第3章

凸集与凸函数

内容提要

- 凸集
- 凸函数
- 凸集与凸函数的关联
- 凸函数的性质
- 凸函数的判定条件
- 强凸函数及其性质

凸集与凸函数是最优化的核心概念，凸函数是最优化领域中最重要的一类函数，在优化算法和理论中处于中心地位.

3.1 凸集

凸集及其相关理论是凸优化的基础，在算法理论的发展中起着重要作用.

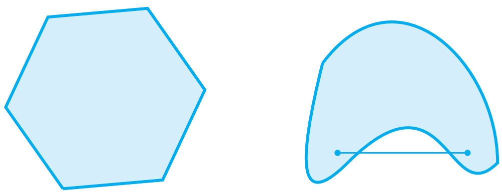
定义 3.1 (凸集)

集合 C 称为凸集，如果对任意 $\boldsymbol{x}$、$\boldsymbol{y} \in C$ 及 $0 \leqslant \lambda \leqslant 1$，都有

$$\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in C$$

♣

定义3.1表明，凸集中任意两点的连线仍然包含在集合中，这称为凸组合，如图3.1所示. 图3.1(a) 中任何两点连线都仍然在集合中，所以是凸集，图3.1(b) 中存在两点的连线超出集合范围，所以不是凸集.



(a) 凸集　　(b) 非凸集

图 3.1　凸集和非凸集

例题 3.1　使用凸集的定义，验证以下集合是凸集.

(1) 仿射空间：设 $\boldsymbol{A} \in \mathbb{R}^{m\times n}, \boldsymbol{b} \in \mathbb{R}^m$，集合 $P = \{\boldsymbol{x}|\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ 为凸集.

(2) 半空间：$H = \{\boldsymbol{x} \in \mathbb{R}^n \mid \langle \boldsymbol{a}, \boldsymbol{x}\rangle \geqslant 0\}$.

(3) 多面体：$M = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \leqslant \boldsymbol{b},\ \boldsymbol{C}\boldsymbol{x} = \boldsymbol{d}\}$.

（1）仿射空间：对于点 $\boldsymbol{x}$、$\boldsymbol{y}\in P$，都有 $\boldsymbol{A}\boldsymbol{x}=b, \boldsymbol{A}\boldsymbol{y}=\boldsymbol{b}$，因此

$$\boldsymbol{A}(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y})=\lambda\boldsymbol{A}\boldsymbol{x}+(1-\lambda)\boldsymbol{A}\boldsymbol{y}=\lambda\boldsymbol{b}+(1-\lambda)\boldsymbol{b}=\boldsymbol{b}$$

即 $\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}\in P$. 因此，$P$ 是凸集.

（2）半空间：对于点 $\boldsymbol{x}$、$\boldsymbol{y}\in P$，都有 $\langle\boldsymbol{a},\boldsymbol{x}\rangle\geqslant 0, \langle\boldsymbol{a},\boldsymbol{y}\rangle\geqslant 0$，因此

$$\langle\boldsymbol{a},\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}\rangle=\lambda\langle\boldsymbol{a},\boldsymbol{x}\rangle+(1-\lambda)\langle\boldsymbol{a},\boldsymbol{y}\rangle.$$

由于 $\langle\boldsymbol{a},\boldsymbol{x}\rangle\geqslant 0, \langle\boldsymbol{a},\boldsymbol{y}\rangle\geqslant 0$，所以 $\langle\boldsymbol{a},\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}\rangle\geqslant 0$，即 $\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}\in P$. 因此，$P$ 是凸集.

（3）多面体：对于点 $\boldsymbol{x}$、$\boldsymbol{y}\in P$，都有 $\boldsymbol{A}\boldsymbol{x}\leqslant\boldsymbol{b}, \boldsymbol{C}\boldsymbol{x}=\boldsymbol{d}$ 并且 $\boldsymbol{A}\boldsymbol{y}\leqslant\boldsymbol{b}, \boldsymbol{C}\boldsymbol{y}=\boldsymbol{d}$

$$\boldsymbol{A}(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y})=\lambda\boldsymbol{A}\boldsymbol{x}+(1-\lambda)\boldsymbol{A}\boldsymbol{y}\leqslant\lambda b+(1-\lambda)\boldsymbol{b}=\boldsymbol{b}$$

$$\boldsymbol{C}(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y})=\lambda\boldsymbol{C}\boldsymbol{x}+(1-\lambda)\boldsymbol{C}\boldsymbol{y}=\lambda\boldsymbol{d}+(1-\lambda)\boldsymbol{d}=\boldsymbol{d}$$

即 $\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}\in P$. 因此，$P$ 是凸集.

例题 3.2 球体是凸集：以 $\boldsymbol{x}_c$ 为中心，r 为半径的球体可以表示为

$$B(\boldsymbol{x}_c,r)=\{\boldsymbol{x}\in\mathbb{R}^n|\ \|\boldsymbol{x}-\boldsymbol{x}_c\|_2\leqslant r\}=\{\boldsymbol{x}\in\mathbb{R}^n|(\boldsymbol{x}-\boldsymbol{x}_c)^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{x}_c)\leqslant r^2\}$$

以原点为中心的球体为 $\{\boldsymbol{x}\in\mathbb{R}^n|\ \|\boldsymbol{x}\|_2\leqslant r\}$.

假设 $\boldsymbol{x}$、$\boldsymbol{y}\in B(0,r)$，则有 $\|\boldsymbol{x}\|\leqslant r,\quad \|\boldsymbol{y}\|\leqslant r$. 对于 $\lambda\in(0,1)$，有

$$\|\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}\|\leqslant\lambda\|\boldsymbol{x}\|+(1-\lambda)\|\boldsymbol{y}\|\leqslant\lambda r+(1-\lambda)r\leqslant r.$$

因此，$\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}\in B(0,r)$，即球体是凸集.

对于一般的球体 $B(\boldsymbol{x}_c,r)$，设 $\boldsymbol{x}$、$\boldsymbol{y}\in B(\boldsymbol{x}_c,r)$，则有

$$\begin{aligned}\|\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}-\boldsymbol{x}_c\|&=\|\lambda(\boldsymbol{x}-\boldsymbol{x}_c)+(1-\lambda)(\boldsymbol{y}-\boldsymbol{x}_c)\|\\&\leqslant\lambda\|\boldsymbol{x}-\boldsymbol{x}_c\|+(1-\lambda)\|\boldsymbol{y}-\boldsymbol{x}_c\|\\&\leqslant\lambda r+(1-\lambda)r\leqslant r\end{aligned}$$

因此，$\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}\in B(\boldsymbol{x}_c,r)$，即 $B(\boldsymbol{x}_c,r)$ 是凸集.

例题 3.3 椭球体是凸集：设 $\boldsymbol{P}$ 是正定矩阵，以 $\boldsymbol{x}_c$ 为中心的椭球体表示为

$$E=\{\boldsymbol{x}\in\mathbb{R}^n|(\boldsymbol{x}-\boldsymbol{x}_c)^{\mathrm{T}}\boldsymbol{P}^{-1}(\boldsymbol{x}-\boldsymbol{x}_c)\leqslant 1\}.$$

下面针对以原点为中心 $\boldsymbol{x}_c=0$ 的椭球体 $E=\{\boldsymbol{x}\in\mathbb{R}^n|\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{x}\leqslant 1\}$，验证 E 是凸集. 对于 $\boldsymbol{x}$、$\boldsymbol{y}\in E$ 和 $\lambda\in(0,1)$，有

$$\begin{aligned}&(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y})^{\mathrm{T}}\boldsymbol{P}^{-1}(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y})\\=&\lambda^2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{x}+(1-\lambda)^2\boldsymbol{y}^{\mathrm{T}}P^{-1}\boldsymbol{y}+2\lambda(1-\lambda)\boldsymbol{x}^{\mathrm{T}}P^{-1}\boldsymbol{y}\\\leqslant&\lambda^2+(1-\lambda)^2+2\lambda(1-\lambda)=1.\end{aligned}$$

其中不等号是由于 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{x}\leqslant 1, \boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{y}\leqslant 1$，以及 $\boldsymbol{P}^{-1}$ 内积的 Cauchy 不等式 $(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{y})^2\leqslant\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{x}\cdot\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{y}$.

凸集满足一些重要运算性质.

命题 3.1

设 C_1、C_2 为凸集，则交集 $C_1 \cap C_2$ 为凸集.

证明 令 $\boldsymbol{x}$、$\boldsymbol{y} \in C_1 \cap C_2$，则 $\boldsymbol{x}$、$\boldsymbol{y} \in C_1$ 并且 $\boldsymbol{x}$、$\boldsymbol{y} \in C_2$. 由于 C_1, C_2 为凸集，有 $\lambda_1 \boldsymbol{x} + \lambda_2 \boldsymbol{y} \in C_1$，及 $\lambda_1 \boldsymbol{x} + \lambda_2 \boldsymbol{y} \in C_2$. 所以，$\lambda_1 \boldsymbol{x} + \lambda_2 \boldsymbol{y} \in C_1 \cap C_2$，即 $C_1 \cap C_2$ 为凸集.

类似地，使用数学归纳法可证明得 $\cap_{i=1}^m C_i$ 为凸集.

在一些重要的映射下，如仿射变换，凸集的像集仍是凸集.

例题 3.4 设 $f: \mathbb{R}^n \to \mathbb{R}^m$ 是仿射变换 $f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$，其中 $\boldsymbol{A} \in \mathbb{R}^{m\times n}, \boldsymbol{b} \in \mathbb{R}^m$，则凸集在仿射变换下的像是凸集：如果 $S \subseteq \mathbb{R}^n$ 是凸集，则 $f(S) = \{f(\boldsymbol{x}) | \boldsymbol{x} \in S\}$ 是凸集.

由于 S 是凸集，对于点 $\boldsymbol{x}$、$\boldsymbol{y} \in S$ 和 $\lambda \in (0,1)$，有 $\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in S$. 因此，$f(\boldsymbol{x}) \in f(S), f(\boldsymbol{y}) \in f(S)$，并且 $f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y})$ 是像集中的一点，$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \in f(S)$. 结论成立.

使用此结论，容易验证前面椭球体 $E = \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{x}^{\mathrm{T}} \boldsymbol{P}^{-1} \boldsymbol{x} \leqslant 1\}$ 是凸集. 实际上，设 $L \in \mathbb{R}^{n\times n}$ 是下三角矩阵，满足 $\boldsymbol{P}^{-1} = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$，则 $\boldsymbol{x}^{\mathrm{T}} \boldsymbol{P}^{-1} \boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}} \boldsymbol{x} = \|\boldsymbol{L}^{\mathrm{T}} \boldsymbol{x}\|^2$. 因此，椭球体 $E = \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{x}^{\mathrm{T}} \boldsymbol{P}^{-1} \boldsymbol{x} \leqslant 1\}$ 可以看作球体 $B = \{\boldsymbol{x} \in \mathbb{R}^n | \|\boldsymbol{x}\| \leqslant 1\}$ 在线性映射 $f(\boldsymbol{x}) = \boldsymbol{L}^{\mathrm{T}} \boldsymbol{x}$ 下的像. 由于球体 B 是凸集，因此可得 E 是凸集.

同理可以验证，凸集在仿射变换下的原像是凸集：如果 $C \subseteq \mathbb{R}^m$ 是凸集，则 $f^{-1}(C) = \{\boldsymbol{x} | f(\boldsymbol{x}) \in C\}$ 是凸集.

凸集具有很好的性质，在理论和实际中起着重要作用. 凸集的一个重要性质是，可以用超平面分离不相交的凸集.

定理 3.1

设 C、D 为两个无交集的凸集，则存在 $\boldsymbol{a} \neq 0, b$，使得

$$\boldsymbol{a}^{\mathrm{T}} \boldsymbol{x} \leqslant b, \boldsymbol{x} \in C, \quad \text{并且 } \boldsymbol{a}^{\mathrm{T}} \boldsymbol{x} \geqslant b, \boldsymbol{x} \in D,$$

超平面 $\boldsymbol{a}^{\mathrm{T}} \boldsymbol{x} = b$ 将两个凸集 C、D 分离在平面两侧.

凸集分离定理表明，如果要划分 $\mathbb{R}^n$ 中的 2 个凸集，只求得一个适当的超平面即可. 凸集分离定理在凸优化理论中有重要作用，是证明 KKT 条件的基础.

定义 3.2

集合 K 称为锥，如果对任意 $\boldsymbol{x} \in K, \alpha > 0$，都有 $\alpha \boldsymbol{x} \in K$.

例题 3.5 所有的半正定矩阵构成锥 $S_+ = \{\boldsymbol{A} \in \mathbb{R}^{n\times n} \mid \boldsymbol{A} \geqslant 0\}$. 设 $\boldsymbol{A} \in S_+$ 为半正定矩阵，则对于任意非零向量 $\boldsymbol{v} \in \mathbb{R}^n$，都有 $\boldsymbol{v}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{v} \geqslant 0$. 对于 $\alpha > 0$，$\boldsymbol{v}^{\mathrm{T}} (\alpha \boldsymbol{A}) \boldsymbol{v} \geqslant 0$，所以 $\alpha \boldsymbol{A}$ 是半正定矩阵，集合 S_+ 构成锥.

同时，S_+ 是凸集. 记所有对称矩阵集合为 $S = \{\boldsymbol{A} \in \mathbb{R}^{n\times n} \mid \boldsymbol{A}^{\mathrm{T}} = \boldsymbol{A}\}$，容易验证 S 是凸集. 半正定矩阵集合 S_+ 可以看作 S 与半空间的交集

$$S_+^n = \bigcap_{x \in \mathbb{R}^n \setminus \{0\}} \{\boldsymbol{A} \in S_n \mid \boldsymbol{x}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{x} \geqslant 0\}.$$

因此, 根据结论, 凸集的交集是凸集, 可得 S_+ 为凸集.

3.2 凸 函 数

凸函数是最优化领域中最重要的一类函数, 在优化算法和理论中处于中心地位. 函数的凸性质保证了所有的局部极小值点都是全局极小值点. 因此, 设计算法时搜索找到局部极小值点即可, 这为优化算法设计带来便利.

定义 3.3 (凸函数)

设集合 Ω 为凸集, 函数 $f:\Omega\subseteq\mathbb{R}^n\to\mathbb{R}$ 称为凸函数, 如果满足: 对于任意 $\boldsymbol{x},\boldsymbol{y}\in D$ 以及 $\lambda\in[0,1]$, 都有

$$f(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y})\leqslant\lambda f(\boldsymbol{x})+(1-\lambda)f(\boldsymbol{y}).\tag{3.1}$$

如果对 $\lambda\in(0,1)$, 上述不等式严格成立, 则称 f 为严格凸函数. ♣

如果不等式(3.1)中的不等号反向, 则成为凹函数. 凸函数的几何意义如图3.2所示. 对于凸函数 f, 连接 $(\boldsymbol{x},f(\boldsymbol{x}))$ 与 $(\boldsymbol{y},f(\boldsymbol{y}))$ 的线段总是在函数 f 的曲线之上.

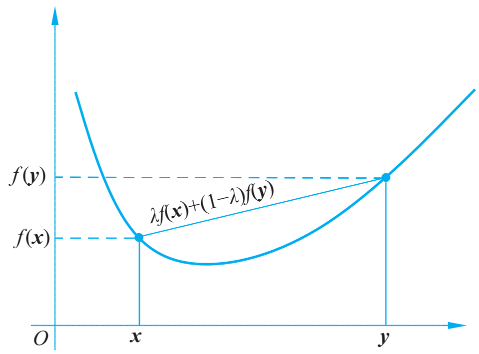


图 3.2 凸函数

例题 3.6 设 $\|\cdot\|:V\to\mathbb{R}$ 是线性空间 V 上的范数, 范数是凸函数. 实际上, 对于任意 $\boldsymbol{x}$、$\boldsymbol{y}\in V$, $\lambda\in[0,1]$, 都有

$$\|\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}\|\leqslant\|\lambda\boldsymbol{x}\|+\|(1-\lambda)\boldsymbol{y}\|=\lambda\|\boldsymbol{x}\|+(1-\lambda)\|\boldsymbol{y}\|.$$

其中不等号是由于三角不等式, 最后等号是由于范数对向量数乘的线性特性.

例题 3.7 如果 $f_i:\mathbb{R}^n\to\mathbb{R}$ 是凸函数, $i=1,2,\cdots,m$, 则 $f(\boldsymbol{x})=\sum\limits_{i=1}^{m}f_i(\boldsymbol{x})$ 是凸函数. 对于任意 $\boldsymbol{x}$、$\boldsymbol{y}\in V$, $\lambda\in[0,1]$, 都有

$$\begin{aligned}f(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y})&=\sum_{i=1}^{m}f_i(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y})\\&\leqslant\sum_{i=1}^{m}(\lambda f_i(\boldsymbol{x})+(1-\lambda)f_i(\boldsymbol{y}))\end{aligned}$$

$$
\begin{aligned}
&= \lambda \sum_{i=1}^{m} f_i(\boldsymbol{x}) + (1-\lambda) \sum_{i=1}^{m} f_i(\boldsymbol{y}) \\
&= \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}).
\end{aligned}
$$

因此 $f(\boldsymbol{x})$ 是凸函数.

例题 3.8 设 $S \subseteq \mathbb{R}^n$ 为凸集, $\boldsymbol{x} \in \mathbb{R}^n$. 定义点到集合的最远距离为

$$
d_S(\boldsymbol{x}) = \sup_{\boldsymbol{y}\in S} \|\boldsymbol{x}-\boldsymbol{y}\|.
$$

证明: $d_S(\boldsymbol{x})$ 是凸函数.

证明 对于 $\boldsymbol{u}$、$\boldsymbol{v} \in \mathbb{R}^n, \lambda \in [0,1]$, 都有

$$
\begin{aligned}
d_S(\lambda \boldsymbol{u} + (1-\lambda)\boldsymbol{v}) &= \sup_{\boldsymbol{y}\in C} \|\lambda \boldsymbol{u} + (1-\lambda)\boldsymbol{v} - \boldsymbol{y}\| \\
&= \sup_{\boldsymbol{y}\in C} \|\lambda(\boldsymbol{u}-\boldsymbol{y}) + (1-\lambda)(\boldsymbol{v}-\boldsymbol{y})\| \\
&\leqslant \sup_{\boldsymbol{y}\in C} (\lambda\|\boldsymbol{u}-\boldsymbol{y}\| + (1-\lambda)\|\boldsymbol{v}-\boldsymbol{y}\|) \\
&\leqslant \lambda \sup_{\boldsymbol{y}\in C} \|\boldsymbol{u}-\boldsymbol{y}\| + (1-\lambda) \sup_{\boldsymbol{y}\in C} \|\boldsymbol{v}-\boldsymbol{y}\| \\
&= \lambda d_S(\boldsymbol{u}) + (1-\lambda) d_S(\boldsymbol{v}).
\end{aligned}
$$

因此, $d_S(\boldsymbol{x})$ 对 $\boldsymbol{x}$ 是凸函数.

例题 3.9 设 $\boldsymbol{X} \in S^n$ 为对称矩阵, $f(\boldsymbol{X}) = \max\limits_{\|\boldsymbol{u}\|=1} \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u}, \boldsymbol{u} \in \mathbb{R}^n$, 证明 f 是 $\boldsymbol{X}$ 的凸函数.

证明 对于 $\boldsymbol{X}$、$\boldsymbol{Y} \in \mathcal{S}^n, \lambda \in [0,1]$, 都有

$$
\begin{aligned}
f(\lambda \boldsymbol{X} + (1-\lambda)\boldsymbol{Y}) &= \max_{\|\boldsymbol{u}\|=1} \boldsymbol{u}^{\mathrm{T}}(\lambda \boldsymbol{X} + (1-\lambda)\boldsymbol{Y})\boldsymbol{u} \\
&= \max_{\|\boldsymbol{u}\|=1} (\lambda \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u} + (1-\lambda)\boldsymbol{u}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{u}) \\
&\leqslant \lambda \max_{\|\boldsymbol{u}\|=1} \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u} + (1-\lambda) \max_{\|\boldsymbol{u}\|=1} \boldsymbol{u}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{u} \\
&= \lambda f(\boldsymbol{X}) + (1-\lambda) f(\boldsymbol{Y}).
\end{aligned}
$$

其中第三行的不等号是由于max函数的性质.

注意, 上述证明中不涉及向量范数的具体形式和性质. 因此, 对于任意给定的向量范数, 函数 $f(\boldsymbol{X})$ 都是凸函数.

例题 3.10 若 f 是凸函数, 则 $f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})$ 是凸函数.

证明 由于 f 是凸函数, 因此

$$
f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leqslant \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}).
$$

所以,

$$
\begin{aligned}
f(\boldsymbol{A}(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) + \boldsymbol{b}) &= f(\lambda \boldsymbol{A}\boldsymbol{x} + (1-\lambda)A\boldsymbol{y} + \boldsymbol{b}) \\
&\leqslant \lambda f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}) + (1-\lambda) f(\boldsymbol{A}\boldsymbol{y}+\boldsymbol{b})
\end{aligned}
$$

所以 $f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})$ 是凸函数.

此结论说明，线性映射与凸函数的复合是凸函数.

例题 3.11 (凸函数的复合) 设 $f: I_1 \subseteq \mathbb{R}^n \to \mathbb{R}$, $g: I_2 \to \mathbb{R}$, 并有 $f(I_1) \subseteq I_2$. 如果 f、g 是凸函数，并且 g 是增函数，则复合函数 $g \circ f$ 是 I_1 上的凸函数.

证明 由于 f 是凸函数，因此有 $f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leqslant \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$. 由于 g 是增函数，因此

$$g(f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y})) \leqslant g(\lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})).$$

再由 g 的凸性质，有

$$\begin{aligned} g(f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y})) &\leqslant g(\lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})) && (g\text{是增函数}) \\ &\leqslant \lambda g(f(\boldsymbol{x})) + (1-\lambda) g(f(\boldsymbol{y})) && (g\text{是凸函数}) \end{aligned}$$

凸函数与单调增凸函数的复合是凸函数. 例如，如果 g 是凸函数，则 $\exp g(x)$ 是凸函数.

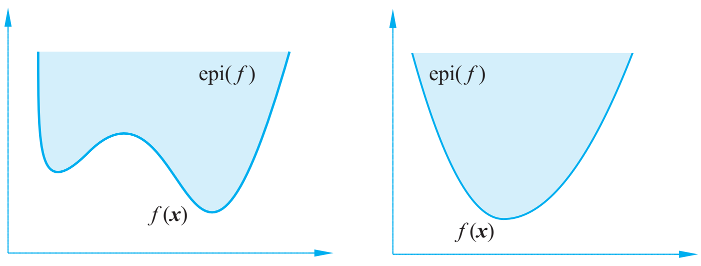
3.2.1 凸集与凸函数的关联

上方图的概念建立起集合的凸性与凸函数之间的密切联系.

定义 3.4

函数 $f: \mathbb{R}^n \to \mathbb{R}$, 集合 $\operatorname{epi}(f) = \{(\boldsymbol{x}, t) \in \mathbb{R}^{n+1} \mid f(\boldsymbol{x}) \leqslant t\}$ 称为上方图 (epigraph). ♣

函数的上方图的重要性在于，它将函数的凸性和集合的凸性联系起来，指出集合 $\operatorname{epi}(f)$ 是凸集与函数 $f(\boldsymbol{x})$ 是凸函数之间具有密切关联. 非凸函数与凸函数的上方图如图3.3所示. 从图3.3(b) 可以看出，如果函数 $f(\boldsymbol{x})$ 是凸集，则对应的上方图为凸集.



(a) (b)

图 3.3 非凸函数与凸函数的上方图

命题 3.2 (凸集与凸函数的关联)

函数 $f: \mathbb{R}^n \to \mathbb{R}$ 是凸函数，当且仅当集合 $\operatorname{epi}(f)$ 是凸集. ♠

证明 命题必要性和充分性可以通过定义直接证明.

必要性：若 f 为凸函数，对任意 $(\boldsymbol{x}_1, t_1)$, $(\boldsymbol{x}_2, t_2) \in \operatorname{epi}(f)$, $\lambda \in (0,1)$, 则 $f(\boldsymbol{x}_1) \leqslant t_1$, $f(\boldsymbol{x}_2) \leqslant t_2$, 因此

$$f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) \leqslant \lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2) \leqslant \lambda t_1 + (1-\lambda) t_2,$$

故 $(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2, \lambda t_1 + (1-\lambda) t_2) \in \operatorname{epi}(f)$.

充分性：若 $\operatorname{epi}(f)$ 是凸集，则对任意 $\boldsymbol{x}_1$、$\boldsymbol{x}_2 \in \mathbb{R}^n, \lambda \in [0,1]$,

$$(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2, \lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2)) \in \mathrm{epi}(f).$$

因此，

$$f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) \leqslant \lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2).$$

所以 f 是凸函数.

例题 3.12 设 f 是凸函数，则 f 的下水平集 $C_\alpha = \{\boldsymbol{x} \in \mathbb{R} | f(\boldsymbol{x}) \leqslant \alpha\}$ 为凸集.

证明 对任意 $\boldsymbol{x}$、$\boldsymbol{y} \in C_\alpha$，都有 $f(\boldsymbol{x}) \leqslant \alpha, f(\boldsymbol{y}) \leqslant \alpha$. 对任意 $\lambda \in (0,1)$，要证明 C_α 是凸集，只需要证明 $\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in C_\alpha$，即 $f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leqslant \alpha$.

根据 f 的凸性，可得

$$\begin{aligned} f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) &\leqslant \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) \\ &\leqslant \lambda \alpha + (1-\lambda)\alpha \\ &= \alpha. \end{aligned}$$

因此，C_α 为凸集.

3.2.2 凸函数的性质

定理 3.2

$f: \mathbb{R}^n \to \mathbb{R}$ 为凸函数，$\boldsymbol{x}^*$ 为其局部极小点，则 $\boldsymbol{x}^*$ 为全局极小点：对于任意 $\boldsymbol{x} \in \mathbb{R}^n$，都有 $f(\boldsymbol{x}) \geqslant f(\boldsymbol{x}^*)$.

定理3.2表明，凸函数的局部极小点即全局极小点．这是极为重要的性质，这确保了，对于凸函数，只寻找到局部极小点，即可确认为全局极小点．这就是为什么凸函数处于极为重要的位置．对于非凸函数，在局部范围内也可以用凸函数近似，并使用相应方法进行求解，获得一个局部极小值点.

证明 使用反证法证明．假设存在 $\boldsymbol{y} \in \mathbb{R}^n$，使得 $f(\boldsymbol{y}) < f(\boldsymbol{x}^*)$．对于 $\lambda \in (0,1)$，令

$$\boldsymbol{y}_0 = \lambda \boldsymbol{x}^* + (1-\lambda)\boldsymbol{y}$$

由于 f 是凸函数，因此

$$f(\boldsymbol{y}_0) = f(\lambda \boldsymbol{x}^* + (1-\lambda)\boldsymbol{y}) \leqslant \lambda f(\boldsymbol{x}^*) + (1-\lambda) f(\boldsymbol{y}) < f(\boldsymbol{x}^*).$$

选 λ 接近 1，使得 $\|\boldsymbol{y}_0 - \boldsymbol{x}^*\| < \epsilon$，这与 $\boldsymbol{x}^*$ 是局部极小点矛盾.

使用相同的方法可以证明，对于严格凸函数，任何两点的连线都不会与函数有两点之外的交点，因此最多只有一个极小值点.

例题 3.13 (Jensen 不等式) 对于凸函数，Jensen 不等式成立．反之亦然，即 Jensen 不等式是凸函数的等价条件．设 f 为凸函数，$\boldsymbol{x}_i \in X$，$\lambda_i > 0$，$\sum_{i=1}^{m} \lambda_i = 1$，则有

$$f\left(\sum_{i=1}^{m} \lambda_i \boldsymbol{x}_i\right) \leqslant \sum_{i=1}^{m} \lambda_i f(\boldsymbol{x}_i). \tag{3.2}$$

对于 $m = 2$，这就是凸函数的定义．对于一般的 $m \geqslant 3$，可以使用归纳法证明.

首先，考虑 $n = 2$ 的情况，即凸函数的定义．对于 $n > 2$，使用数学归纳法证明．假设

Jensen 不等式对 $m=k$ 成立, 即

$$f\left(\sum_{i=1}^{k}\lambda_i\boldsymbol{x}_i\right)\leqslant\sum_{i=1}^{k}\lambda_i f(\boldsymbol{x}_i)$$

下面证明不等式对 $m=k+1$ 也成立.

设 $\lambda_1,\lambda_2,\cdots,\lambda_{k+1}$ 是一组正权重, 满足 $\sum\limits_{i=1}^{k+1}\lambda_i=1$. 令 $\mu=\sum\limits_{i=1}^{k}\lambda_i$, 则有 $\lambda_{k+1}=1-\mu$. 因此

$$f\left(\sum_{i=1}^{k+1}\lambda_i\boldsymbol{x}_i\right)=f\left(\mu\left(\frac{1}{\mu}\sum_{i=1}^{k}\lambda_i\boldsymbol{x}_i\right)+(1-\mu)\boldsymbol{x}_{k+1}\right).$$

根据凸函数的定义, 有

$$f\left(\mu\left(\frac{1}{\mu}\sum_{i=1}^{k}\lambda_i\boldsymbol{x}_i\right)+(1-\mu)\boldsymbol{x}_{k+1}\right)\leqslant\mu f\left(\frac{1}{\mu}\sum_{i=1}^{k}\lambda_i\boldsymbol{x}_i\right)+(1-\mu)f(\boldsymbol{x}_{k+1}).$$

再利用归纳假设

$$f\left(\frac{1}{\mu}\sum_{i=1}^{k}\lambda_i\boldsymbol{x}_i\right)\leqslant\frac{1}{\mu}\sum_{i=1}^{k}\lambda_i f(\boldsymbol{x}_i),$$

因此, 有

$$\begin{aligned}f\left(\sum_{i=1}^{k+1}\lambda_i\boldsymbol{x}_i\right)&\leqslant\mu\left(\frac{1}{\mu}\sum_{i=1}^{k}\lambda_i f(\boldsymbol{x}_i)\right)+(1-\mu)f(\boldsymbol{x}_{k+1})\\&=\sum_{i=1}^{k}\lambda_i f(\boldsymbol{x}_i)+\lambda_{k+1}f(\boldsymbol{x}_{k+1})=\sum_{i=1}^{k+1}\lambda_i f(\boldsymbol{x}_i).\end{aligned}$$

这就证明了 Jensen 不等式对 $n=k+1$ 也成立. 因此, Jensen 不等式对任意 m 均成立.

3.2.3 凸函数的判定条件

对于很多复杂的问题, 尤其是机器学习中以向量和矩阵为变量的问题, 直接使用凸函数的定义难以判定是否为凸函数. 将函数限制在某个特定方向, 分析其凸性, 可以获得函数整体的凸性, 这就是凸函数判定定理, 它给出了分析复杂变量函数的凸性的一种可行性方案.

定理 3.3 (凸函数判定定理)

$f:\mathbb{R}^n\to\mathbb{R}$ 为凸函数, 当且仅当对任意 $\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{v}\in\mathbb{R}^n$, 函数 $g:\mathbb{R}\to\mathbb{R}$

$$g(t)=f(\boldsymbol{x}+t\boldsymbol{v}),\ \ \mathrm{dom}(g)=\{t\in\mathbb{R}|\boldsymbol{x}+t\boldsymbol{v}\in\mathbb{R}^n\}\tag{3.3}$$

对于变量 t 是凸函数.

证明 必要性: 设 $f(\boldsymbol{x})$ 是凸函数, 证明 $g(t)=f(\boldsymbol{x}+t\boldsymbol{v})$ 是凸函数.

先证明函数 g 的定义域 $\mathrm{dom}(g)$ 是凸集. 对任意的 t_1、$t_2\in\mathrm{dom}(g)$ 以及 $\theta\in(0,1)$, 有 $\boldsymbol{x}+t_1\boldsymbol{v}\in\mathbb{R}^n$ 和 $\boldsymbol{x}+t_2\boldsymbol{v}\in\mathbb{R}^n$. 由 f 的定义域 $\mathbb{R}^n$ 是凸集, 对以上两式分别乘以 $\lambda,(1-\lambda)$, 其线性组合 $\boldsymbol{x}+(\lambda t_1+(1-\lambda))t_2\boldsymbol{v}\in\mathbb{R}^n$. 这说明 $\lambda t_1+(1-\lambda)t_2\in\mathrm{dom}(g)$, 即 $\mathrm{dom}(g)$ 是凸集. 对于 $\lambda t_1+(1-\lambda)t_2\in\mathrm{dom}(g)$, 根据函数 g 与函数 f 的关系, 有

$$
\begin{aligned}
g(\lambda t_1+(1-\lambda)t_2) &= f(\boldsymbol{x}+(\lambda t_1+(1-\lambda))t_2\boldsymbol{v}) \\
&= f((\lambda(\boldsymbol{x}+t_1\boldsymbol{v})+(1-\lambda))(\boldsymbol{x}+t_2\boldsymbol{v})) \\
&\leqslant \lambda f(\boldsymbol{x}+t_1\boldsymbol{v})+(1-\lambda)f(\boldsymbol{x}+t_2\boldsymbol{v}) \\
&= \lambda g(t_1)+(1-\lambda)g(t_2).
\end{aligned}
$$

结合以上两点, 得到函数 $g(t)$ 是凸函数.

充分性: 首先 f 的定义域 $\mathbb{R}^n$ 是凸集. 设 $g(t)=f(\boldsymbol{x}+t\boldsymbol{v})$ 是凸函数, 有

$$
\begin{aligned}
g(1-\lambda) &= g(\lambda t_1+(1-\lambda)t_2) \\
&\leqslant \lambda g(t_1)+(1-\lambda)g(t_2) \\
&= \lambda g(0)+(1-\lambda)g(1) \\
&= \lambda f(\boldsymbol{x})+(1-\lambda)f(\boldsymbol{y}).
\end{aligned}
$$

在函数 $g(t)=f(\boldsymbol{x}+t\boldsymbol{v})$ 中, 令 $\boldsymbol{v}=\boldsymbol{y}-\boldsymbol{x}$, 则不等式左边为

$$
g(1-\lambda)=f(\boldsymbol{x}+(1-\lambda)(\boldsymbol{y}-\boldsymbol{x}))=f(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}).
$$

代入上述不等式, 得函数 $f(\boldsymbol{x})$ 是凸函数.

上述定理给出了对于复杂函数是凸函数的判定条件, 即使用函数 $g(t)=f(\boldsymbol{x}+t\boldsymbol{v})$ 将问题转换为单变量函数, 通过判断函数 $g(t)$ 的性质判断函数 $f(\boldsymbol{x})$ 是否为凸函数. 这为我们判断一般多变量乃至矩阵变量函数的凸性提供了有力工具.

例题 3.14 设 $\boldsymbol{X}$ 为对称正定矩阵 $\boldsymbol{X}\in\mathcal{S}_{++}^n$, 函数 $f(\boldsymbol{X})=-\log\det\boldsymbol{X}$ 是凸函数.

证明 对于任意 $\boldsymbol{X}\succ 0$ 以及方向 $\boldsymbol{V}\in\mathcal{S}_{++}^n$, 将 f 限制在直线 $\boldsymbol{X}+t\boldsymbol{V}$（$t$ 满足 $\boldsymbol{X}+t\mathbb{V}>0$）上, 那么

$$
\begin{aligned}
g(t) &= -\log\det(\boldsymbol{X}+t\boldsymbol{V}) \\
&= -\log\det\left(\boldsymbol{X}^{1/2}(\boldsymbol{I}+t\boldsymbol{X}^{-1/2}\boldsymbol{V}\boldsymbol{X}^{-1/2})\boldsymbol{X}^{1/2}\right) \\
&= -\log\det\boldsymbol{X}-\log\det(\boldsymbol{I}+t\boldsymbol{X}^{-1/2}\boldsymbol{V}\boldsymbol{X}^{-1/2}) \\
&= -\log\det\boldsymbol{X}-\sum_{i=1}^{n}\log(1+t\lambda_i)
\end{aligned}
$$

其中 λ_i 是 $\boldsymbol{X}^{-1/2}\boldsymbol{V}\boldsymbol{X}^{-1/2}$ 的第 i 个特征值, 因此 $\lambda_i>0$. 对于 $a>0$, 函数 $h(t)=-\log(1+at)$ 是凸函数. 因此, 对于不同的 $\lambda_i>0$, 上式最右边是凸函数.

因此, 对任意 $\boldsymbol{X}\succ\boldsymbol{0}$ 以及方向 $\boldsymbol{V}$, 函数 g 关于 t 是凸的, 因此 f 是凸函数.

对于一般函数, 直接使用凸函数的定义或判定定理通常较为复杂, 因此需要更简单、快捷的判定方法. 一阶条件通过建立凸函数与梯度之间的关系, 为判定凸性提供了更方便的工具.

定理 3.4 (凸函数的一阶条件)

设函数 $f:\mathbb{R}^n\to\mathbb{R}$ 是可微函数, $f(\boldsymbol{x})$ 是凸函数, 当且仅当对任意 $\boldsymbol{y}\in\mathbb{R}^n$, 都有

$$
f(\boldsymbol{y})\geqslant f(\boldsymbol{x})+\nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}). \tag{3.4}
$$

♡

定理的几何意义是函数 $f(\boldsymbol{x})$ 在点 $\boldsymbol{x}$ 处的切线始终在函数图像下方, 如图3.4所示.

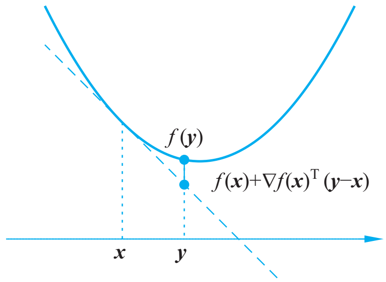


图 3.4　凸函数总是在图像上任意点的切线上方

证明　必要性: 设 f 是凸函数, 则对于任意的 $\boldsymbol{x}$、$\boldsymbol{y}\in\mathbb{R}^n$ 以及 $\lambda\in(0,1)$, 都有

$$\lambda f(\boldsymbol{y})+(1-\lambda)f(\boldsymbol{x})\geqslant f(\boldsymbol{x}+\lambda(\boldsymbol{y}-\boldsymbol{x})).$$

将上式移项, 两边同时除以 λ, 注意 $\lambda>0$, 则

$$f(\boldsymbol{y})-f(\boldsymbol{x})\geqslant\frac{f(\boldsymbol{x}+\lambda(\boldsymbol{y}-\boldsymbol{x}))-f(\boldsymbol{x})}{\lambda}.$$

令 $\lambda\to 0$, 由极限保号性可得

$$f(\boldsymbol{y})-f(\boldsymbol{x})\geqslant\lim_{t\to 0}\frac{f(\boldsymbol{x}+\lambda(\boldsymbol{y}-\boldsymbol{x}))-f(\boldsymbol{x})}{\lambda}=\nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}).$$

这里, 最后一个等式成立是由于方向导数的性质.

充分性: 对任意的 $\boldsymbol{x}$、$\boldsymbol{y}\in\mathbb{R}^n$ 以及任意的 $\lambda\in(0,1)$, 定义 $\boldsymbol{z}=\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}$, 对 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 分别在 $\boldsymbol{z}$ 处应用一阶条件, 有

$$f(\boldsymbol{x})\geqslant f(\boldsymbol{z})+\nabla f(\boldsymbol{z})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{z}),$$
$$f(\boldsymbol{y})\geqslant f(\boldsymbol{z})+\nabla f(\boldsymbol{z})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{z}).$$

将第一个不等式两边同时乘以 λ, 第二个不等式两边同时乘以 $1-\lambda$, 然后两者相加得

$$\lambda f(\boldsymbol{x})+(1-\lambda)f(\boldsymbol{y})\geqslant f(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}).$$

这正是凸函数的定义, 因此充分性成立.

单变量凸函数的一个基本性质是随着自变量的增大, 函数的导数逐渐增大, 即导数具有单调性. 此结论可以推广到多变量的情形.

例题 3.15　设 $f(\boldsymbol{x})=\dfrac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x}$, 其中 $\boldsymbol{x}\in\mathbb{R}^n$, $\boldsymbol{Q}\in\mathbb{R}^{n\times n}$ 为正定矩阵. 使用一阶条件证明 $f(\boldsymbol{x})$ 是凸函数.

证明　首先, $f(\boldsymbol{x})$ 的梯度为 $\nabla f(\boldsymbol{x})=\boldsymbol{Q}\boldsymbol{x}$. 要证明 $f(\boldsymbol{x})$ 是凸函数, 只需证明

$$f(\boldsymbol{y})-f(\boldsymbol{x})-\nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x})\geqslant 0.$$

将函数表达式和梯度代入, 直接计算可得

$$\begin{aligned}&f(\boldsymbol{y})-f(\boldsymbol{x})-\nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x})\\=&\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{y}-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}(\boldsymbol{y}-\boldsymbol{x})\\=&\frac{1}{2}\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{y}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x}-2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{y}+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x}\right)\end{aligned}$$

$$
\begin{aligned}
&=\frac{1}{2}\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{y}+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x}-2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{y}\right)\\
&\geqslant\frac{1}{2}(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{y})\geqslant 0.
\end{aligned}
$$

最后的不等号是由于Cauchy不等式和矩阵$\boldsymbol{Q}$的正定性. 结论得证.

命题 3.3 (梯度单调性)

设$f:X\subset\mathbb{R}^n\to\mathbb{R}$为可微函数, 则$f$为凸函数, 当且仅当$X$为凸集, 且对于任意$\boldsymbol{x}$、$\boldsymbol{y}\in X$

$$
(\nabla f(\boldsymbol{x})-\nabla f(\boldsymbol{y}))^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y})\geqslant 0,\ \forall\boldsymbol{x}、\boldsymbol{y}\in X. \tag{3.5}
$$

证明 必要性: 若f可微且为凸函数, 根据一阶条件, 有

$$
\begin{aligned}
f(\boldsymbol{y})&\geqslant f(\boldsymbol{x})+\nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x})\\
f(\boldsymbol{x})&\geqslant f(\boldsymbol{y})+\nabla f(\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}).
\end{aligned}
$$

将两式不等号左右两边相加, 整理可得到结论式(3.5).

充分性: 若∇f为单调映射, 构造 $g(t)=f(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})),t\in(0,1)$, 则有$g(0)=f(\boldsymbol{x}),g(1)=f(\boldsymbol{y})$, 并且$g(t)$的导数为

$$
g'(t)=\nabla f(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x})
$$

由梯度$\nabla f(\boldsymbol{x})$的单调性可知

$$
\langle\nabla f(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))-\nabla f(\boldsymbol{x}),t(\boldsymbol{y}-\boldsymbol{x})\rangle\geqslant 0
$$

两边除以t, 并展开, 可得

$$
\langle\nabla f(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})),\boldsymbol{y}-\boldsymbol{x}\rangle\geqslant\langle\nabla f(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle.
$$

根据$g(t)$的导数, 梯度为单调映射, 等价于$g'(1)\geqslant g'(0)$. 根据微积分基本定理, 并由$f(\boldsymbol{y})=g(1),f(\boldsymbol{x})=g(0)$, 有

$$
\begin{aligned}
f(\boldsymbol{y})=g(1)=g(0)+\int_0^1 g'(t)\mathrm{d}t&\\
\geqslant g(0)+g'(0)&=f(\boldsymbol{x})+\nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}).
\end{aligned}
$$

此即凸函数的一阶条件.

定理3.5给出了凸函数判定的二阶条件. 与一阶条件相比, 二阶条件要求目标函数具备二阶连续可微性, 并通过利用其 Hessian 矩阵的二阶信息来判定凸性.

定理 3.5 (凸函数的二阶条件)

设$f:\mathbb{R}^n\to\mathbb{R}$是二阶连续可微函数,

(1) f 是凸函数, 当且仅当$\nabla^2 f(\boldsymbol{x})$是半正定矩阵

$$
\nabla^2 f(\boldsymbol{x})\succeq\mathbf{0},\forall\boldsymbol{x}\in\mathbb{R}^n. \tag{3.6}
$$

(2) 如果 $\nabla^2 f(\boldsymbol{x})\succ\mathbf{0},\forall\boldsymbol{x}\in\mathbb{R}^n$, 则$f$是严格凸函数.

证明 必要性: 已知 f 是凸函数, 证明 Hessian 矩阵 $\nabla^2 f(\boldsymbol{x})$ 是半正定矩阵. 令 $g(t) = f(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))$, 其导数和二阶导数为

$$g'(t)=\nabla f(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}),\quad g''(t)=(\boldsymbol{y}-\boldsymbol{x})^{\mathrm{T}}\nabla^2 f(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))(\boldsymbol{y}-\boldsymbol{x})$$

首先证明, 如果 f 是凸函数, 总有 $g''(0)\geqslant 0$. 注意到, 对于所有足够小的 δ, 都有

$$\begin{aligned}\frac{1}{\delta}(g'(\delta)-g'(0)) &= \frac{1}{\delta}(\nabla f(\boldsymbol{x}+\delta\boldsymbol{v})-\nabla f(\boldsymbol{x}))^{\mathrm{T}}\boldsymbol{v}\\ &= \frac{1}{\delta^2}\left((\nabla f(\boldsymbol{x}+\delta\boldsymbol{v})-\nabla f(\boldsymbol{x}))^{\mathrm{T}}\delta\boldsymbol{v}\right)\\ &\geqslant 0.\end{aligned}$$

第二行实际是凸函数 f 的梯度单调性表达式. 根据命题3.3, 对于凸函数 f, 最后的不等号成立. 因此, 根据极限的保号性, $g''(0)=\lim\limits_{\delta\to 0}\frac{1}{\delta}(g'(\delta)-g'(0))\geqslant 0$.

为证明 $\nabla^2 f(\boldsymbol{x})\succeq \mathbf{0}$, 对于任意 $\boldsymbol{v}\in\mathbb{R}^n$, 都存在 $\boldsymbol{y}\in\mathbb{R}$, 使得 $\boldsymbol{v}=\boldsymbol{y}-\boldsymbol{x}$, 于是 $\boldsymbol{v}^{\mathrm{T}}\nabla^2 f(\boldsymbol{x})\boldsymbol{v}=g''(0)\geqslant 0$. 由此即可得, 矩阵 $\nabla^2 f(\boldsymbol{x})\succeq\mathbf{0}$.

充分性: 已知 $\nabla^2 f(\boldsymbol{x})\succeq\mathbf{0}$ 是半正定矩阵, 证明 f 是凸函数. 对于函数 $g(t)$, 对 $g'(t)$ 在 $[0,1]$ 上使用微分中值定理, 存在 $\xi\in(0,1)$, 使得

$$g'(1)-g'(0)=g''(\xi).$$

代入 $g(t)$ 与函数 $f(\boldsymbol{x})$ 的关系, 可得

$$\begin{aligned}(\nabla f(\boldsymbol{y})^{\mathrm{T}}-\nabla f(\boldsymbol{x})^{\mathrm{T}})\boldsymbol{v} &= \nabla f(\boldsymbol{y})^{\mathrm{T}}\boldsymbol{v}-\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{v}\\ &= \boldsymbol{v}^{\mathrm{T}}\nabla^2 f(\boldsymbol{x}+\xi\boldsymbol{v})\boldsymbol{v}\\ &\geqslant 0.\end{aligned}$$

最后的不等号是由于 $\nabla^2 f(\boldsymbol{z})\geqslant 0$ 对于所有 $\boldsymbol{z}\in\mathbb{R}^n$ 都成立. 由此证明了梯度的单调性, 进而根据命题3.3得 f 是凸函数.

直观地, 二阶可微函数 f 在局部范围可以用二次函数近似, 此时函数的凸性就对应了该二次函数的凸性. 对于二次函数 $f(\boldsymbol{x})=\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x}$, 其 Hessian 矩阵 $\nabla^2 f(\boldsymbol{x})=\boldsymbol{Q}$, 因此 f 为凸函数, 当且仅当 $\boldsymbol{Q}\succeq\mathbf{0}$.

例题 3.16 (Log-Sum-Exp 函数) 考虑函数 $f(\boldsymbol{x})=\log\sum\limits_{i=1}^{n}\exp(x_i)$, 计算其 Hessian 矩阵, 并判断其凸性.

首先计算函数的梯度. 由于 $\frac{\partial f}{\partial x_i}=\exp(x_i)/K$, 其中 $K=\sum\limits_{i=1}^{n}\mathrm{e}^{x_i}$. 因此梯度可以写成

$$\nabla f(\boldsymbol{x})=\boldsymbol{z},\quad \boldsymbol{z}=\frac{1}{K}(\mathrm{e}^{x_1},\mathrm{e}^{x_2},\cdots,\mathrm{e}^{x_n})^{\mathrm{T}}.$$

在此基础上, 对梯度的每个分量计算偏导数, 可得

$$\frac{\partial^2 f}{\partial x_i\partial x_j}=\begin{cases}\dfrac{\mathrm{e}^{x_i}}{K^2}-\dfrac{(\mathrm{e}^{x_i})^2}{K^2}, & i=j.\\ -\dfrac{\mathrm{e}^{x_i}\mathrm{e}^{x_j}}{K^2}, & i\neq j.\end{cases}$$

使用上述向量 $\boldsymbol{z}$ 的记号, Hessian 矩阵可以写成

$$\nabla^2 f(\boldsymbol{x}) = \mathrm{diag}(\boldsymbol{z}) - \boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}.$$

记 $\boldsymbol{H} = \nabla^2 f(\boldsymbol{x})$, 则 $\boldsymbol{H}$ 的分量可以写成 $h_{ij} = z_i\delta_{ij} - z_i z_j$, 其中 δ_{ij} 是 Kronecker 符号. 容易验证, $h_{ii} > 0, h_{ij} < 0, i \neq j$, 并且矩阵每一行所有元素的和为零

$$\sum_{j=1}^{n} h_{ij} = \sum_{j=1}^{n}(z_i\delta_{ij} - z_i z_j) = 0.$$

最后一个等号是由于 $\sum_{i=1}^{n} z_i = 1$.

下面证明矩阵 $\boldsymbol{H}$ 是正定矩阵. 对于 $\forall \boldsymbol{u} \in \mathbb{R}^n, \boldsymbol{u} \neq 0$, 有

$$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{u} = \boldsymbol{u}^{\mathrm{T}}(\mathrm{diag}(\boldsymbol{z}) - \boldsymbol{z}\boldsymbol{z}^{\mathrm{T}})\boldsymbol{u} = \sum_{i=1}^{n} z_i u_i^2 - \sum_{j=1}^{n}(z_i u_i)^2 \geqslant 0.$$

最后的不等号是由于 Cauchy 不等式. 由此可见, Hessian 矩阵是正定矩阵. 所以函数 $f(\boldsymbol{x})$ 是凸函数.

3.2.4 强凸函数及其性质

定义 3.5 (强凸函数)

设 $f: \mathbb{R}^n \to \mathbb{R}$ 为可微函数, 如果存在 $m > 0$, 使得

$$f(\boldsymbol{y}) \geqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \boldsymbol{y} \in \mathbb{R}^n, \tag{3.7}$$

则称 $f(\boldsymbol{x})$ 为强凸函数, 其中 m 为强凸参数. ♣

不等式右边可以看作变量 $\boldsymbol{y}$ 的二次函数强凸函数的定义说明, 函数 $f(\boldsymbol{x})$ 具有二次函数下界, 如图3.5所示. 强凸的几何意义为: 对任意 $\boldsymbol{x}$ 函数, f 的图形总是在 $(\boldsymbol{x}, f(\boldsymbol{x}))$ 处相切的二次函数之上.

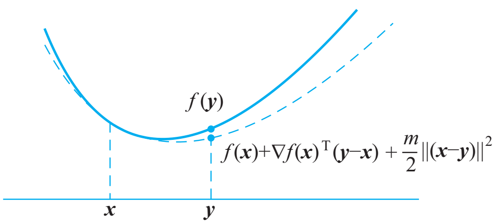


图 3.5 强凸函数具有二次函数下界

例题 3.17 考虑函数 $f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$, 其梯度为 $\nabla f(\boldsymbol{x}) = 2\boldsymbol{x}$, 则有

$$\begin{aligned} f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) &= \|\boldsymbol{y}\|^2 - \|\boldsymbol{x}\|^2 - 2\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) \\ &= \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - 2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = \|\boldsymbol{y} - \boldsymbol{x}\|^2. \end{aligned}$$

这说明, 函数 $f(\boldsymbol{x})$ 是一个强凸函数.

使用定义判定强凸函数较为困难, 对于二阶连续可微函数, 可以通过其二阶信息判定.

定理 3.6 (强凸函数的条件)

设 $f: \mathbb{R}^n \to \mathbb{R}$ 为二阶连续可微函数, $f(\boldsymbol{x})$ 是 m-强凸函数, 当且仅当 $\nabla^2 f(\boldsymbol{x}) \succeq m\boldsymbol{I}$. ♡

证明 如果 $\nabla^2 f(\boldsymbol{x}) \succeq m\boldsymbol{I}$, 则有

$$(\boldsymbol{y}-\boldsymbol{x})^{\mathrm{T}}\nabla^2 f(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}) \geqslant (\boldsymbol{y}-\boldsymbol{x})^{\mathrm{T}}(m\boldsymbol{I})(\boldsymbol{y}-\boldsymbol{x}) = m\|\boldsymbol{y}-\boldsymbol{x}\|^2 > 0.$$

代入 $f(\boldsymbol{x})$ 的 Taylor 公式中, 得

$$\begin{aligned} f(\boldsymbol{y}) &= f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})^{\mathrm{T}}\nabla^2 f(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))(\boldsymbol{y}-\boldsymbol{x}) \\ &\geqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}) + \frac{m}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2. \end{aligned}$$

因此, $f(\boldsymbol{x})$ 是 m-强凸函数. 上述证明可逆, 因此为等价条件.

例题 3.18 二次函数 $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}$, 其中 $\boldsymbol{Q}$ 是正定矩阵. 根据定理3.6, $\nabla^2 f(\boldsymbol{x}) = \boldsymbol{Q} \succeq \lambda_{\min}\boldsymbol{I}$, 其中 $\lambda_{\min}$ 是 $\boldsymbol{Q}$ 的最小特征值. 由于 $\boldsymbol{Q}$ 是正定矩阵, $\lambda_{\min} > 0$, 所以 $\nabla^2 f(\boldsymbol{x}) \succ 0$, 因此 $f(\boldsymbol{x})$ 是强凸函数, 并且 $m = \lambda_{\min}$.

可以证明, 对于强凸函数, 极小点是唯一的. 设 $f(\boldsymbol{x})$ 为二阶连续可微函数, 对于二阶连续可微函数, 有 $\nabla^2 f(\boldsymbol{x}) \succeq \alpha\boldsymbol{I} \succ \boldsymbol{0}$, 因此强凸函数必然是严格凸函数, 具有唯一全局极小值点.

强凸函数与凸函数有密切联系. 可以证明如下结论.

例题 3.19 函数 $f:\mathbb{R}^n \to \mathbb{R}$ 是 m-强凸函数, 则 $h(\boldsymbol{x}) = f(\boldsymbol{x}) - \frac{m}{2}\|\boldsymbol{x}\|^2$ 是凸函数.

实际上, 直接对 $h(\boldsymbol{x})$ 计算 Hessian 矩阵, 有 $\nabla^2 h(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}) - m\boldsymbol{I}$. 根据定理3.6, 由于 $f(\boldsymbol{x})$ 是 m-强凸函数, $\nabla^2 f(\boldsymbol{x}) \succeq m\boldsymbol{I}$, 所以 $\nabla^2 h(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}) - m\boldsymbol{I} \succeq \boldsymbol{0}$. 根据定理3.5, $h(\boldsymbol{x})$ 是凸函数.

第3章练习

1. 设 $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|^2$, 证明: $S = \{\boldsymbol{x}|f(\boldsymbol{x}) \leqslant 1\}$ 是凸集.

2. 证明以下结论:

(a) 设 $f(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}$, 其中 $\boldsymbol{Q}$ 为正定矩阵. 给定 $\boldsymbol{d} \in \mathbb{R}^n, \boldsymbol{d} \neq 0$, 定义 $h(t) = f(\boldsymbol{x}+t\boldsymbol{d})$. 证明: $h(t)$ 是严格凸函数.

(b) 对于矩阵 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, 矩阵范数定义为 $\|\boldsymbol{A}\| = \max_{\|\boldsymbol{x}\|=1}\|\boldsymbol{A}\boldsymbol{x}\|$. 证明: $\|\cdot\|$ 是 $\mathbb{R}^{m\times n}$ 上的凸函数.

(c) 设 $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|^2 + \|\boldsymbol{x}\|_1$. 证明: $f(\boldsymbol{x})$ 是凸函数.

3. 设函数 $f(\boldsymbol{x}) = -\log(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x})$, 其中 $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{A} \in \mathbb{R}^{n\times n}$ 是一个对称正定矩阵. 判断 $f(\boldsymbol{x})$ 是否为凸函数.

4. 使用凸函数的一阶条件证明如下结论:

(a) $f(\boldsymbol{x}) = (\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x})^2 + (\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x})^2$ 是凸函数.

(b) 设 $f(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}$ 是凸函数, 其中 $\boldsymbol{Q}$ 为正定矩阵.

5. 使用凸函数的二阶条件证明 $f(\boldsymbol{x}) = -\sum_{k=1}^{n} x_k \log x_k$ 是凸函数, $x_k > 0$.

6. 设 $f:\mathbb{R}^n \to \mathbb{R}$ 是可微函数, 证明以下结论:

(a) f 是 m-强凸的充分必要条件是: $h(\boldsymbol{x}) = f(\boldsymbol{x}) - \frac{m}{2}\|\boldsymbol{x}\|^2$ 是凸函数.

（b）$f(\boldsymbol{x})$ 为强凸函数的充分必要条件是: 存在常数 $m>0$, 使得对于任意 $\boldsymbol{x}$、$\boldsymbol{y}\in X$ 以及 $\theta\in(0,1)$, 都有

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

（c）f 为强凸函数, 当且仅当 $\nabla f(\boldsymbol{x})$ 满足

$$(\nabla f(\boldsymbol{x})-\nabla f(\boldsymbol{y}))^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y})\geqslant m\|\boldsymbol{x}-\boldsymbol{y}\|^2,\ \forall\boldsymbol{x}、\boldsymbol{y}\in\mathbb{R}^n.$$

7. 判断线性回归的损失函数 $L(\boldsymbol{w})=\dfrac{1}{2}\sum\limits_{i=1}^{m}(y_i-\boldsymbol{w}^T\boldsymbol{x}_i)^2$ 是否为凸函数, 并给出理由.

8. 判断逻辑回归的损失函数 $L(\boldsymbol{w})=-\sum\limits_{i=1}^{m}[y_i\log(f(\boldsymbol{x}_i;\boldsymbol{w}))+(1-y_i)\log(1-f(\boldsymbol{x}_i;\boldsymbol{w}))]$ 是否为凸函数, 并给出理由.

9. 判断带有正则化的线性回归损失函数 $L(\boldsymbol{w})=\dfrac{1}{2}\sum\limits_{i=1}^{m}(y_i-\boldsymbol{w}^T\boldsymbol{x}_i)^2+\dfrac{\lambda}{2}\|\boldsymbol{w}\|^2$ 是否为凸函数, 并给出理由.

10. 判断 Softmax 函数 $\sigma(\boldsymbol{z})=\dfrac{\exp(\boldsymbol{z})}{\sum\limits_{j=1}^{k}\exp(z_j)}$ 是否为凸函数, 并给出理由.

11. 给定一个凸集 $C=\{\boldsymbol{x}\in\mathbb{R}^n\mid\boldsymbol{A}\boldsymbol{x}\leqslant\boldsymbol{b}\}$, 其中 $\boldsymbol{A}\in\mathbb{R}^{m\times n}$, $\boldsymbol{b}\in\mathbb{R}^m$. 定义函数

$$\phi(\boldsymbol{x})=-\sum_{i=1}^{m}\log(b_i-\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x}),$$

其中 $\boldsymbol{a}_i$ 是 $\boldsymbol{A}$ 的第 i 行. 判断 $\phi(\boldsymbol{x})$ 是否为凸函数, 并计算其梯度 $\nabla\phi(\boldsymbol{x})$.