凸分析基础与梯度型算法

课后习题和复习

作者：张鑫航

组织：国防科技大学

时间：2023 年 12 月 5 日

版本：1.0

联系方式：zhangxinhang19@foxmail.com

更换校徽为矢量图

目录

第 0 章 基础知识 1
第 0 章 练习 1

第 1 章 凸集与投影 4
1.1 凸集 4
第 1 章 练习 4

第 2 章 凸函数 6
2.1 凸函数 6
2.2 次微分理论 7
2.2.1 次微分与次梯度 7
2.2.2 最速下降方向 8
第 2 章 练习 9

第 3 章 最优化条件与对偶 11
3.1 最优化条件 11
3.1.1 无约束优化情形 11

第 4 章 梯度下降法 13
4.1 梯度法的描述 13
4.1.1 迭代格式 13
4.1.2 步长选取 14
4.2 收敛理论 15
4.2.1 基本收敛性质 15

参考文献 17

附录 A 基本数学工具 18
A.1 求和算子与描述统计量 18

第 0 章 基础知识

第 0 章 练习

1. 证明 Cauchy-Schwarz 不等式：

$$\langle y,x\rangle \leqslant \|y\|\|x\|, \forall x,y \in \mathbb{R}^n$$

证明 若 $x=0$ 或 $y=0$，则结论显然成立。设 $x\neq 0$，对任意数 λ，取向量 $y-\lambda x$，则有[1]

$$\langle y-\lambda x, y-\lambda x\rangle \geqslant 0$$

即 $\langle y,y\rangle + \lambda^2\langle x,x\rangle - 2\lambda\langle y,x\rangle \geqslant 0$ 由 $x\neq 0$，可取到

$$\lambda = \frac{\langle y,x\rangle}{\langle x,x\rangle}$$

代入上式，即得

$$\langle y,y\rangle\langle x,x\rangle \geqslant (\langle y,x\rangle)^2$$

从而有

$$\langle y,x\rangle \leqslant \|y\|\|x\|$$

等式成立的充要条件为

$$y-\lambda x = 0$$

即 y 和 x 线性相关。证毕！

欧氏距离的三角不等式：

证明

$$\begin{aligned}
\langle x+y, x+y\rangle &= \|x\|^2 + \|y\|^2 + 2\langle x,y\rangle \\
&\leqslant \|x\|^2 + \|y\|^2 + 2(\|x\|\|y\|)^2 \\
&= (\|x\| + \|y\|)^2
\end{aligned}$$

所以，有 $\|x+y\| \leqslant \|x\| + \|y\|$。

等式成立的充要条件为 y 和 x 线性相关。证毕！

(i) 欧氏范数 ℓ_1- 范数 $\|\cdot\|$ 的等价性：

$$\frac{1}{\sqrt{n}}\|x\| \leqslant \|x\|_\infty \leqslant \|x\| \leqslant \|x\|_1 \leqslant \sqrt{n}\cdot\|x\|_2,\ \forall x \in \mathbb{R}^n$$

证明

$$\begin{aligned}
\frac{1}{\sqrt{n}}\|x\| &= \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}} \leq \sqrt{\frac{n x_{\max}^2}{n}} = \|x\|_\infty \\
&\leq \sum_{i=1}^n \sqrt{|x|_i^2} = \|x\| \\
&\leq \sqrt{\sum_{i=1}^n |x|_i^2 + 2\sum_{i\neq j}|x|_i|x|_j} = \sqrt{(|x_1|+\cdots+|x_n|)^2} = \|x\|_1 \\
&= (1\cdot|x_1| + \cdots + 1\cdot|x_n|) \leq \sqrt{n}\cdot\|x\|_2
\end{aligned}$$

证毕！

(ii) 对任意的序列 $\{a_j\}_{j=1}^n \subset \mathbb{R}$，均有：

$$(\sum_{j=1}^n a_j)^2 + (\sum_{j=1}^n (-1)^j a_j)^2 \leqslant (n+1)\sum_{j=1}^n a_j^2.$$

证明 $n=2m$ 时，

$$\begin{aligned}&\left(\sum_{i=1}^{n} a_i\right)^2+\left(\sum_{i=1}^{n}(-1)^j a_i\right)^2=\left(\sum_{i=1}^{m} a_{2i-1}+\sum_{i=1}^{m} a_{2i}\right)^2+\left(\sum_{i=1}^{m} a_{2i}-\sum_{i=1}^{m} a_{2i-1}\right)^2\\&=2\left(\sum_{i=1}^{m} a_{2i-1}\right)^2+2\left(\sum_{i=1}^{m} a_{2i}\right)^2 \text{由柯西不等式}\\&\le 2\cdot m\sum_{i=1}^{m} a_{2i-1}^2+2\cdot m\sum_{i=1}^{m} a_{2i}^2\le(2m+1)\sum_{i=1}^{2m} a_i^2=(n+1)\sum_{i=1}^{n} a_i^2\end{aligned}$$

$n=2m+1$ 时，

$$\begin{aligned}&\left(\sum_{i=1}^{n} a_i\right)^2+\left(\sum_{i=1}^{n}(-1)^j a_i\right)^2=\left(\sum_{i=1}^{m+1} a_{2i-1}+\sum_{i=1}^{m} a_{2i}\right)^2+\left(\sum_{i=1}^{m} a_{2i}-\sum_{i=1}^{m+1} a_{2i-1}\right)^2\\&=2\left(\sum_{i=1}^{m+1} a_{2i-1}\right)^2+2\left(\sum_{i=1}^{m} a_{2i}\right)^2 \text{由柯西不等式}\\&\le 2\cdot(m+1)\sum_{i=1}^{m} a_{2i-1}^2+2\cdot m\sum_{i=1}^{m} a_{2i}^2\le(2m+2)\sum_{i=1}^{2m+1} a_i^2=(n+1)\sum_{i=1}^{n} a_i^2\end{aligned}$$

综上所述，$(\sum_{j=1}^n a_j)^2+(\sum_{j=1}^n(-1)^j a_j)^2\leqslant(n+1)\sum_{j=1}^n a_j^2$。

证毕！

2. 设 $A\in\mathbb{R}^{n\times n}$ 为实对称矩阵，$b\in\mathbb{R}^n$。令

$$f(x)=x^{\mathrm{T}}Ax-2b^{\mathrm{T}}x.$$

证明下述结论等价：

(i) $\inf\{f(x):x\in\mathbb{R}^n\}>-\infty$

(ii) $A\succeq 0$ 且 $b\in\operatorname{Im}A$

(iii) $\operatorname{Arg}\min\{f(x):x\in\mathbb{R}^n\}\neq\emptyset$

证明 *(i) ⇒ (ii)*，首先证明 $A\succeq 0$：

假设 $A\succeq 0$ 不成立，则 $\exists v\in|\mathbb{R}^n$, s.t. $v^{\mathrm{T}}Av<0$，那么可以构建极小化序列 $x=tv$ 使得

$$\begin{aligned}f(x)&=t^2v^{\mathrm{T}}Av-2tb^{\mathrm{T}}v\\&=t(tv^{\mathrm{T}}Av-2b^{\mathrm{T}}v)\end{aligned}$$

那么 $f(x)\to-\infty, t\to\infty$，与 $\inf\{f(x):x\in\mathbb{R}^n\}>-\infty$ 矛盾。

接下来证明 $b\in\operatorname{Im}A$：

假设 $b\notin\operatorname{Im}A$，那么 b 可以写为 $b=c+d$ 使得 $c\in\operatorname{Im}A, d\in(\operatorname{Im}A)^{\perp}=\operatorname{Null}A^{\mathrm{T}}=\operatorname{Null}A$

那么可以构建极小化序列 $x=kd$ 使得

$$\begin{aligned}f(x)&=k^2d^{\mathrm{T}}Ad-2kb^{\mathrm{T}}d\\&=-2kc^{\mathrm{T}}d-2kd^{\mathrm{T}}d\\&=-2k\|d\|^2\end{aligned}$$

则 $f(x)\to-\infty, k\to\infty$，与 $\inf\{f(x):x\in\mathbb{R}^n\}>-\infty$ 矛盾。

(ii) ⇒ (i)，存在正交阵 P，s.t. $A=P^{\mathrm{T}}DP$，其中 $D=\operatorname{diag}\{\lambda_1,\cdots,\lambda_n\}$ 为对角阵，不妨设 $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_n$，则

$$f(x)=x^{\mathrm{T}}P^{\mathrm{T}}DPx-2b^{\mathrm{T}}P^{\mathrm{T}}Px$$

令 $y=Px$，则 $f(x)=g(y)=y^{\mathrm{T}}Dy-2b^{\mathrm{T}}P^{\mathrm{T}}y$

设 $Pb=(b_1,\cdots,b_n)^{\mathrm{T}}$，则

$$g(y)=\sum_{i=1}^{n}\lambda_i y_i^2-2\sum_{i=1}^{n}b_iy_i=\sum_{i=1}^{n}(\lambda_i y_i^2-2b_iy_i)$$

因为 $b\in\operatorname{Im}A$，存在 $\alpha\in|\mathbb{R}^n$, s.t. $Pb=PA\alpha=PP^{\mathrm{T}}DP\alpha=DP\alpha$。由 $A\succeq 0$，有 $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_n\ge 0$，

那么当 $\lambda_i = 0$ 时，$b_i = 0$，故而

$$\inf\{f(x) : x \in \mathbb{R}^n\} = \inf\{g(y) : y \in \mathbb{R}^n\} > -\infty$$

$(ii) \Rightarrow (iii)$，进行 $(ii) \Rightarrow (i)$中的变量替换后

$$\begin{aligned}&\operatorname{Arg}\min\{g(y) : y \in \mathbb{R}^n\}\\ &= \{y = [y_1, \cdots, y_n] : (y_i = \frac{b_i}{\lambda_i}, \textbf{if}\lambda_i > 0); (y_i \in |\mathbb{R}^n \textbf{if}\lambda_i = 0)\}\end{aligned}$$

故

$$\operatorname{Arg}\min\{f(x) : x \in \mathbb{R}^n\} = \left\{x = P^{\mathrm{T}} y : y \in \operatorname{Arg}\min\{g(y) : y \in \mathbb{R}^n\}\right\}$$

$(iii) \Rightarrow (i)$，显然成立。

证毕！

3. 设 $A \in \mathbb{R}^{m\times n}$ 是一个秩为 $r > 0$ 的给定矩阵，$1 \le k \le r$ 为给定的自然数. 试求如下秩约束优化问题的解:

$$\begin{aligned}&\text{minimize} \quad \|A - X\|_F\\ &\text{subject to} \quad \operatorname{rank} X = k.\end{aligned}$$

解 $A = UDV^{\mathrm{T}} \Rightarrow U^{\mathrm{T}}AV = D$ 根据矩阵 Frobenius 范数的酉不变性

$$\begin{aligned}\|A - X\|_F &= \|U^{\mathrm{T}}(A - X)V\|_F\\ &= \|D - U^{\mathrm{T}}XV\|_F\end{aligned},$$

令 $Y = U^{\mathrm{T}}XV$ 上述优化问题可以转化为

$$\begin{aligned}&\text{minimize} \quad \|D - Y\|_F\\ &\text{subject to} \quad \operatorname{rank} Y = k.\end{aligned},$$

其中 $D = \operatorname{diag}(\sigma_1, \cdots, \sigma_r)$，$\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r$。易得，当 $Y = \Sigma_k = \operatorname{diag}(\sigma_1, \cdots, \sigma_k, \cdots, 0)$ 时，达到最优解，此时

$$X = U\Sigma_k V^{\mathrm{T}}$$

第 1 章 凸集与投影

1.1 凸集

定义 1.1 (凸集)

设 X 为 $\mathbb{R}^n$ 中的非空子集。若对任意的 $x, y \in X$，以及参数 $\alpha \in (0,1)$ 均有

$$\alpha x + (1-\alpha) y \in X, \tag{1.1}$$

则称 X 为 $\mathbb{R}^n$ 中的凸集。我们约定空集 $\emptyset$ 为凸集。

♣

例 1.1 下述集合均为凸集

1. 仿射线 $I := \{\tau x^1 + (1-\tau)x^2 : \tau \in \mathbb{R}\}$，其中 $x^1, x^2 \in \mathbb{R}^n, x^1 \neq x^2$ 给定
2. 超平面 $H := \{x \in \mathbb{R}^n : \langle a, x\rangle = b\}$，其中 $a \in \mathbb{R}^n \backslash \{0\}, b \in \mathbb{R}$
3. 半空间 $H^- := \{x \in \mathbb{R}^n : \langle a, x\rangle \subseteq b\}$，其中 $a \in \mathbb{R}^n \backslash \{0\}, b \in \mathbb{R}$
4. 范数球 $B := \{x \in \mathbb{R}^n : ||x-a|| \leq b\}$，其中 $a \in \mathbb{R}^n, b > 0$.
5. 闭椭球 $\tilde{B} := \{x \in \mathbb{R}^n : x^T Q x + 2b^T x + c \leq 0\}$，$Q \in \mathbb{R}^{n\times n}$ 半正定，$b \in \mathbb{R}^n, c \in \mathbb{R}$.
6. 半正定矩阵集 $H_+ := \{A \in \mathbb{R}^{n\times n} : \langle x, Ax\rangle \geq 0, \forall x \in \mathbb{R}^n\}$.

证明 先证明仿射线的凸性。在仿射线 I 中任取两点：$x = \tau_1 x^1 + (1-\tau_1)x^2$，$y = \tau_2 x^1 + (1-\tau_2)x^2$。对任意的参数 $\forall\alpha \in (0,1)$。我们进行如下推导：

$$\begin{aligned}
\alpha x + (1-\alpha)y &= \alpha_1\tau_1 x^1 + \alpha(1-\tau_1)x^2 + (1-\alpha)\tau_2 x^1 + (1-\alpha)(1-\tau_2)x^2 \\
&= (\alpha\tau_1 + (1-\alpha)\tau_2)x^1 + [1-(\alpha\tau_1 + (1-\alpha)\tau_2)]x^2 \\
&= \tau_3 x^1 + (1-\tau_3)x^2 \in I,
\end{aligned}$$

其中 $\tau_3 = \alpha\tau_1 + (1-\alpha)\tau_2$。因此据凸集定义可知仿射线 I 为凸集。

下面证明超平面的凸性。任取 x, $y \in H$, $\alpha \in (0,1)$，则由内积的线性性可知

$$\langle a, \alpha x + (1-\alpha)y\rangle = \alpha\langle a, x\rangle + (1-\alpha)\langle a, y\rangle = b.$$

因此 $\alpha x + (1-\alpha)y \in H$，从而超平面为凸集。半空间 H^- 和半正定矩阵集 H_+ 的凸性可类似的证明。

先证范数球的凸性。任取 $x, y \in B$, $\alpha \in (0,1)$，则由范数的三角不等式可知

$$\begin{aligned}
\|(\alpha x + (1-\alpha)y) - a\| &= \|\alpha(x-a) + (1-\alpha)(y-a)\| \\
&\leq \alpha\|x-a\| + (1-\alpha)\|y-a\| \\
&\leq \alpha b + (1-\alpha)b = b.
\end{aligned}$$

因此，$\alpha x + (1-\alpha)y \in B$，从而范数球为凸集。

最后证闭椭球的凸性。由 Q 为半正定矩阵可知存在矩阵 Γ 使得 $Q = \Gamma^{\mathrm{T}}\Gamma$。任取 $x, y \in \tilde{B}, \alpha \in (0,1)$，则有

$$\begin{aligned}
&[\alpha x + (1-\alpha)y]^{\mathrm{T}} Q[\alpha x + (1-\alpha)y] + 2b^{\mathrm{T}}[\alpha x + (1-\alpha)y] + c \\
&= \alpha(x^{\mathrm{T}}Qx + 2b^{\mathrm{T}}x + c) + (1-\alpha)(y^{\mathrm{T}}Qy + 2b^{\mathrm{T}}y + c) \\
&+ \alpha(1-\alpha)(xQ^{\mathrm{T}}y + y^{\mathrm{T}}Qx - x^{\mathrm{T}}Qx - y^{\mathrm{T}}Qy) \\
&\leq \alpha(1-\alpha)(x^{\mathrm{T}}Qy + y^{\mathrm{T}}Qx - x^{\mathrm{T}}Qx - y^{\mathrm{T}}Qy) \\
&= -\alpha(1-\alpha)\|\Gamma x - \Gamma y\|^2 \leq 0.
\end{aligned}$$

第 1 章 练习

1. 证明下述集合为非凸的

$$A = \{x \in \mathbb{R}^2 : x_1^2 - x_2^2 + x_1 + x_2 \leq 4\}$$

证明 存在 $a = (2, 2)$，$b = (2, -1)$，满足

$$2^2 - 2^2 + 2 - 2 \leq 4$$
$$2^2 - (-1)^2 + 2 - (-1) \leq 4$$

有 $\alpha = \frac{1}{2}$，使得 $y = \alpha a + (1-\alpha)b = (2, 0.5)$

$$2^2 - (0.5)^2 + 2 - (0.5) = 5.25 > 4$$

即 $y = \alpha a + (1-\alpha)b \notin A$

证毕！

2. 证明 $\mathrm{conv}\{e_1, e_2, -e_1, -e_2\} = \{x \in \mathbb{R}^2 : \|x\|_1 \leq 1\}$。其中 $e_1 = (1,0)^{\mathrm{T}}, e_2 = (0,1)^{\mathrm{T}}$

证明 首先我们证明 $\mathrm{conv}\{e_1, e_2, -e_1, -e_2\} \subseteq \{x \in \mathbb{R}^2 : \|x\|_1 \leq 1\}$。

假设 $y \in \mathrm{conv}\{e_1, e_2, -e_1, -e_2\}$，即存在 $\lambda_i \geq 0$，满足 $\sum_{i=1}^4 \lambda_i = 1$ 且

$$\begin{aligned} y &= \lambda_1 e_1 + \lambda_2 e_2 + \lambda_3(-e_1) + \lambda_4(-e_2) \\ &= (\lambda_1 - \lambda_3)e_1 + (\lambda_2 - \lambda_4)e_2 \end{aligned}$$

有

$$\begin{aligned} \|y\|_1 &= |\lambda_1 - \lambda_3| + |\lambda_2 - \lambda_4| \\ &\leq |\lambda_1 + \lambda_3| + |\lambda_2 + \lambda_4| \\ &\leq \lambda_1 + \lambda_3 + \lambda_2 + \lambda_4 = 1 \end{aligned}$$

接下来我们证明 $\{x \in \mathbb{R}^2 : \|x\|_1 \leq 1\} \subseteq \mathrm{conv}\{e_1, e_2, -e_1, -e_2\}$

假设有 $y = (y_1, y_2) \in \{x \in \mathbb{R}^2 : \|x\|_1 \leq 1\}$，满足 $|y_1| + |y_2| \leq 1$，那么有

$$\begin{aligned} y &= y_1 e_1 + y_2 e_2 \\ &= a e_1 + b e_2 - c e_1 - d e_2 \end{aligned}$$

满足

$$\begin{cases} a + b + c + d = 1 \\ a - c = y_1 \\ b - d = y_2 \\ a, b, c, d \in (0, 1) \end{cases}$$

有解。

综上所述，$\mathrm{conv}\{e_1, e_2, -e_1, -e_2\} = \{x \in \mathbb{R}^2 : \|x\|_1 \leq 1\}$。证毕！

第 2 章 凸函数

2.1 凸函数

我们引入上图的概念。设 $f:\mathbb{R}^n\to\bar{\mathbb{R}}$，则 f 的上图（epigraph）定义为：

$$\mathrm{epi}(f):=\{(x,\gamma)\in\mathbb{R}^n\times\mathbb{R}:f(x)\le\gamma\}$$

而 f 的有效域（effective domain）定义为：

$$\mathrm{dom}(f):=\{x\in\mathbb{R}^n:f(x)<+\infty\}$$

不难看出，f 的有效域为其上图在 $\mathbb{R}^n$ 空间中的投影，下面基于上图的概念给出凸函数的现代定义。

定义 2.1 (凸函数)

设 $f:\mathbb{R}^n\to\bar{\mathbb{R}}$。若 $\mathrm{epi}(f)\subset\mathbb{R}^n\times\mathbb{R}$ 为凸集，则称 f 为凸函数。若 $-f$ 为凸的，则称 f 为凸函数。♣

上述定义推广了第 1 章中一维凸函数的定义. 事实上, 我们有凸函数的如下等价条件.

定理 2.1

设 $f:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ 且 $\mathrm{dom}(f)\ne\emptyset$。则 f 为凸的当且仅当对任意 $x,y\in\mathbb{R}^n,0<\alpha<1$，均有

$$f(\alpha x+(1-\alpha)y)\le\alpha f(x)+(1-\alpha)f(y). \tag{2.1}$$

♡

证明 先证明必要性。设 f 为凸函数并任取 $x,y\in\mathbb{R}^n$，由定义可知 $\mathrm{epi}(f)$ 为凸集。若 $x,y\in\mathbb{R}^n$ 其中之一不属于 $\mathrm{dom}(f)$，则 (2.1) 显然成立。故不妨设 $x,y\in\mathrm{dom}(f)$，于是 $\begin{bmatrix}x\\f(x)\end{bmatrix}$ 和 $\begin{bmatrix}y\\f(y)\end{bmatrix}$ 均属于 $\mathrm{epi}(f)$，从而其凸组合亦属于 $\mathrm{epi}(f)$，也即

$$\begin{bmatrix}\alpha x+(1-\alpha)y\\\alpha f(x)+(1-\alpha)f(y)\end{bmatrix}\in\mathrm{epi}(f). \tag{2.2}$$

因而 $f(\alpha x+(1-\alpha)y)\le\alpha f(x)+(1-\alpha)f(y)$。

反之，上式蕴含了 (2.2) 式，从而 $\mathrm{epi}(f)$ 为凸集，故 f 为凸函数。证毕！

注

1. 若 (2.1) 式为严格不等式，则称 f 为严格凸函数
2. 限制在凸集上的凸函数可拓展为全空间的凸函数
3. 若 f 为凸，则 $\mathrm{dom}(f)$ 必为凸集

定义 2.2

设 $\Omega\subset\mathbb{R}^n$ 为凸集并设 $f:\Omega\to(-\infty,+\infty]$ 为定义在 Ω 上的拓展的实值函数。若对任意的 $x,y\in\Omega$ 及 $\alpha\in(0,1)$ 均有 (2.1) 式成立，则称 f 为 Ω 上的凸函数。♣

限制在凸集 Ω 上的凸函数 f 可以延拓为全空间中的凸函数。事实上，可定义 $\tilde{f}:\mathbb{R}^n\to(-\infty,+\infty]$ 为

$$\tilde{f}(x):=\begin{cases}f(x), & x\in\Omega;\\+\infty, & x\notin\Omega.\end{cases}$$

则不难验证：

$$\mathrm{dom}(\bar{f})=\mathrm{dom}(f)$$
$$\mathrm{epi}(\tilde{f})=\mathrm{epi}(f).$$

由此可知，f 的凸性等同于 $\tilde{f}$ 的凸性。

例 2.1 下面介绍几类常见的简单凸函数

(i) 放射函数：$f(x)=a^{\mathrm{T}}x+b$，其中 $a\in\mathbb{R}^n, b\in\mathbb{R}$.

(ii) 范数函数：$f(x)=\|x\|$，其中 $\|\cdot\|$ 为 $\mathbb{R}^n$ 上的范数

(iii) 示性函数：$\delta_\Omega(x)=\begin{cases}0, & x\in\Omega\\ +\infty, & x\notin\Omega\end{cases}$，其中 $\Omega\subset\mathbb{R}^n$ 为凸集

(iv) 距离函数:$d_\Omega=\min\limits_{z\in\Omega}\|x-z\|$，其中 $\Omega\subset\mathbb{R}^n$ 为闭凸集

证明 先应用 (2.1) 验证（ii）与（iv）。对任意的 $x,y\in\Omega$ 及 $\alpha\in(0,1)$，由范数的三角不等式及齐次性可进行如下推导：

$$
\begin{aligned}
f(\alpha x+(1-\alpha)y) &= \|\alpha x+(1-\alpha)y\| \\
&\le \|\alpha x\|+\|(1-\alpha)y\| \\
&= \alpha\|x\|+(1-\alpha)\|y\| \\
&= \alpha f(x)+(1-\alpha)f(y).
\end{aligned}
$$

从而可知范数函数 $f(x)=\|x\|$ 为凸函数。下证（iv）。设 $x,y\in\mathbb{R}^n$ 及 $\alpha\in(0,1)$，并记 $v=\prod_\Omega(x), w=\prod_\Omega(y)$，则由 Ω 的凸性以及投影的定义可知

$$\alpha v+(1-\alpha)w\in\Omega$$

且 $d_\Omega(x)=\|x-v\|, d_\Omega(y)=\|y-w\|$. 据此，

$$
\begin{aligned}
d_\Omega(\alpha x+(1-\alpha)y) &= \min_{z\in\Omega}\|\alpha x+(1-\alpha)y-z\| \\
&\le \|\alpha x+(1-\alpha)y-\alpha v-(1-\alpha)w\| \\
&\le \alpha\|x-v\|+(1-\alpha)\|y-w\| \\
&= \alpha d_\Omega(x)+(1-\alpha)d_\Omega(y).
\end{aligned}
$$

从而距离函数 $d_\Omega(x)$ 必为凸函数。

最后，（i）易证而（iii）成立是因为 $\mathrm{epi}(\delta_\Omega)=\Omega\times\mathbb{R}_+$ 为凸集。

凸函数的定义只依赖于函数值的信息. 但若给定的函数连续可微，则我们可以分别基于函数的一阶信息（梯度）和二阶信息（海森矩阵）来判断函数的凸性.

定理 2.2

设 $f:\Omega\to\mathbb{R}$ 为连续可微的函数，其中 $\Omega\subset\mathbb{R}^n$ 为开凸集，则 f 为凸函数当且仅当如下条件之一成立：

1. 当任意 $x,y\in\Omega$ 均有
$$\langle x-y,\nabla f(x)-\nabla f(y)\rangle\ge 0; \tag{2.3}$$
2. 当任意 $x,y\in\Omega$ 均有
$$f(y)\ge f(x)+\langle\nabla f(x),y-x\rangle; \tag{2.4}$$
3. 若 f 二阶连续可微时，
$$\nabla^2 f(x)\succeq 0,\forall x\in\Omega \tag{2.5}$$

2.2 次微分理论

2.2.1 次微分与次梯度

下面，我们给出凸函数次梯度和次微分的定义。

定义 2.3

设 $f:\mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ 为凸函数且 $\mathrm{dom}(f)\neq\emptyset$ 并令 $x\in\mathrm{dom}(f)$。若向量 $g\in\mathbb{R}^n$ 满足

$$f(y)\geq f(x)+\langle g,y-x\rangle\,,\forall y\in\mathbb{R}^n$$

则称 g 为 f 在 x 处的梯度。函数 f 在 x 处的全体次梯度构成的集合称为 f 在 x 处的次微分，记为 $\partial f(x)$。♣

下例指出：并非任意凸函数在其有效域 $\mathrm{dom}(f)$ 中总可微。

例 2.2 定义 $f:\mathbb{R}\to(-\infty,+\infty]$ 为

$$f(x)=\begin{cases}-\sqrt{x}, & x\geq 0,\\ \infty, & \text{否则}.\end{cases}$$

往证 f 在 $x=0$ 处的次梯度不存在。反正存在 $g\in\partial f(0)$。则

$$f(y)\geq f(0)+g(y-0),\forall y>0$$

也即

$$-\sqrt{y}\geq gy,\ \forall y>0$$

从而

$$g\sqrt{y}\leq -1,\ \forall y>0$$

令 $y\to 0_+$ 可得 $0\geq -1$ 矛盾！从而 $\partial f(0)=\emptyset$

尽管如此，当限定到 f 有效域的内点时，次梯度总存在，也即有关系式：

$$\mathrm{int}\,\mathrm{dom}(f)\subseteq\mathrm{dom}(\partial f) \tag{2.6}$$

定理 2.3

设 $f:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ 为凸函数其 $\mathrm{dom}(f)\neq\emptyset$，则对任意 $x\in\mathrm{int}\,\mathrm{dom}(f)$ 及 $d\in\mathbb{R}^n$ 均为有

$$f'(x;d)=\max\{\langle g,d\rangle : g\in\partial f(x)\}$$

注 条件 $x\in\mathrm{int}\,\mathrm{dom}(f)$ 可弱化为 $x\in\mathrm{dom}(f)$ ♡

证明 由次梯度不等式可知，对 $\forall\in\partial f(x)$ 均有

$$\begin{aligned} f'(x;d) &= \lim_{\tau\to 0_+}\frac{f(x+\tau d)-f(x)}{\tau}\\ &\geq \lim_{\tau\to 0_+}\langle g,d\rangle\\ &=\langle g,d\rangle\end{aligned}$$

故

$$f'(x;d)\geq\max\{\langle g,d\rangle : g\in\partial f(x)\}$$

2.2.2 最速下降方向

再给出次微分集中下降的次梯度方向前，我们先回顾作为最速下降方向的负梯度方向。假设 $\nabla f(x)\neq 0$。固定方向 $d\in\mathbb{R}^n$ 且 $\|d\|=1$。考虑沿此方向函数值的变化。由可微的定义可知

$$f(x+\tau d)=f(x)+\tau\nabla f(x)^{\mathrm{T}}d+o(\tau),\ \tau>0$$

当 $\nabla f(x)^{\mathrm{T}}d<0$ 时，由于 $\frac{o(\tau)}{\tau}\to 0,\tau\to 0_+$ 可知，存在 $\varepsilon>0$，使得当 $0<\tau<\varepsilon$ 时，

$$\frac{o(\tau)}{\tau}\leq -\frac{\nabla f(x)^{\mathrm{T}}d}{2}$$

此时

$$f(x+\tau d)-f(x)\leq\frac{\tau}{2}\cdot\nabla f(x)^{\mathrm{T}}d<0$$

上式表明，从 x 点眼 d 方向出发，当步长 $\tau < \varepsilon$ 时，

$$f(x+\tau d) < f(x)$$

因此，称满足条件 $\nabla f(x)^{\mathrm{T}}d$ 的方向为下降方向。而称下降最快或 $\nabla f(x)^{\mathrm{T}}d$ 最小的方向为最速下降方向，也即

$$\hat{d} = \arg\min\{\nabla f(x)^{\mathrm{T}}d\}$$
$$\text{s.t.}\|d\| = 1$$

则 $d = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$。事实上，有 Cauchy-Schwarz 不等式

$$|\nabla f(x)^{\mathrm{T}}d| \leq \|\nabla f(x)\| \cdot \|d\| \leq \|\nabla f(x)\|$$

可知

$$-\|\nabla f(x)\| \leq \nabla f(x)^{\mathrm{T}}d \leq \|\nabla f(x)\|$$

从而 $\hat{d} = \frac{\nabla f(x)}{\|\nabla f(x)\|}$ 可取得下界 $-\|\nabla f(x)\|$。对一般的凸函数（不一定可微），我们用次线性的方向导数 $f'(x;d)$ 代替线性项 $\nabla f(x)^{\mathrm{T}}d$ 以类似地定义最速下降方向。

定理 2.4

设 $f: \mathrm{R}^n \to \mathbb{R}\cup\{+\infty\}$ 在 x 处次可微且 $0 \notin \partial f(x), x \in \operatorname{int}\operatorname{dom}(f)$。定义

$$\Lambda = \arg\min\{f'(x;d) : \|d\| \leq 1\}$$

为最速下降方向集。则

$$\hat{d} = -\frac{g}{\|g\|} \in \Lambda$$

其中 $g = \arg\min\{\|z\| : z \in \partial f(x)\}$

第 2 章 练习

1. 设 f 为凸函数且存在 $\beta_0 \in \mathbb{R}$ 使得 $M_{\beta_0} := \{x : f(x) < \beta_0\} \neq \emptyset$ 且有界。证明：$\forall \beta \in \mathbb{R}$，$M_\beta$ 均为有界集。

   证明 对于 $\beta \leq \beta_0$，显然 M_β 为有界集。

   不妨设 $\beta > \beta_0$，令 $f(x^*) = \alpha < f(x)\ \forall x \in \mathbb{R}^n$。

   令 $r = \max\limits_{x\in M_{\beta_0}}$，则对 $\forall x\|x-x^*\| = r, f(x) \geq \beta_0$ 所以对 $\forall x \in M_\beta$

$$\begin{aligned}\beta_0 &\leq f(x^* + \frac{r}{\|x-x^*\|}(x-x^*)) = f(\frac{r}{\|x-x^*\|}x + (1-\frac{r}{\|x-x^*\|})x^*)\\ &\geq \frac{r}{\|x-x^*\|}f(x) + (1-\frac{r}{\|x-x^*\|})f(x^*)\\ &< \frac{r}{\|x-x^*\|}\beta + (1-\frac{r}{\|x-x^*\|})\beta_0\end{aligned}$$

   因此，$\|x-x^*\| \leq r\cdot\dfrac{\beta-\alpha}{\beta_0-\alpha}$ M_β 有界。

   证毕！

2. 对 $x \in \mathbb{R}^n$，用 $x_{[i]}$ 表示 x 中第 i 大的分量：

$$x_{[1]} \geq x_{[2]} \geq \cdots \geq x_{[n]}$$

   证明：对任意的 $1 \leq k \leq n$，函数

$$f_k(x) = \sum_{j=1}^{k} x_{[j]}$$

   为凸函数

   证明 对 $x^1 = (x_1^1, x_2^1, \cdots, x_n^1)', x^2 = (x_1^2, x_2^2, \cdots, x_n^2)'$，$\forall \alpha \in (0,1)$，令 $x^\alpha = \alpha x^1 + (1-\alpha)x^2$，设 $x_{[i]}^\alpha = x_{rj}^\alpha$，

则

$$\begin{aligned}
f_k(x^\alpha) &= \sum_{j=1}^{k} x^\alpha_{[j]} = \sum_{j=1}^{k} x^\alpha_{rj} = \sum_{j=1}^{k} (\alpha x^1_{rj} + (1-\alpha)x^2_{rj}) \\
&= \alpha \sum_{j=1}^{k} x^1_{rj} + (1-\alpha) \sum_{j=1}^{k} x^2_{rj} \\
&\leq \alpha f_k(x^1) + (1-\alpha) f_k(x^2)
\end{aligned}$$

证毕！

第 3 章 最优化条件与对偶

3.1 最优化条件

3.1.1 无约束优化情形

首先给出极小点的定义。

定义 3.1 (极小点)

设 $f:\mathbb{R}^n\to\bar{\mathbb{R}}, X\subset\mathbb{R}^n$。若 $\hat{x}\in X$ 满足条件：存在 $\varepsilon>0$，使得 $f(y)\ge f(\hat{x}),\forall y\in X$ 且 $\|y-\hat{x}\|\le\varepsilon$，则称 $\hat{x}$ 为 f 在 X 上的局部极小点。若 $f(y)\ge f(\hat{x})$ 对任意 $y\in X$ 恒成立。则称 $\hat{x}$ 为 f 在 X 上的全局极小点。

♣

定理 3.1

设 $f:\mathbb{R}^n\to\bar{\mathbb{R}}$ 在 $\hat{x}$ 处可微，若 $\hat{x}$ 是 f 在 $\mathbb{R}^n$ 上的局部极小点，则有 $\nabla f(\hat{x})=0$；若假设 f 二阶连续可微，则还有 $\nabla^2 f(\hat{x})\succeq 0$

♡

证明 反设 $\nabla f(\hat{x})\neq 0$，由 f 在 $\hat{x}$ 处可微知

$$f(y)=f(\hat{x})+\langle\nabla f(\hat{x}),y-\hat{x}\rangle+o(\|y-\hat{x}\|)$$

考虑 $y(\tau)=\hat{x}-\tau\cdot\nabla f(\hat{x}),\tau>0$，带入上式可得

$$f(y(\tau))=f(\hat{x})-\tau\cdot\|\nabla f(\hat{x})\|^2+o(\tau). \tag{3.1}$$

由 $\lim_{\tau\to 0_+}\frac{o(\tau)}{\tau}=0$ 知存在 $\tau_0>0$ 使得当 $0<\tau<\tau_0$，

$$\frac{o(\tau)}{\tau}<\frac{1}{2}\|\nabla f(\hat{x})\|^2$$

于是由 (3.1) 知 $f(y(\tau))<f(\hat{x})-\frac{\tau}{2}\|\nabla f(\hat{x})\|^2<f(\hat{x})$. 矛盾！

若 f 在 $\hat{x}$ 处二阶可微，则

$$\begin{aligned}f(y)&=f(\hat{x})+\langle\nabla f(\hat{x}),y-\hat{x}\rangle+\frac{1}{2}(y-\hat{x})^T\nabla^2 f(\hat{x})(y-\hat{x})+o(\|y-\hat{x}\|^2)\\&=f(\hat{x})+\frac{1}{2}(y-\hat{x})^T\nabla^2 f(\hat{x})(y-\hat{x})+o(\|y-\hat{x}\|^2).\end{aligned} \tag{3.2}$$

反设 $\nabla^2 f(\hat{x})\prec 0$，固定非零向量 $d\in\mathbb{R}^n$ 并考虑

$$y(\tau)=\hat{x}+\tau d.$$

将其代入 (3.2) 式可得

$$f(y(\tau))=f(\hat{x})+\frac{\tau^2}{2}d^T\nabla^2 f(\hat{x})d+o(\tau^2).$$

类似的，存在 $\tau>0$，使得

$$o(\tau^2)<-\frac{\tau}{4}d^T\nabla^2 f(\hat{x})d.$$

从而

$$f(y(\tau))<f(\hat{x})+\frac{\tau^2}{4}d^T\nabla^2 f(\hat{x})d<f(\hat{x}).$$

矛盾！

定理 3.2

设 $f:\mathbb{R}^n \to \bar{\mathbb{R}}$ 为正规凸函数，以下结论成立

(i) 若 $\hat{x}$ 为 f 的局部极小，则 $\hat{x}$ 也为 f 的全局极小点，且

$$0 \in \partial f(\hat{x}) \tag{3.3}$$

(ii) 若 (3.3) 成立，则 $\hat{x}$ 为 f 的全局极小点

♡

证明

(i) 设 $\hat{x}$ 为 f 的局部极小点，对任意固定的 $y \in \mathbb{R}^n$，必存在 $1 > \alpha_0 > 0$，使得当 $0 < \alpha < \alpha_0$ 时，

$$\begin{aligned} f(\hat{x}) &\le f(\hat{x} + \alpha(y - \hat{x})) \\ &= f(\alpha y + (1-\alpha)\hat{x}) \\ &\le \alpha f(y) + (1-\alpha) f(\hat{x}) \end{aligned}$$

从而 $f(\hat{x}) \le f(y)$，可知 $\hat{x}$ 为 f 的全局极小点，进而

$$f(y) \ge f(\hat{x}) + \langle 0, y - \hat{x} \rangle, \forall y \in \mathbb{R}^n \tag{3.4}$$

因此，$0 \in \partial f(\hat{x})$

(ii) 当 (3.4) 式成立，(3.3) 式必成立，从而 $\hat{x}$ 为 f 的全局极小点。证毕！

第 4 章 梯度下降法

4.1 梯度法的描述

4.1.1 迭代格式

本章介绍的梯度法将用于近似求解下述的无约束优化问题

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\, f(x), \tag{4.1}$$

其中目标函数 $f:\mathbb{R}^n\to\mathbb{R}$ 为连续可微的函数并假设我们可以计算其在任意一点 x 处的梯度 $\nabla f(x)$。基于最优性条件，最优化问题 (4.1) 满足必要性条件 $\nabla f(x)=0$。从而将最优化问题转化为非线性方程组的求解，然后再所有可行解中找出目标函数值最小的解，即可得到最优化问题的最优解。但实际上，$\nabla f(x)=0$ 的求解是非常困难的，因为它本质上需要计算梯度算子的逆算子，相对解一个逆问题。特别地，当 f 为二次函数时，梯度算子的逆算子对应矩阵的逆矩阵，对应大规模问题而言，稠密矩阵的逆矩阵计算也是很困难的. Cauchy 当初提出梯度法的动机正是为了克服逆问题的求解困难而采取的一种正向计算（计算梯度）的方法来逐渐逼近问题的解。具体而言，梯度法的基本格式为：

$$x^{k+1}=x^k-\tau_k\cdot\nabla f(x^k),k\geq 0, \tag{4.2}$$

其中 τ_k 为步长参数，$x^0\in\mathbb{R}^n$ 为给定的初始点。由于 $-\nabla f(x^k)$ 为当前迭代点 x^k 的最速下降方向，当步长参数 τ_k 落在某个足够小的区间 $(0,\tau)$ 时，在更新点 x^{k+1} 出的函数值 $f(x^{k+1})$ 必然减少，也即 $f(x^{k+1})<f(x^k)$。因此，梯度法也通常称为梯度下降法。

为了进一步看出迭代格式 (4.2) 的内在机制，我们将其等价地写成二次优化问题最优解的形式

$$x^{k+1}:=\arg\min_x\{f_k(x):=\frac{1}{\tau_k}\|x-(x^k-\tau_k\cdot\nabla f(x^k))\|^2+f(x^k)-\frac{\tau_k}{2}\|\nabla f(x^k)\|^2\},$$

其中添加的常数项 $f(x^k)-\frac{\tau_k}{2}\|\nabla f(x^k)\|^2$ 是使目标函数 $f_k(x)$ 凑成如下形式

$$f_k(x)=f(x^k)+\langle\nabla f(x^k),x-x^k\rangle+\frac{1}{2\tau_k}\|x-x^k\|^2.$$

基于该形式，我们可以看出梯度法本质上是求解一系列二次优化问题，它们的目标函数 f_k 是原目标函数 f 的二阶近似。不难预见，近似程度的好坏将决定梯度法的收敛性与收敛率。特别地，当 f 为凸的 L-光滑函数时，f_k 与 f 的近似程度可以被一个二次函数控制。事实上，由凸函数和 L-光滑性的定义可知

$$f(x^k)+\langle\nabla f(x^k),x-x^k\rangle\leq f(x)\leq f(x^k)+\langle\nabla f(x^k),x-x^k\rangle+\frac{L}{2}\|x-x^k\|^2.$$

上式同减 f_k 后可得

$$-\frac{1}{2\tau_k}\|x-x^k\|^2\leq f(x)-f_k(x)\leq(\frac{L}{2}-\frac{1}{2\tau_k})\|x-x^k\|^2.$$

因此，$f(x)-f_k(x)$ 被二次函数 $\|x-x^k\|^2$ 控制。下述的下降引理表明：只需目标函数 f 满足 L-光滑性，就可通过控制步长参数的选取以保证梯度法生成的函数值序列下降.

定理 4.1 (基本下降引理)

假设目标函数 f 为满足 L-光滑性条件：

$$f(y)\geq f(x)+\langle\nabla f(x),y-x\rangle+\frac{L}{2}\|y-x\|^2 \tag{4.3}$$

则梯度法格式 (4.3) 生成的序列 $\{x^k\}$ 满足如下下降性质：

$$f(x^k)-f(x^{k+1})\geq\tau_k(1-\frac{L}{2}\tau_k)\|\nabla f(x^k)\|^2 \tag{4.4}$$

♡

证明 首先，在光滑性条件 (4.3) 中取 $y = x^{k+1}, x = x^k$ 可得

$$f(x^{k+1}) \le f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2.$$

然后，将迭代式 $x^{k+1} = x^k - \tau_k \cdot \nabla f(x^k)$ 待入上式的右边得到

$$\begin{aligned} f(x^k) - f(x^{k+1}) &\ge \langle \nabla f(x^k), \tau_k \nabla f(x^k) \rangle - \tfrac{L}{2}\| - \tau_k \nabla f(x^k)\|^2 \\ &= \tau_k (1 - \frac{L}{2}\tau_k)\|\nabla f(x^k)\|^2 \end{aligned}$$

证毕！

4.1.2 步长选取

梯度法迭代法 (4.2) 中补偿参数 τ_k 的设计直接影响到梯度法的收敛效率。若步长取得过于保守，每步迭代使函数值的下降量微乎其微，算法将收敛的非常缓慢；而若取得太过激进，单独迭代可能使函数值上升，算法将不稳定甚至不收敛。因此合理的步长应该介于保守和激进之间。此处，我们将介绍三种常见的选取策略。

第一种是常数步长策略。其理论依据是下降关系 (4.4) 中右边的步长函数

$$h(\tau) := \tau(1 - \frac{L}{2}\tau)$$

一方面，为了保证单步迭代均能使函数值下降，我们要求 $h(\tau) > 0$，由于步长参数的非负性，可知 $\tau \in (0, \tau)$，换言之在区间 $(0, \frac{2}{L})$ 中任选常数步长均可保证函数值下降（梯度未消失之前）；另一方面，为了使单步迭代的函数值下降最多，我们可取

$$\tau_k := \arg \max_{\tau \in (0, \frac{2}{L})} h(\tau)$$

此时常数步长 $\tau_k \equiv \frac{1}{L}$，并取该步长的梯度法为最优常数步长梯度法。相应的， (4.4) 式变为

$$f(x^k) - f(x^{k+1}) = f(x^k) - f(x^k - \frac{1}{L}\nabla f(x^k)) \ge \frac{1}{2L}\|\nabla f(x^k)\|^2 \tag{4.5}$$

改策略的局限在于目标函数的光滑常数 L 难以估计。

第二种是精确线搜索策略。该策略尝试从当前点 x^k 出发，沿最速下降方向 $-\nabla f(x^k)$ 前进，以求尽可能降低函数值。具体而言，

$$\tau_k \in \operatorname{Arg}\min_{\tau > 0}\{g(\tau) := f(x^k - \tau \nabla f(x^k))\}$$

此时，我们需要在正区间上极小化单变量的连续可微函数 $g(\tau)$ 或者求方程 $g'(\tau) = 0$ 的正根。具体到 f 为二次函数是，$g(\tau)$ 是关于 τ 的二次函数，而单变量的二次函数极小化问题可以解析地求解。对应一般情形，通常需要基于数值方法估计方程 $g(\tau) = 0$ 的根。

由于该策略的目标是搜索下降方向 $-\nabla f(x^k)$ 上的最下函数值，因此精确线搜索的步长 τ_k 必然满足

$$f(x^{k+1}) = f(x^k - \tau \nabla f(x^k)) \le f(x^k - \frac{1}{L}\nabla f(x^k))$$

因而，结合 (4.5) 可知精确线性搜索的梯度法满足与最优常数步长梯度法相同的下降性质

$$f(x) - f(x^{k+1}) \ge f(x) - f(x^k - \frac{1}{L}\nabla f(x^k)) \ge \frac{1}{2L}\|\nabla f(x^k)\|^2$$

第三种是反向追踪线搜索策略。该方法既不需要估计光滑常数 L，也不需要求解单变量的最优化问题，因为在实际应用中更为常见。其设计的理论依据是 f 的 L-光滑性条件，因为该条件可以保证当步长 $\tau_k \le \dfrac{1}{L}$ 时，必有

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\tau_k}\|y - x\|^2$$

在上述中分别取 $x = x^k, y = x^{k+1} = x^k - \tau_k \nabla f(x^k)$ 可得

$$f(x^{k+1}) \le f(x^k) - \frac{\tau}{2}\|\nabla f(x^k)\|^2 \tag{4.6}$$

下面尝试去掉 τ_k 对于光滑常数 L 的依赖。为此给定一个 τ_k 的初始估计值 $\sigma > 0$ 以及缩减参数 $\beta \in (0, 1)$。

1. 将 $\tau_k = \sigma$ 带入梯格式中，判断 (4.6) 是否成立

2. 若不成立，缩减步长 $\tau_k = \beta\sigma$ 重复之前过程
3. 若成立，停止

如此重复，由于 $\beta \in (0,1)$，必然存在自然数 m 使得 $\beta^m \leq \frac{1}{L}$ 使得 (4.6) 成立。此时，要么有 $\tau_k = \sigma$ 要么 $\frac{\tau_k}{\beta} > \frac{1}{L}$。因此

$$\tau_k \geq 2c := \min\{\sigma, \frac{\beta}{L}\}$$

反向追踪线搜索梯度法成立如下下降性质：

$$f(x^k) - f(x^{k+1}) \geq \frac{c}{2}\|\nabla f(x^k)\|^2 \tag{4.7}$$

4.2 收敛理论

判断梯度法的收敛性的三种度量标准

1. 目标间隙 $f(x^k) - \bar{f}$
2. 梯度范数 $\|f(x^k)\|$
3. 距离间隙 $d(x^k, X*)$

4.2.1 基本收敛性质

定理 4.2

设 $f : \mathbb{R}^n \to \mathbb{R}$ 满足 L-光滑条件且有下下界，也即存在常数 $\ell \in \mathbb{R}$，使得

$$\ell < f(y) < f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$$

令 $\{x^k\}$ 为梯度法 (4.2) 应用于 f 生成的序列，其中步长参数 τ_k 可选为常数步长 $\tau \in (0, \frac{1}{L})$ 或按精确线搜索或 backtracking 线搜索。则有

$$\lim_{k\to\infty} = \|\nabla f(x^k)\| = 0$$

且

$$\min_{k=0,1,\cdots,T} \|\nabla f(x^k)\|^2 \leq \frac{c_1}{T+1}$$

证明 由 (4.7) 可知序列 $\{f(x^K)\}$ 单调下降且有下界 ℓ。从而必为收敛序列，其极限记为 $\bar{f}$。因而

$$f(x^k) - f(x^{k+1}) \to 0, k \to 0$$

再结合可知

$$\lim_{k\to\infty} = \|\nabla f(x^k)\| = 0$$

成立。现在从 $k = 0, 1, \cdots, T$ 求和

$$\begin{cases} f(x^0) - f(x^1) \geq c\|\nabla f(x^0)\|^2 \\ f(x^1) - f(x^2) \geq c\|\nabla f(x^1)\|^2 \\ \vdots \\ f(x^T) - f(x^{T+1}) \geq c\|\nabla f(x^T)\|^2 \end{cases}$$

可得

$$f(x^0) - f(x^{T+1}) \geq c\sum_{k=0}^{T} \|\nabla f(x^k)\|^2$$

又因为 $f(x^{T+1}) \geq \bar{f}$，于是

$$\begin{aligned} f(x^0) - \bar{f} \quad &\geq c \sum_{k=0}^{T} \|\nabla f(x^k)\|^2 \\ &\geq c \cdot (T+1) \cdot \min_{k=0,1,\cdots,T} \|\nabla f(x^k)\|^2 \end{aligned}$$

令 $c_1 = \frac{f(x^0)-\bar{f}}{c}$，即可推出 $\min\limits_{k=0,1,\cdots,T} \|\nabla f(x^k)\|^2 \leq \dfrac{c_1}{T+1}$

参考文献

[1] 杨明, 刘先忠. 矩阵论[M]. 2 版. 武汉: 华中科技大学, 2006.

附录 A 基本数学工具

本附录包括了计量经济学中用到的一些基本数学，我们扼要论述了求和算子的各种性质，研究了线性和某些非线性方程的性质，并复习了比例和百分数。我们还介绍了一些在应用计量经济学中常见的特殊函数，包括二次函数和自然对数，前 4 节只要求基本的代数技巧，第 5 节则对微分学进行了简要回顾；虽然要理解本书的大部分内容，微积分并非必需，但在一些章末附录和第 3 篇某些高深专题中，我们还是用到了微积分。

A.1 求和算子与描述统计量

求和算子是用以表达多个数求和运算的一个缩略符号，它在统计学和计量经济学分析中扮演着重要作用。如果 $\{x_i : i = 1, 2, \ldots, n\}$ 表示 n 个数的一个序列，那么我们就把这 n 个数的和写为：

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \cdots + x_n \tag{A.1}$$