机器学习中的优化算法

Lecture02: 凸分析

张立平

清华大学数学科学系

办公室：理科楼#A302, Tel: 62798531

E-mail: lipingzhang@tsinghua.edu.cn

Contents and Acknowledgement

- 教材：**最优化：建模、算法与理论**

  http://bicmr.pku.edu.cn/ wenzw/bigdata2021.html

- 致谢：北京大学**文再文**教授

Outline of Lecture02

- 向量范数和矩阵范数
- 凸集
- 凸函数
- 共轭函数
- 次梯度

向量范数

- 最常用的向量范数ℓ_p范数($p \geq 1$): 令$v \in \mathbb{R}^n$,

$$\|v\|_p = \left(\sum_{i=1}^{n} |v_i|^p\right)^{\frac{1}{p}}.$$

- ℓ_∞范数: $\|v\|_\infty = \max_{1 \leq j \leq n} |v_{(j)}|$.

- 由正定矩阵A诱导的向量范数: $\|v\|_A = \sqrt{v^T A v}$.

- Cauchy不等式：设$a, b \in \mathbb{R}^n$, 则$|a^{\mathrm{T}} b| \leqslant \|a\|_2 \|b\|_2$,且等号成立的条件是a与b线性相关.

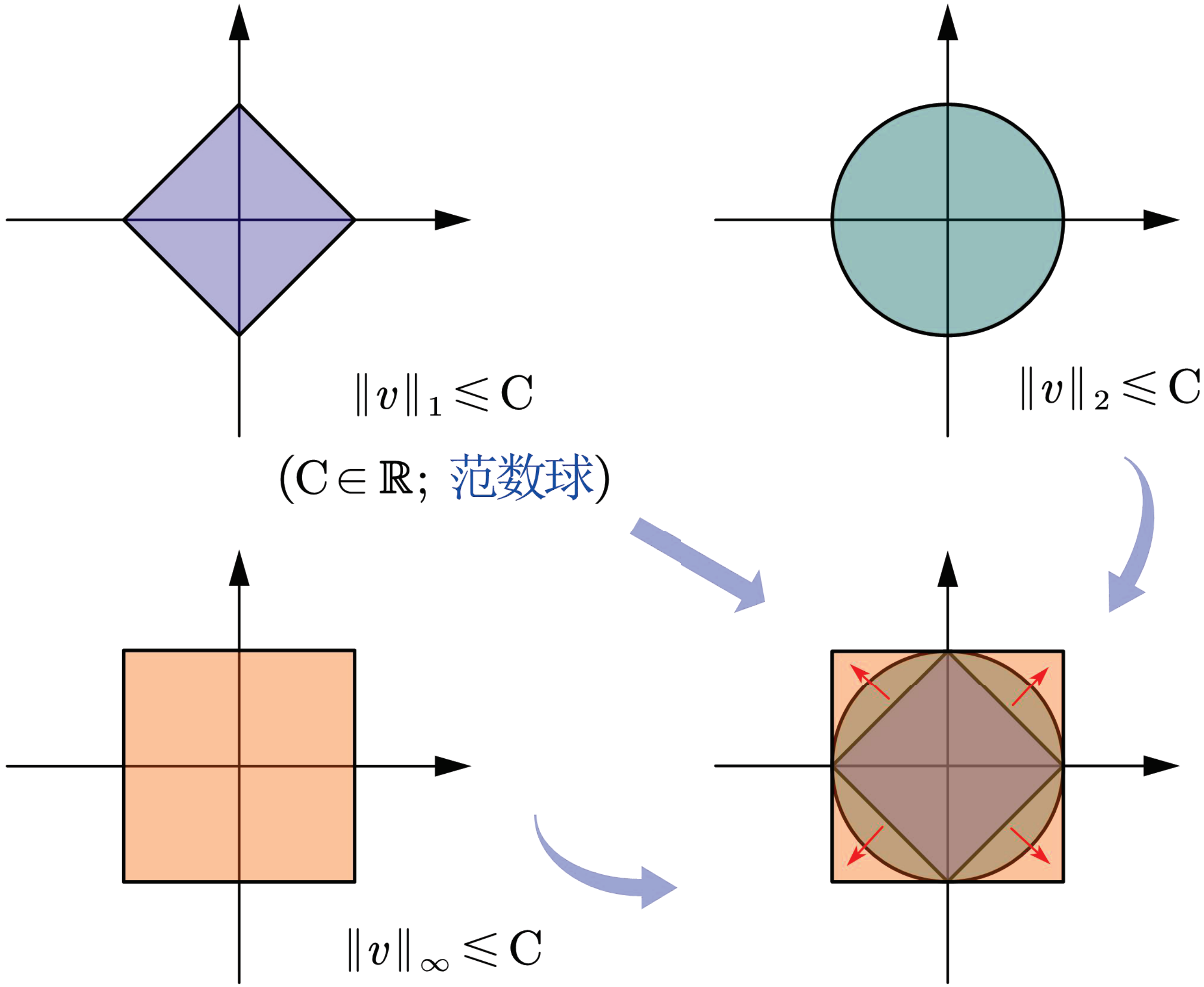


Figure 1: ℓ_1范数, ℓ_2范数和ℓ_∞范数

矩阵范数、核范数

Definition 1. *如果函数*$\|\cdot\|:\mathbb{R}^{m\times n}\to\mathbb{R}^+$*满足:*

- *正定性: 对*$\forall A\in\mathbb{R}^{m\times n}$*, 有*$\|A\|\geq 0$*且*$\|A\|=0\Leftrightarrow A=0_{m\times n}$*;*
- *齐次性: 对任意*$A\in\mathbb{R}^{m\times n}$*和*$\alpha\in\mathbb{R}$*, 有*$\|\alpha A\|=|\alpha|\,\|A\|$*;*
- *三角不等式: 对于任意*$A,B\in\mathbb{R}^{m\times n}$*, 有*$\|A+B\|\leq\|A\|+\|B\|$*.*

则称$\|\cdot\|$*是定义在向量空间*$\mathbb{R}^{m\times n}$*上的矩阵范数.*

矩阵的核范数以衡量矩阵的秩的大小.

Definition 2. *给定矩阵*$A\in\mathbb{R}^{m\times n}$*, 其核范数定义为*

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

其中$r=\mathbf{rank}\,(A)$*,* $\sigma_i(i=1,\cdots,r)$*为*A*的所有非零奇异值.*

Question: 核范数是矩阵范数吗?

矩阵ℓ_p范数

类似于向量ℓ_p范数，矩阵的ℓ_p范数：ℓ_1-范数，F-范数. 令$A \in \mathbb{R}^{m\times n}$,

- ℓ_1-范数：$\|A\|_1 = \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|$.
- Frobenius范数(F-范数): $\|A\|_F = \sqrt{\mathbf{Tr}\left(AA^{\mathrm{T}}\right)} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}$.
- F-范数具有正交不变性:

  对于任意的正交矩阵$\mathcal{U} \in \mathbb{R}^{m\times m}$和$\mathcal{V} \in \mathbb{R}^{n\times n}$, 有

$$\|\mathcal{U}A\mathcal{V}\|_F^2 = \|A\|_F^2 .$$

- 矩阵迹的性质：$\mathbf{Tr}\left(AB\right) = \mathbf{Tr}\left(BA\right)$.

- 矩阵的内积: 设$A, B \in \mathbb{R}^{m\times n}$, 矩阵$A, B$的内积定义为

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

- F-范数的Cauchy不等式: 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.

- 矩阵范数的自相容性：如果对于可乘的有限维矩阵A, B, 有

$$\|AB\| \leqslant \|A\| \|B\|,$$

则称矩阵范数$\|\cdot\|$是自相容的.

Theorem 1. *矩阵的ℓ_1范数和F-范数都是自相容的.*

矩阵的算子范数

Definition 3. *给定矩阵*$A \in \mathbb{R}^{m\times n}$, $\mathbb{R}^m$*中的向量范数*$\|\cdot\|_{(m)}$*和*$\mathbb{R}^n$*中的向量范数*$\|\cdot\|_{(n)}$, *其诱导的矩阵范数为*

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

将$\|\cdot\|_{(m)}$和$\|\cdot\|_{(n)}$取向量的ℓ_p范数, 可诱导下面的矩阵的p范数.

- A的1-范数：$\|A\|_1 = \max\limits_{\|x\|_1=1} \|Ax\|_1 = \max\limits_{1\leqslant j\leqslant n} \sum_{i=1}^m |a_{ij}|$.
- A的谱范数: $\|A\|_2 = \max\limits_{\|x\|_2=1} \|Ax\|_2 = \sqrt{\lambda_{\max}(A^{\mathrm{T}}A)}$.
- A的∞范数：$\|A\|_\infty = \max\limits_{\|x\|_\infty=1} \|Ax\|_\infty = \max\limits_{1\leq i\leq m} \sum_{j=1}^n |a_{ij}|$.

矩阵算子范数的性质

矩阵算子范数的相容性: $\|Ax\|_{(m)} \leq \|A\|_{(m,n)} \|x\|_{(n)}$.

具体地说,

$$\|Ax\|_2 \leq \|A\|_2 \|x\|_2 .$$

Theorem 2. *设$A \in \mathbb{R}^{n\times n}$, 则对$A$的谱范数$\|A\|_2 = \sqrt{\lambda_{\max}(A^{\mathrm{T}}A)}$,*

- $\|A\|_2^2 = \|A^{\mathrm{T}}\|_2^2 = \|A^{\mathrm{T}}A\|_2 = \|AA^{\mathrm{T}}\|_2 .$

- *对于任意n阶酉矩阵C, D有*

$$\|CA\|_2 = \|AD\|_2 = \|CAD\|_2 = \|A\|_2 .$$

凸集

- 仿射集: $x_1, x_2 \in \mathcal{C} \Rightarrow \theta x_1 + (1-\theta) x_2 \in \mathcal{C}, \quad \forall\, \theta \in \mathbb{R}.$
- 凸集: $x_1, x_2 \in \mathcal{C} \Rightarrow \theta x_1 + (1-\theta) x_2 \in \mathcal{C}, \forall 0 \leqslant \theta \leqslant 1.$
- 仿射集当然都是凸集.
- 若$\mathcal{S}$是凸集, 则$k\mathcal{S} = \{ks | k \in \mathbb{R}, s \in \mathcal{S}\}$是凸集.
- 若$\mathcal{S}$和$\mathcal{T}$均是凸集, 则$\mathcal{S} + \mathcal{T} = \{s + t | s \in \mathcal{S}, t \in \mathcal{T}\}$是凸集.
- 若$\mathcal{S}$和$\mathcal{T}$均是凸集, 则$\mathcal{S} \cap \mathcal{T}$是凸集.
- 设$\mathcal{S}$是凸集, 则$\mathring{\mathcal{S}}, \bar{\mathcal{S}}$均是凸集.

保凸运算

- 仿射变换的保凸性: 设$f:\mathbb{R}^n\to\mathbb{R}^m$是仿射变换, 即$f(x)=Ax+b$, $A\in\mathbb{R}^{m\times n}, b\in\mathbb{R}^m$, 则

  (i) 凸集在f下的像是凸集:

$$S\subseteq\mathbb{R}^n\text{是凸集}\Rightarrow f(\mathcal{S})=\{f(x)|x\in\mathcal{S}\}\text{是凸集}.$$

  (ii) 凸集在f下的原像是凸集:

$$\mathcal{C}\subseteq\mathbb{R}^m\text{是凸集}\Rightarrow f^{-1}(\mathcal{C})=\{x|f(x)\in\mathcal{C}\}\text{是凸集}.$$

- 透视变换的保凸性: 集合$\{(x,t)\,|x\in\mathbb{R}^n,t>0\}$的透视变换所得的集合$\{\frac{x}{t}|x\in\mathbb{R}^n,t>0\}$ 是凸集.

- 分式线性变换的保凸性: 若集合$X = \{x|x \in \mathbb{R}^n\}$是凸集, 则其分式线性变换

$$f(x) = \left\{ \frac{Ax+b}{c^{\mathrm{T}}x+d} | x \in \mathbb{R}^n, A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, c \in \mathbb{R}^n, d \in \mathbb{R}, c^{\mathrm{T}}x+d > 0 \right\}$$

也是凸集.

注: 分式线性变换不是线性变换. 可先利用仿射变换, 再利用透视变换, 可以证明分式线性变换的保凸性确实成立. 保凸运算的复合仍然是保凸运算.

光滑函数性质

Theorem 3 (泰勒展开). *设$f:\mathbb{R}^n\to\mathbb{R}$连续可微, $p\in\mathbb{R}^n$, 则$\exists t\in(0,1)$ 使得*

$$f(x+p)=f(x)+\nabla f(x+tp)^T p.$$

若f二阶连续可微, 则$\exists t\in(0,1)$ 使得

$$\nabla f(x+p)=\nabla f(x)+\int_0^1 \nabla^2 f(x+tp)p\,\mathbf{d}t,$$

$$f(x+p)=f(x)+\nabla f(x)^T p+\frac{1}{2}p^T\nabla^2 f(x+tp)p.$$

L-光滑函数

Definition 4 (梯度利普希茨连续). *设f连续可微, 若存在$L > 0$, 对$\forall x, y \in \mathrm{dom}\ f$有*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|,$$

则称f是L-光滑函数.

- 二次上界: 设f是L-光滑函数且$\mathrm{dom}\ f = \mathbb{R}^n$, 则函数$f$有二次上界:

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

- 设f可微, $\mathrm{dom}\ f = \mathbb{R}^n$, 且存在一个全局极小点$x^*$. 若$f$是L-光滑的，则对$\forall x \in \mathrm{dom}\ f$有

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

适当函数和凸函数

Definition 5 (适当函数). 给定*广义实值函数* $f : \mathbb{R}^n \to \overline{\mathbb{R}} \overset{\text{def}}{=\!=} \mathbb{R} \cup \{\pm\infty\}$ *和非空集合*$\mathcal{X}$. *如果存在*$x \in \mathcal{X}$ *使得*$f(x) < +\infty$，*并且对任意的*$x \in \mathcal{X}$，*都有*$f(x) > -\infty$，*那么称函数*f *关于集合*$\mathcal{X}$ *是适当的.*

Definition 6. *设*f *为适当函数，若*$\operatorname{dom} f$ *是凸集，*

- f *是凸函数: 若对所有*$x, y \in \operatorname{dom} f$*和*$0 \le \theta \le 1$ *有*

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

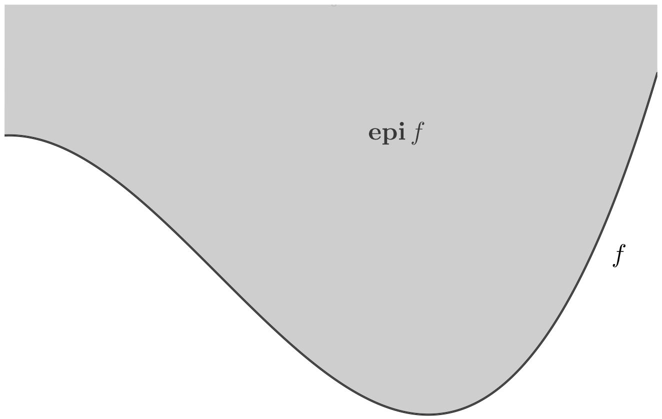
- f *是严格凸函数: 若对所有*$x, y \in \operatorname{dom} f$*和*$x \ne y, 0 < \theta < 1$，*有*

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

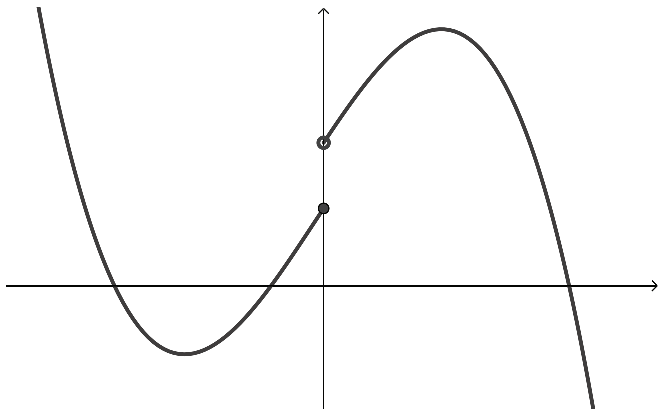
下水平集和上方图

Definition 7. *设广义实值函数$f : \mathbb{R}^n \to \overline{\mathbb{R}}$.*

- *f 的α-下水平集:* $C_\alpha = \{x \mid f(x) \le \alpha\}$.
- *f 的上方图:* $\text{epi } f = \{(x,t) \in \mathbb{R}^{n+1} \mid f(x) \le t\}$.
- *f为闭函数:* *epi f为闭集.*
- *f为下半连续函数:* *对任意的$x \in \mathbb{R}^n$，有*$\liminf_{y \to x} f(y) \ge f(x)$.



(a) 上方图epi$f$



(b) 下半连续函数$f(x)$

闭函数与下半连续函数

Theorem 4. *设广义实值函数*$f:\mathbb{R}^n\to\overline{\mathbb{R}}$*，则下列命题等价：*

① $f(x)$*的任意*α*-下水平集都是闭集；*

② $f(x)$*是下半连续的；*

③ $f(x)$*是闭函数.*

下半连续函数的性质：

- 加法: 若f与g均为适当的下半连续函数，且$\mathrm{dom} f\cap\mathrm{dom} g\neq\emptyset$，则$f+g$也是下半连续函数.
- 仿射函数的复合: 若f为下半连续函数, 则$f(Ax+b)$也为下半连续函数；
- 上确界: 若$f_\lambda,\ \lambda\in\Lambda$均为下半连续函数，则$\sup_{\lambda\in\Lambda} f_\lambda(x)$也为下半连续函数.

凸函数的性质

- 设f为凸函数, 则f的所有α-下水平集都是凸集.

- **Jensen不等式**: 设f 是凸函数，则对于$1 \le i \le m, x_i \in \mathrm{dom} f$, $0 \le \theta_i \le 1$ 且$\sum_{i=1}^m \theta_i = 1$, 有

$$f(\sum_{i=1}^m \theta_i x_i) \le \sum_{i=1}^m \theta_i f(x_i).$$

- **概率Jensen 不等式**: 设f 是凸函数，则对任意随机变量z,

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

- 设$f: \mathbb{R}^n \to (-\infty, +\infty]$为凸函数，则$f$在$\mathrm{intdom} f$连续.

凸函数的判定

- 函数$f(x)$为凸函数当且仅当其上方图epi$f$是凸集.
- 设f为可微函数且dom$f$是凸集，则f 是凸函数当且仅当

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) \quad \forall x, y \in \mathrm{dom} f. \tag{1}$$

$$(\nabla f(x) - \nabla f(y))^{\mathrm{T}}(x - y) \geq 0, \quad \forall\, x, y \in \textbf{dom}\ f. \tag{2}$$

f 是严格凸函数当且仅当(1)或(2)对所有$x, y \in \mathrm{dom} f$且$x \neq y$严格成立.

- 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.

Theorem 5. *f是凸函数当且仅当对$\forall x \in \mathrm{dom} f, v \in \mathbb{R}^n$，函数$g : \mathbb{R} \to \mathbb{R}$,*

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

是凸函数

【例】函数$f(X) = -\log\det X \ \ (\mathrm{dom}\, f = \mathbb{S}^n_{++})$是凸函数.

任取$X \succ 0, V \in \mathbb{S}^n$以及$t \in \mathbb{R}$满足$X + tV \succ 0$, 则

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

强凸函数

- 强凸函数: 若存在常数$m > 0$, 使得$g(x) = f(x) - \frac{m}{2}\|x\|^2$为凸函数.
- 强凸函数: 若存在常数$m > 0$,使得对$\forall x, y \in \mathrm{dom} f, \theta \in (0,1)$有

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta)f(y) - \frac{m}{2}\theta(1-\theta)\|x-y\|^2.$$

- 设f为可微函数且$\mathrm{dom} f$是凸集，则f是m-强凸函数当且仅当

$$(\nabla f(x) - \nabla f(y))^{\mathrm{T}}(x-y) \ge m\|x-y\|^2, \quad \forall\, x, y \in \textbf{dom}\ f.$$

- 二次下界: 设f为可微m-强凸函数，则f的所有α-下水平集有界,

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

保凸函数运算

- 非负加权和: 若f_1, f_2是凸函数且$\alpha_1, \alpha_2 \geq 0$, 则$\alpha_1 f_1 + \alpha_2 f_2$是凸函数.
- 与仿射函数的复合: 若f是凸函数，则$f(Ax+b)$是凸函数.

  【例】 $f(x) = \|Ax + b\|$.
- 逐点取最大值: 若$f_1, \ldots, f_m$是凸函数，则$\max\{f_1(x), \ldots, f_m(x)\}$是凸函数. 【例】$f(x) = \max_{i=1,\ldots,m}(a_i^T x + b_i)$.
- 取下确界: 若$f(x,y)$关于(x,y)整体是凸函数，C是凸集，则

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

  是凸函数.

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

- 取上确界: 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】

1. 集合C的支撑函数：$S_C(x) = \sup_{y\in C} y^T x$ 是凸函数.
2. 点x到集合C的最远距离：$f(x) = \sup_{y\in C} \|x-y\|$是凸函数.
3. 对称矩阵$X \in \mathbb{S}^n$的最大特征值

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

是凸函数.

- 与标量函数的复合: 给定函数$g:\mathbb{R}^n\to\mathbb{R}$, $h:\mathbb{R}\to\mathbb{R}$, 令$f(x)=h(g(x))$. 若$g$是凸函数且$h$是单调增凸函数，则$f$是凸函数; 若$g$是凹函数且$h$是单调减凸函数，则$f$是凸函数.

【例】 若g是凸函数，则$\exp g(x)$ 是凸函数；若g是正值凹函数，则$1/g(x)$ 是凸函数.

- 与向量函数的复合: 给定函数$g:\mathbb{R}^n\to\mathbb{R}^k$, $h:\mathbb{R}^k\to\mathbb{R}$, 令

$$f(x)=h(g(x))=h(g_1(x),g_2(x),\ldots,g_k(x)).$$

若g_i是凸函数, h是凸函数且关于每个分量单调增，则f是凸函数; 若g_i是凹函数, h是凸函数且关于每个分量单调减，则f是凸函数.

【例】 若g_i 是正值凹函数，则$\sum_{i=1}^m \ln g_i(x)$ 是凹函数; 若g_i 是凸函数，则$\ln\sum_{i=1}^m \exp g_i(x)$是凸函数.

- 透视函数: 定义$f:\mathbb{R}^n\to\mathbb{R}$ 的**透视函数** $g:\mathbb{R}^n\times\mathbb{R}\to\mathbb{R}$如下:

$$g(x,t)=tf\left(\frac{x}{t}\right),\quad \mathrm{dom}\, g=\{(x,t)|\frac{x}{t}\in\mathrm{dom}\, f, t>0\}.$$

若f 是凸函数，则g 是凸函数.

【例】

1. 相对熵函数$g(x,t)=t\log t-t\log x$是$\mathbb{R}^2_{++}$上的凸函数.

2. 若f 是凸函数，则

$$g(x)=(c^Tx+d)f\left(\frac{Ax+b}{c^Tx+d}\right)$$

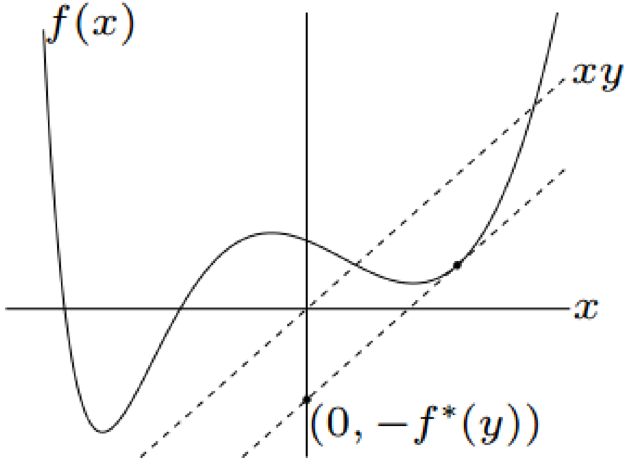
是区域$\left\{x|c^Tx+d>0,\frac{Ax+b}{c^Tx+d}\in\mathrm{dom}\, f\right\}$上的凸函数.

共轭函数

Definition 8. *适当函数f的*共轭函数*为:*

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

- f^* 恒为凸函数，无论f 是否是凸函数.
- Fenchel不等式：$f(x) + f^*(y) \geq x^T y.$



常见函数的共轭函数

- 强凸二次函数$f(x) = \frac{1}{2}x^T Qx,\ Q \in \mathbb{S}^n_{++}$的共轭函数:

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

- 凸集C的示性函数$I_C(x)$和其共轭函数$I_C^*(y)$:

$$I_C(x) = \begin{cases} 0 & x \in C, \\ +\infty & x \notin C. \end{cases}$$

$$I_C^*(y) = \sup_x \{y^T x - I_C(x)\} = \sup_{x \in C} y^T x.$$

$I_C^*(y)$恰是凸集C 的支撑函数.

- 范数$f(x) = \|x\|$的共轭函数:

$$f^*(y) = \sup_x \{y^T x - \|x\|\} = \begin{cases} 0 & \|y\|_* \le 1, \\ +\infty & \|y\|_* > 1. \end{cases}$$

对偶范数: $\|y\|_* = \sup_{\|x\| \le 1} y^T x.$

范数与对偶范数的关系: $x^T y \le \|x\| \|y\|_*, \quad \forall x, y \in \mathbb{R}^n.$

Proof. 若$\|y\|_* \le 1$, 则对$\forall x \in \mathbb{R}^n$有$x^T y \le \|x\|$, 且当$x = 0$时等号成立, 从而$f^*(y) = \sup_x \{y^T x - \|x\|\} = 0.$

若$\|y\|_* > 1$, 则至少存在一个x使得$\|x\| \le 1, \ x^T y > 1$. 从而对$\forall t > 0$,

$$f^*(y) = \sup_x \{y^T x - \|x\|\} \ge y^T (tx) - \|tx\| = t(y^T x - \|x\|), \quad (3)$$

当$t \to +\infty$时，(3)的右端趋于$+\infty$. $\square$

二次共轭函数

Definition 9. *函数f的二次共轭函数为:*

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

注:

(i) f^{**} 恒为闭凸函数，且由Fenchel不等式知

$$f^{**}(x) \leq f(x) \ \ \forall x; \quad \textbf{或等价地,} \quad \text{epi} f \subseteq \text{epi} f^{**}.$$

(ii) 若f为闭凸函数，则

$$f^{**}(x) = f(x) \ \ \forall x; \quad \textbf{或等价地,} \quad \text{epi} f = \text{epi} f^{**}.$$

Proof. 用反证法. 若$(x, f^{**}(x)) \notin \mathrm{epi} f$, 则$\exists a \in \mathbb{R}^n, b, c \in \mathbb{R}$, $(a, b) \neq 0$且$b \leq 0$ (若$b > 0$, 令$s \to +\infty$, 可推出矛盾), 使得

$$\begin{bmatrix} a \\ b \end{bmatrix}^T \begin{bmatrix} z - x \\ s - f^{**}(x) \end{bmatrix} \leq c < 0 \quad \forall (z, s) \in \mathrm{epi} f. \tag{4}$$

若$b < 0$, 在(4)中取$s = f(z)$, 则

$$a^T z + b f(z) - a^T x - b f^{**}(x) \leq c. \tag{5}$$

令$y = -\frac{a}{b}$, 由(4)得$f^*(y) - x^T y + f^{**}(x) \leq -\frac{c}{b} < 0$, which contradicts with Fenchel Inequality.

若$b = 0$, 取$\hat{y} \in \mathrm{dom}\, f^*$, 则对$\forall \varepsilon > 0$, $\forall (z, s) \in \mathrm{epi} f$, 由Fenchel不等式知, $\hat{y}^T z - s \leq \hat{y}^T z - f(z) \leq f^*(\hat{y})$, 从而有

$$\begin{bmatrix} a + \varepsilon \hat{y} \\ -\varepsilon \end{bmatrix}^T \begin{bmatrix} z - x \\ s - f^{**}(x) \end{bmatrix} \leq c + \varepsilon (f^*(\hat{y}) - x^T \hat{y} + f^{**}(x)) < 0,$$

转化为$b < 0$的情形，可推出矛盾. $\square$

次梯度

Definition 10. 设f 为适当凸函数，$x \in \mathrm{dom}\, f$.

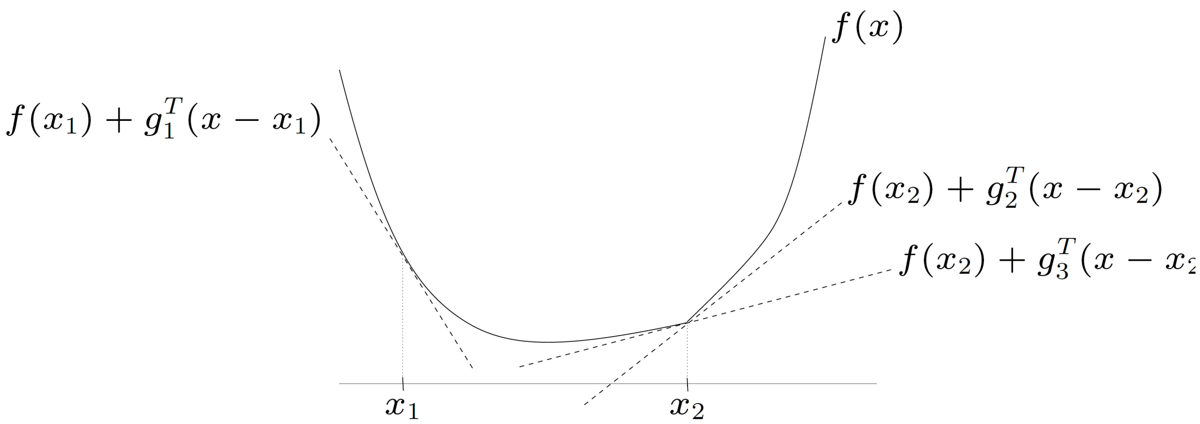
- 若向量$g \in \mathbb{R}^n$ 满足

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

  则称g 为函数f 在点x 处的一个次梯度.

- f 在点x 处的次微分:

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



- 若f 是可微凸函数, 则$\nabla f(x)$ 是f 在点x 处的一个次梯度.
- 次梯度g可提供$f(y)$ 的一个全局下界: $f(x) + g^{\mathrm{T}}(y - x)$.
- 次梯度g可诱导出上方图$\mathbf{epi}\, f$ 在点$(x, f(x))$ 处的一个支撑超平面:

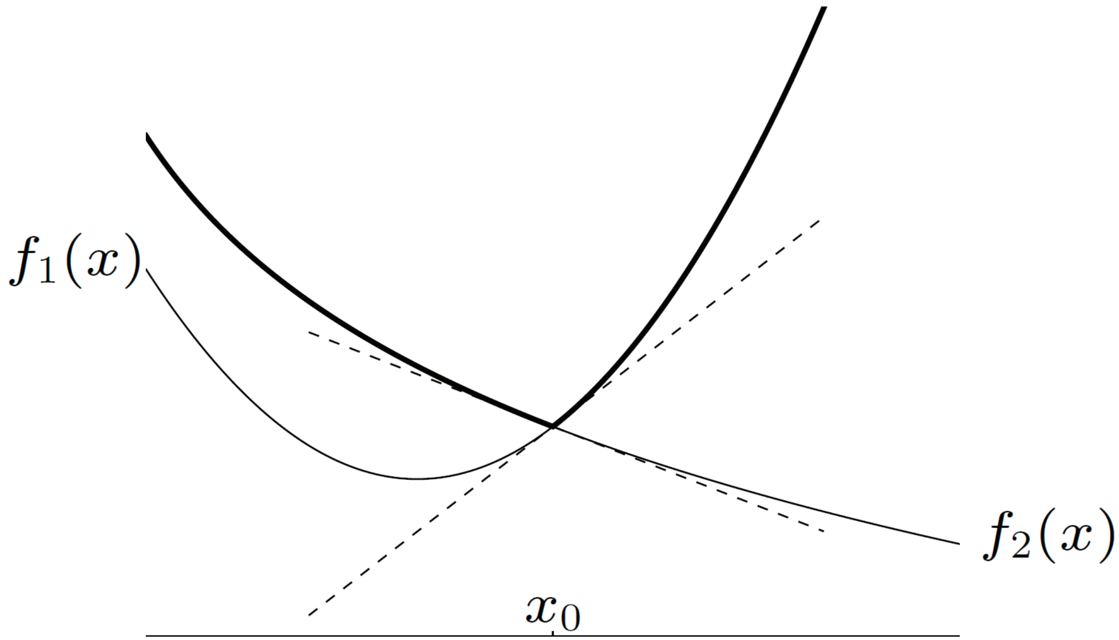
$$\begin{bmatrix} g \\ -1 \end{bmatrix}^T \left(\begin{bmatrix} y \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \leq 0 \quad \forall\, (y, t) \in \mathbf{epi}\, f.$$

- 次梯度存在性: 设f 是凸函数, 若$x \in \mathbf{intdom} f$, 则$\partial f(x) \neq \emptyset$.
- ℓ_2范数$f(x) = \|x\|_2$ 的次微分:

$$\partial \|x\|_2 = \begin{cases} \left\{ \dfrac{x}{\|x\|_2} \right\} & \textbf{if } x \neq 0, \\ \{g : \ \|g\|_2 \leq 1\} & \textbf{if } x = 0. \end{cases}$$

- 绝对值函数$f(x) = |x|$在点$x = 0$处的次微分: $\partial f(0) = [-1, 1]$.

【例】 $f(x) = \max\{f_1(x),\ f_2(x)\}$ f_1, f_2 **是可微凸函数.**



- $f(x)$在点x_0 处的次梯度可取范围$[\nabla f_1(x_0), \nabla f_2(x_0)]$.
- 若$f_1(\hat{x}) > f_2(\hat{x})$, 则$f$ 在点$\hat{x}$ 处的次梯度等于$\nabla f_1(\hat{x})$.
- 若$f_1(\hat{x}) < f_2(\hat{x})$, 则$f$ 在点$\hat{x}$ 处的次梯度等于$\nabla f_2(\hat{x})$.

次梯度的性质

- 设f 是凸函数, 则对$\forall x \in \mathbf{dom}\ f$, $\partial f(x)$是闭凸集(可能为空集).
- 设f 是凸函数, 则对$\forall x \in \mathbf{intdom}\ f$, $\partial f(x)$是非空有界集.
- 设凸函数$f(x)$在$x_0 \in \mathbf{intdom}\ f$处可微, 则$\partial f(x_0) = \{\nabla f(x_0)\}$.
- 次梯度的单调性 设$f : \mathbb{R}^n \to \mathbb{R}$ 是凸函数, $x, y \in \mathbf{dom}\ f$, 则

$$(u - v)^T(x - y) \geq 0, \quad \forall u \in \partial f(x), \forall v \in \partial f(y).$$

- 次梯度的连续性 设$f(x)$ 是闭凸函数且$\partial f(\bar{x}) \neq \emptyset$. 若

$$\lim_{k\to\infty} x^k = \bar{x}, \quad g^k \in \partial f(x^k) \text{且} \lim_{k\to\infty} g^k = \bar{g},$$

则$\bar{g} \in \partial f(\bar{x})$.

凸函数的方向导数

Definition 11. *对于凸函数f，给定点$x_0 \in \mathbf{dom}\, f$ 以及方向$d \in \mathbb{R}^n$，其方向导数定义为*

$$\partial f(x_0; d) = \inf_{t>0} \frac{f(x_0 + td) - f(x_0)}{t}.$$

注: 对于可微函数, $\partial f(x_0; d) = \nabla f(x_0)^{\mathrm{T}} d.$

- 方向导数有限: 设$f(x)$为凸函数，$x_0 \in \mathbf{int\ dom}\, f$, 则对$\forall d \in \mathbb{R}^n$, $\partial f(x_0; d)$ 有限.
- 方向导数和次梯度: 设$f : \mathbb{R}^n \to (-\infty, +\infty]$ 为凸函数, $x_0 \in \mathbf{int\ dom}\, f$, 则对$\forall d \in \mathbb{R}^n$, 有

$$\partial f(x_0; d) = \max_{g \in \partial f(x_0)} g^{\mathrm{T}} d.$$

次梯度的计算

- 凸函数的非负线性组合： 设$\alpha_1, \alpha_2 \geq 0$, 凸函数$f_1, f_2$ 满足$\mathbf{int\ dom}\ f_1 \cap \mathbf{dom}\ f_2 \neq \emptyset$, 若

$$f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x), \quad x \in \mathbf{dom}\ f_1 \cap \mathbf{dom}\ f_2$$

  则$f(x)$的次微分

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

- 线性变量替换： 设h 为适当凸函数, $f(x) = h(Ax + b)$. 若存在$x^\sharp \in \mathbb{R}^m$使得$Ax^\sharp + b \in \mathbf{int\ dom}\ h$，则

$$\partial f(x) = A^{\mathrm{T}} \partial h(Ax + b), \quad \forall\, x \in \mathbf{int\ dom}\ f.$$

▶两个函数之和的次梯度

Theorem 6 (Moreau-Rockafellar定理). 设$f_1, f_2 : \mathbb{R}^n \to (-\infty, +\infty]$ 是凸函数, 则对任意的$x_0 \in \mathbb{R}^n$,

$$\partial f_1(x_0) + \partial f_2(x_0) \subseteq \partial (f_1 + f_2)(x_0).$$

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

$$\partial (f_1 + f_2)(x_0) = \partial f_1(x_0) + \partial f_2(x_0).$$

▶凸函数族最大值的次梯度

Theorem 7. 设$f_1, \ldots, f_m : \mathbb{R}^n \to (-\infty, +\infty]$ 为凸函数, 令

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

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

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

【例】 $f(x) = \max_{i=1,2,\cdots,m}\{a_i^{\mathrm{T}} x + b_i\}$的次梯度.

▶逐点上确界函数的次梯度

Theorem 8. 设$\{f_\alpha \mid \mathbb{R}^n \to (-\infty, +\infty]\}_{\alpha\in\mathcal{A}}$ 是一族凸函数,令

$$f(x) = \sup_{\alpha\in\mathcal{A}} f_\alpha(x).$$

对$x_0 \in \cap_{\alpha\in\mathcal{A}} \textbf{int dom } f_\alpha$, 定义$I(x_0) = \{\alpha \in \mathcal{A} \mid f_\alpha(x_0) = f(x_0)\}$, 则

$$\textbf{conv} \bigcup_{\alpha\in I(x_0)} \partial f_\alpha(x_0) \subseteq \partial f(x_0).$$

若$\mathcal{A}$ 是紧集且f_α 关于α 连续, 则

$$\textbf{conv} \bigcup_{\alpha\in I(x_0)} \partial f_\alpha(x_0) = \partial f(x_0).$$

▶固定分量的函数极小值的次梯度

Theorem 9. 设$h : \mathbb{R}^n \times \mathbb{R}^m \to (-\infty, +\infty]$是关于$(x, y)$的凸函数,
$f(x) = \inf_y h(x, y)$. 对$\hat{x} \in \mathbb{R}^n$, 设$\hat{y} \in \mathbb{R}^m$ 满足$h(\hat{x}, \hat{y}) = f(\hat{x})$, 且存在$g \in \mathbb{R}^n$ 使得$(g, 0) \in \partial h(\hat{x}, \hat{y})$, 则$g \in \partial f(\hat{x})$.

▶复合函数的次梯度

Theorem 10. 设$f_1, f_2, \cdots, f_m : \mathbb{R}^n \to (-\infty, +\infty]$ 为m 个凸函数, $h : \mathbb{R}^m \to (-\infty, +\infty]$ 为关于各分量单调递增的凸函数，令

$$f(x) = h(f_1(x), f_2(x), \cdots, f_m(x)).$$

设$z = (z_1, \ldots, z_m) \in \partial h(f_1(\hat{x}), \ldots, f_m(\hat{x}))$, $g_i \in \partial f_i(\hat{x})$, 则

$$g \stackrel{\text{def}}{=\!=} z_1 g_1 + z_2 g_2 + \cdots + z_m g_m \in \partial f(\hat{x}).$$

强凸函数共轭函数的性质

Theorem 11. *设$f(x)$是适当且闭的强凸函数, 强凸参数为$\mu > 0$, 则$f^*(y)$在全空间$\mathbb{R}^n$上有定义, 且是$\frac{1}{\mu}$-光滑函数.*

Proof. 对任意的$y \in \mathbb{R}^n$，$f(x) - x^{\mathrm{T}}y$是强凸函数，因此对任意的$y \in \mathbb{R}^n$，存在唯一的$x \in \mathbf{dom}\, f$，使得$f^*(y) = x^{\mathrm{T}}y - f(x)$．根据最优性条件

$$y \in \partial f(x) \Leftrightarrow f^*(y) = x^{\mathrm{T}}y - f(x).$$

由于$f(x)$是闭凸函数，二次共轭为其本身，于是对同一组x, y有

$$x^{\mathrm{T}}y - f^*(y) = f(x) = f^{**}(x) = \sup_y \left\{x^{\mathrm{T}}y - f^*(y)\right\}.$$

这说明y也使得$x^{\mathrm{T}}y - f^*(y)$取到最大值．根据一阶最优性条件，

$$x \in \partial f^*(y).$$

再根据x的唯一性容易推出$\partial f^*(y)$中只含一个元素，故$f^*(y)$可微且$\nabla f^*(y) = x$．

下证$f^*(y)$为梯度$\frac{1}{\mu}$-利普希茨连续的．对任意的y_1, y_2，存在唯一的$x_1, x_2 \in \mathbf{dom}\, f$使得

$$y_1 \in \partial f(x_1), \quad y_2 \in \partial f(x_2).$$

根据次梯度性质以及$f(x) - \frac{\mu}{2}\|x\|^2$是凸函数，

$$f(x_2) \geq f(x_1) + (y_1 - \mu x_1)^{\mathrm{T}}(x_2 - x_1),$$
$$f(x_1) \geq f(x_2) + (y_2 - \mu x_2)^{\mathrm{T}}(x_1 - x_2).$$

两式相加得

$$(y_1 - y_2)^{\mathrm{T}}(x_1 - x_2) \geq \mu\|x_1 - x_2\|^2.$$

因$x_1 = \nabla f^*(y_1), \quad x_2 = \nabla f^*(y_2)$, 故有

$$(y_1 - y_2)^{\mathrm{T}}(\nabla f^*(y_1) - \nabla f^*(y_2)) \geq \mu\|\nabla f^*(y_1) - \nabla f^*(y_2)\|^2.$$

因此，$\nabla f^*(y)$是$\dfrac{1}{\mu}$-利普希茨连续的． $\square$