall x \in \mathrm{dom}\ f$$

  则 f 是**严格凸函数**

- **例** 最小二乘函数 $f(x) = \|Ax - b\|_2^2$

$$\nabla f(x) = 2A^\top (Ax - b), \quad \nabla^2 f(x) = 2A^\top A$$

  对任意 A, 函数 f 都是凸函数

- **定理** 函数 $f(x)$ 为凸函数当且仅当其上方图 $\mathbf{epi}\ f$ 是凸集

证明 (必要性) 若 f 为凸函数, 则对任意 $(x_1, y_1), (x_2, y_2) \in \mathbf{epi}\ f, t \in [0, 1]$ 有

$$ty_1 + (1-t)y_2 \geq tf(x_1) + (1-t)f(x_2) \geq f(tx_1 + (1-t)x_2)$$

故 $(tx_1 + (1-t)x_2, ty_1 + (1-t)y_2) \in \mathbf{epi}\ f, t \in [0, 1]$

(充分性) 若 $\mathbf{epi}\ f$ 是凸集, 则对任意 $x_1, x_2 \in \mathbf{dom}\ f, t \in [0, 1]$ 有

$$(tx_1 + (1-t)x_2, tf(x_1) + (1-t)f(x_2)) \in \mathbf{epi}\ f$$

$$\Downarrow$$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

凸函数的判断方法

- 用定义验证（通常将函数限制在一条直线上）
- 利用一阶条件、二阶条件
- 直接研究 f 的上方图 $\mathrm{epi}\ f$
- 说明 f 可由简单的凸函数通过保凸运算得到
  - 非负加权和
  - 与仿射函数复合
  - 逐点取最大值
  - 与标量向量函数复合

非负加权和与仿射函数的复合

- **定理** (1) 若 f 是凸函数, 则 αf 是凸函数, 其中 $\alpha \geq 0$
- **定理** (2) 若 f_1, f_2 是凸函数, 则 $f_1 + f_2$ 是凸函数
- **定理** (3) 若 f 是凸函数, 则 $f(Ax + b)$ 是凸函数

例

- 线性不等式的对数障碍函数

$$f(x) = -\sum_{i=1}^{m} \log(b_i - a_i^\top x), \quad \mathrm{dom}\ f = \{x \mid a_i^\top x < b_i, i = 1, \cdots, m\}$$

- 仿射函数的（任意）范数 $f(x) = \|Ax + b\|$

逐点取最大值

- **定理** (4) 若 $f_1, \cdots, f_m$ 是凸函数, 则

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

是凸函数

例

- 分段线性函数

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

- $x \in \mathbb{R}^n$ 的前 r 个最大分量之和

$$f(x) = x_{[1]} + x_{[2]} + \cdots + x_{[r]}$$

$$\Updownarrow$$

$$f(x) = \max\{x_{i_1} + x_{i_2} + \cdots + x_{i_r} \mid 1 \leq i_1 < i_2 < \cdots < i_r \leq n\}$$

逐点取上界

- **定理** (5) 若对每个 $y \in \mathcal{A}$, $f(x, y)$ 是关于 x 的凸函数, 则

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

是凸函数

例

- 集合 C 上点到给定点 x 的最远距离

$$f(x) = \sup_{y \in C} \|x - y\|$$

- 对称矩阵 $X \in \mathcal{S}^n$ 的最大特征值

$$\lambda_{\max}(X) = \sup_{\|y\|_2 = 1} y^\top X y$$

与函数的复合

- **定理** (6) 给定函数 $g:\ \mathbb{R}^n \to \mathbb{R}$ 和 $h:\ \mathbb{R} \to \mathbb{R}$, 令

$$f(x) = h(g(x))$$

若 $\begin{array}{l} g \text{ 是凸函数}, h \text{ 是凸函数且单调不减} \\ g \text{ 是凹函数}, h \text{ 是凸函数且单调不增} \end{array}$，那么 f 是凸函数

例

- 如果 g 是凸函数, 则 $\exp g(x)$ 是凸函数
- 如果 g 是正值凹函数, 则 $1/g(x)$ 是凸函数

定理 (7) 若 $f(x,y)$ 关于 (x,y) 整体是凸函数, $\mathcal{C}$ 是凸集, 则

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

是凸函数

例

- 考虑函数 $f(x,y) = x^\top Ax + 2x^\top By + y^\top Cy$, 海瑟矩阵满足

$$\begin{bmatrix} A & B \\ B^\top & C \end{bmatrix} \succeq 0, \quad C \succ 0$$

则 $f(x,y)$ 为凸函数. 对 y 求最小值得

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

- 点 x 到凸集 $\mathcal{S}$ 的距离 $\mathrm{dist}(x,\mathcal{S}) = \inf_{y\in\mathcal{S}} \|x-y\|$ 是凸函数

凸函数的性质

- **命题**　设 $f(x)$ 是凸函数, 则 $f(x)$ 的所有的 α -下水平集为凸集
- **引理**　设 $f(x)$ 是参数为 m 的可微强凸函数, 则如下不等式成立

$$g(y) \ge f(x) + \nabla f(x)^\top (y-x) + \frac{m}{2}\|y-x\|^2, \quad \forall x, y \in \mathrm{dom}\ f$$

证明　由强凸函数的定义有 $g(x) = f(x) - \frac{m}{2}\|y-x\|^2$ 是凸函数. 根据凸函数的一阶条件知

$$g(y) \ge g(x) + \nabla g(x)^\top (y-x)$$

$$\Downarrow$$

$$\begin{aligned} f(y) &\ge f(x) - \frac{m}{2}\|x\|^2 + \frac{m}{2}\|y\|^2 + (\nabla f(x) - mx)^\top (y-x) \\ &= f(x) + \nabla f(x)^\top (y-x) + \frac{m}{2}\|y-x\|^2 \end{aligned}$$

目录

次梯度

- 回顾可微凸函数 f 的一阶条件

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x)$$

- **定义** 设 f 为适当凸函数, x 为 $\mathrm{dom}\ f$ 中的一点. 若向量 $g \in \mathbb{R}^n$ 满足

$$f(y) \geq f(x) + g^\top (y - x), \quad \forall y \in \mathrm{dom}\ f$$

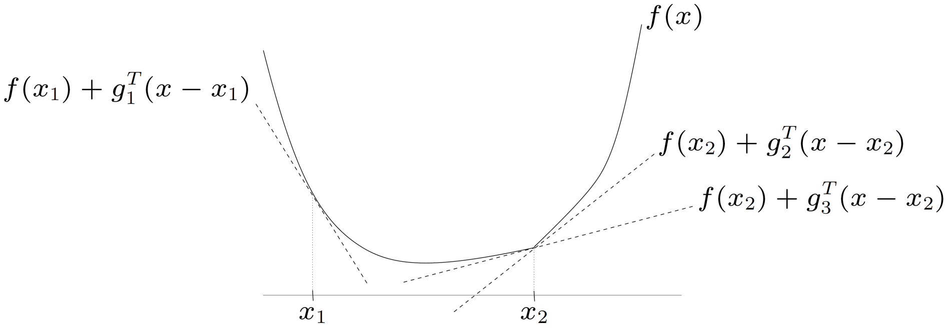
则称 g 为函数 f 在点 x 处的一个**次梯度**. 进一步, 称集合

$$\partial f(x) = \{g \mid g \in \mathbb{R}^n, f(y) \geq f(x) + g^\top (y - x), \forall y \in \mathrm{dom}\ f\}$$

为 f 在点 x 处的**次微分**

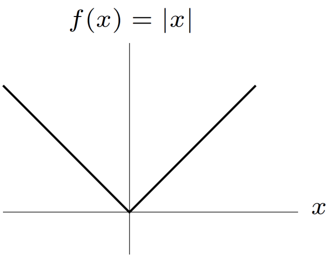
次梯度

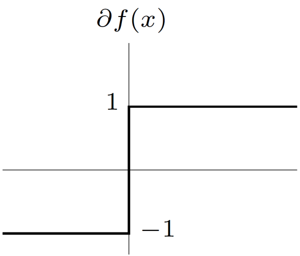
- g_1 是点 x_1 处的次梯度
- g_2, g_3 是点 x_2 处的次梯度



例

- 绝对值函数 $f(x) = |x|$





- 欧几里得范数 $f(x) = \|x\|_2$

$$\text{若 } x \neq 0,\ \partial f(x) = \frac{1}{\|x\|_2} x, \quad \text{若 } x = 0,\ \partial f(x) = \{g \mid \|g\|_2 \leq 1\}$$

次梯度的性质

- **定理** 设 f 是凸函数，则 $\partial f(x)$ 有如下性质
  - 对任何 $x \in \mathrm{dom}\ f$，$\partial f(x)$ 是一个闭凸集（可能为空集）
  - 如果 $x \in \mathrm{int\ dom}\ f$，则 $\partial f(x)$ 非空有界集
- **命题** 设凸函数 $f(x)$ 在 $x \in \mathrm{int\ dom}\ f$ 处可微，则

$$\partial f(x) = \{\nabla f(x)\}$$

- **定理** 设 $f: \mathbb{R}^n \to \mathbb{R}$ 为凸函数，$x, y \in \mathrm{dom}\ f$，则

$$(u - v)^\top (x - y) \geq 0$$

其中 $u \in \partial f(x)$，$v \in \partial f(y)$

两个函数之和的次梯度

- **定理** 设 $f_1, f_2 : \mathbb{R}^n \to (-\infty, +\infty]$ 是凸函数, 则对任意的 $x \in \mathbb{R}^n$ 有

$$\partial f_1(x) + \partial f_2(x) \subseteq \partial(f_1 + f_2)(x)$$

  进一步, 若 $\mathrm{int\ dom}\ f_1 \cap \mathrm{dom}\ f_2 \neq \emptyset$, 则对任意的 $x_0 \in \mathbb{R}^n$ 有

$$\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x)$$

- 若 $f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x), \quad \alpha_1, \alpha_2 \geq 0$, 则 $f(x)$ 的次微分

$$\partial f(x) = \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x)$$

- Moreau-Rockafellar 定理

函数族的上确界

- **定理** 设 $f_1, f_2, \cdots, f_m: \mathbb{R}^n \to (-\infty, +\infty]$ 均为凸函数, 令

$$f(x) = \max\{f_1(x), f_2(x), \cdots, f_m(x)\}, \quad \forall x \in \mathbb{R}^n$$

对 $x_0 \in \bigcap_{i=1}^{m} \operatorname{int} \operatorname{dom} f_i$，定义 $I(x_0) = \{i \mid f_i(x_0) = f(x_0)\}$，则

$$\partial f(x_0) = \operatorname{conv} \bigcup_{i \in I(x_0)} \partial f_i(x_0)$$

  - $I(x_0)$ 表示点 x_0 处“有效”函数的指标
  - $\partial f(x_0)$ 是点 x_0 处“有效”函数的次微分并集的凸包

- 如果 f_i 可微, $\partial f(x_0) = \operatorname{conv}\{\nabla f_i(x_0) \mid i \in I(x_0)\}$

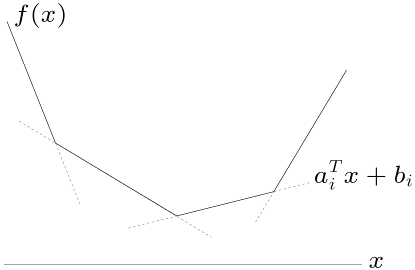
例

- **分段线性函数**

$$f(x) = \max_{i=1,2,\cdots,m} \{a_i^\top x + b_i\}$$

点 x 处的次微分是一个多面体

$$\partial f(x) = \mathrm{conv}\{a_i \mid i \in I(x)\}, \quad I(x) = \{i \mid a_i^\top x + b_i = f(x)\}$$



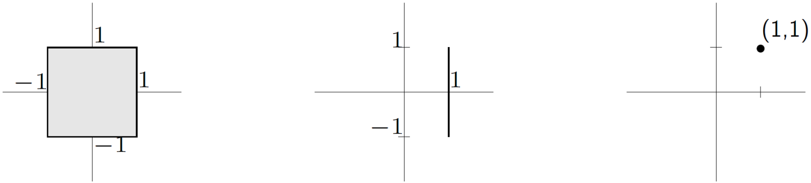
例

- ℓ_1 -范数

$$f(x) = \|x\|_1 = \max_{s\in\{-1,1\}^n} s^\top x$$

点 x 处的次微分是

$$\partial f(x) = J_1 \times \cdots \times J_n, \quad J_k = \begin{cases} [-1,1], & x_k = 0 \\ \{1\}, & x_k > 0 \\ \{-1\}, & x_k < 0 \end{cases}$$

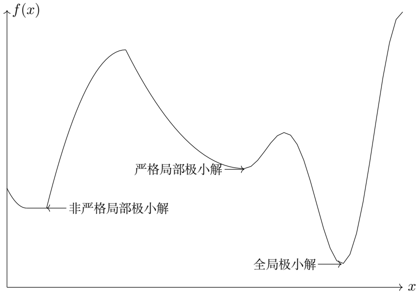


$\partial f(0,0) = [-1,1] \times [-1,1]$ $\partial f(1,0) = \{1\} \times [-1,1]$ $\partial f(1,1) = \{(1,1)\}$

目录

全局和局部最优解

- 如果 $f(\bar{x}) \le f(x), \forall x \in \mathcal{X}$, 则称 $\bar{x}$ 为全局极小解
- 如果存在 $N_\varepsilon(\bar{x})$ 使得 $f(\bar{x}) \le f(x), \forall x \in N_\varepsilon(\bar{x}) \cap \mathcal{X}$, 则称 $\bar{x}$ 为局部极小解
- 进一步, 如果有 $f(\bar{x}) < f(x), \forall x \in N_\varepsilon(\bar{x}) \cap \mathcal{X}$ 且 $x \ne \bar{x}$ 成立, 则称 $\bar{x}$ 为严格局部极小解



收敛性

- 给定初始点 x^0, 记算法迭代产生的点列为 $\{x^k\}$
  - 如果 $\{x^k\}$ 在某种范数 $\|\cdot\|$ 的意义下满足

$$\lim_{k\to\infty}\|x^k - x^*\| = 0$$

    且收敛的点 x^* 为一个局部（全局）极小解, 则称该算法依点列收敛到局部（全局）极小解
  - 如果从任意初始点 x^0 出发, 算法都是依点列收敛到局部（全局）极小解的, 则称该算法全局依点列收敛到局部（全局）极小解
  - 记对应的函数值序列 $\{f(x^k)\}$, 则称该算法（全局）依函数值收敛到局部（全局）极小值

收敛准则

- 对于无约束优化问题，常用的收敛准则有

$$\frac{f(x^k)-f^*}{\max\{|f^*|,1\}} \le \varepsilon_1, \quad \|\nabla f(x^k)\| \le \varepsilon_2$$

  如果最优解未知，通常使用相对误差

$$\frac{\|x^{k+1}-x^k\|}{\max\{\|x^k\|,1\}} \le \varepsilon_3, \quad \frac{|f(x^{k+1})-f(x^k)|}{\max\{|f(x^k)|,1\}} \le \varepsilon_4$$

- 对于约束优化问题，还需要考虑约束违反度

$$c_i(x^k) \le \varepsilon_5,\ i=1,2,\cdots,m$$
$$|c_i(x^k)| \le \varepsilon_6,\ i=m+1,m+2,\cdots,m+l$$

渐进收敛速度

- 设 $\{x^k\}$ 为算法产生的迭代点列且收敛于 x^*
  - Q-线性收敛

$$\frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|}\le a,\quad a\in(0,1)$$

  - Q-次线性收敛

$$\lim_{k\to\infty}\frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|}=1$$

  - Q-超线性收敛

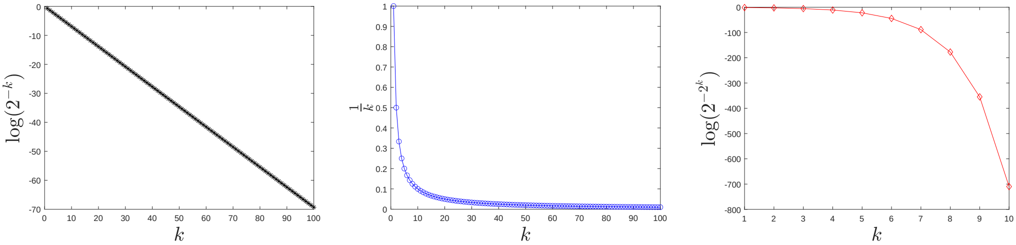
$$\lim_{k\to\infty}\frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|}=0$$

  - Q-二次收敛

$$\frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|^2}\le a,\quad a>0$$

渐进收敛速度

- 点列 $\{2^{-k}\}$ 是 Q-线性收敛的
- 点列 $\{1/k\}$ 是 Q-次线性收敛的
- 点列 $\{2^{-2^k}\}$ 是 Q-二次收敛的，也是 Q-超线性收敛的



一般来说，选择 Q-超线性收敛和 Q-二次收敛的算法

Q&A

Thank you!

感谢您的聆听和反馈