数值计算方法

扩充教程

童伟华　编

中国科学技术大学

目录

第九章　函数逼近

在理论和工程领域, 经常会遇到**函数逼近**问题, 即如何寻找简单的函数 $\varphi(x)$ 去近似地代替一个复杂的函数 $f(x)$, 其中近似代替又称为**逼近**, 函数 $f(x)$ 和 $\varphi(x)$ 分别称为**被逼近**和**逼近**函数. 利用简单函数去逼近复杂函数的一个常用目的: 使得一些常用的操作, 譬如函数求值、微分甚至积分, 可以变得更容易执行. 另一个常用目的: 利用函数的部分信息, 譬如函数值表, 重建或恢复一个函数. 常用于构造逼近函数的类包括: 多项式, 三角多项式, 分片多项式等. 下面是函数逼近的典型例子:

(1) 在区间 $[-1,1]$ 上, 确定具有最低次数的多项式 $p(x)$ 使得 $|p(x)-\arccos(x)| \leqslant 10^{-7}$ 成立. 更一般地, 给定函数 $f(x)$ 和正数 ε, 确定多项式 $p(x)$, 使在区间 $[a,b]$ 上有 $|p(x)-f(x)| \leqslant \varepsilon$;

(2) 通过观察或测量函数 $f(x)$ 得到一组离散数据: $\{(x_i,y_i) \mid i=1,2,\ldots,n\}$, 在函数空间 $\Phi = \text{span}\{\varphi_i(x) \mid j=1,2,\ldots,m\}$ 中选择 $\varphi(x)=\sum_{j=1}^{m} c_j\varphi_j(x)$ 使得逼近误差最小, 即

$$\min_{\varphi\in\Phi}\sum_{i=1}^{n}\Big|y_i-\varphi(x_i)\Big|^2 = \min_{c_1,c_2,\cdots,c_m\in\mathbb{R}}\sum_{i=1}^{n}\Big|y_i-\sum_{j=1}^{m}c_j\varphi_j(x_i)\Big|^2.$$

9.1　逼近问题的描述

在逼近问题中几乎都涉及到从一个集合中选择一个元素, 使它在某种意义上**接近**该集合外的一个预先给定的元素. 因此, 若要确切的描述逼近问题, 需要明确两个元素之间的**距离**是如何度量的. 为了在统一的框架下描述逼近问题, 下面引入赋范线性空间.

定义 9.1 设集合 V 是实数域 $\mathbb{R}$ 上的线性空间, 如果 V 中任意一个元素 f 都按某一法则对应一个实数, 记作 $\|f\|$, 并且它满足下列条件:

(1) **正定性**: $\|f\| \geqslant 0, \quad \forall f \in V; \|f\| = 0$ 当且仅当 $f = 0$ 成立;

(2) **齐次性**: $\|cf\| = |c|\|f\|, \quad \forall c \in \mathbb{R}, \forall f \in V$;

(3) **三角不等式**: $\|f + g\| \leqslant \|f\| + \|g\|, \quad \forall f, g \in V$.

上述对应关系可视为 $V \to \mathbb{R}$ 的映射, 称为线性空间 V 的**范数**, 并简记为 $\|\cdot\|$. 定义了范数的线性空间称为**赋范线性空间**.

下面对常用的有限维线性空间 $\mathbb{R}^n$ 和无穷维线性空间 $C[a, b]$ 分别引入范数.

例 9.1 记 $\mathbb{R}^n$ 为 n 维线性空间, 在 $\mathbb{R}^n$ 中定义

$$\|\mathbf{x}\|_2 = \left(x_1^2 + x_2^2 + \cdots + x_n^2\right)^{1/2}, \quad \forall \mathbf{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}} \in \mathbb{R}^n.$$

易验证 $\|\cdot\|_2$ 满足条件 (1) $\sim$ (3). 因此, $\mathbb{R}^n$ 按 $\|\cdot\|_2$ 构成一赋范线性空间. 事实上, 若 n 维线性空间 $\mathbb{R}^n$ 按常用的内积 $(\cdot, \cdot)$ 构成欧氏空间, 则范数 $\|\mathbf{x}\|_2$ 可视为向量 $\mathbf{x}$ 自己与自己内积的平方根, 即 $\|\mathbf{x}\|_2 = \sqrt{(\mathbf{x}, \mathbf{x})}$, 称 $\|\mathbf{x}\|_2$ 为向量的**2-范数**或**欧几里得范数**. 另外, 不难验证 $\mathbb{R}^n$ 还可按如下范数

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|, \quad \forall \mathbf{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}} \in \mathbb{R}^n,$$

$$\|\mathbf{x}\|_\infty = \max\{|x_1|, |x_2|, \cdots, |x_n|\}, \quad \forall \mathbf{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}} \in \mathbb{R}^n,$$

分别构成不同的赋范线性空间. 更一般地, 在 $\mathbb{R}^n$ 中定义

$$\|\mathbf{x}\|_p = \left(|x_1|^p + |x_2|^p + \cdots + |x_n|^p\right)^{1/p}, \quad \forall \mathbf{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}} \in \mathbb{R}^n,$$

构成向量 $\mathbf{x}$ 的**p-范数**, 前面的范数分别对应 $p = 1, 2, \infty$ 的情形.

例 9.2 记 $C[a, b]$ 为区间 $[a, b]$ 上连续函数的全体, 按通常的函数加法与数乘运算构成线性空间. 在 $C[a, b]$ 中定义

$$\|f\|_\infty = \max_{a \leqslant x \leqslant b} |f(x)|, \quad \forall f \in C[a, b].$$

易验证 $\|\cdot\|_\infty$ 满足条件 (1) $\sim$ (3). 因此, $C[a, b]$ 按 $\|\cdot\|_\infty$ 构成一赋范线性空间, 范数 $\|\cdot\|_\infty$ 称为**一致范数**或 **Chebyshev 范数**.

例 9.3 记 $C^r[a,b]$ 为区间 $[a,b]$ 上 r 次连续可微函数的全体. 定义 $C^r[a,b]$ 的范数

$$\|f\|_\infty = \max_{x\in[a,b]} \Big\{|f(x)|, |f'(x)|, \cdots, |f^{(r)}(x)|\Big\}, \quad \forall f \in C^r[a,b].$$

显然, $C[a,b]$ 是 $C^r[a,b]$ 的一个特殊情形.

例 9.4 记 $L^p[a,b]$ 为区间 $[a,b]$ 上所有满足

$$\int_a^b |f(x)|^p \,\mathrm{d}x < +\infty, \quad p \geqslant 1,$$

的 Lebesgue 可积函数 f 构成的函数类 (Lebesgue 积分是 Riemann 积分的推广). 因区间 $[a,b]$ 上所有的连续函数都是 Riemann 可积的, 故 $C[a,b] \subset L[a,b]$. 在 $L^p[a,b]$ 中定义

$$\|f\|_p = \left(\int_a^b |f(x)|^p \,\mathrm{d}x\right)^{1/p}, \quad \forall f \in L^p[a,b], \tag{9.1}$$

可以证明 $\|\cdot\|_p$ 是 $L^p[a,b]$ 的一个范数. 注意, 在 $L^p[a,b]$ 中约定: 将几乎处处相等的两个可测函数 f, g 视为同一函数.

在赋范线性空间中, 可以按照下面的方式引入向量之间的距离.

定义 9.2 在赋范线性空间 V 中, 定义函数

$$d(f,g) = \|f-g\|, \quad \forall f,g \in V,$$

称为 f 与 g 之间的**距离**. 不难验证 $d(f,g)$ 满足距离定义所要求的条件:

(1) **正定性**: $d(f,g) \geqslant 0, \quad \forall f \in V$; $d(f,g) = 0$ 当且仅当 $f = g$ 成立;

(2) **对称性**: $d(f,g) = d(g,f), \quad \forall f,g \in V$;

(3) **三角不等式**: $d(f,g) \leqslant d(f,h) + d(h,g), \quad \forall f,g,h \in V$.

有了距离的定义, 便可讨论函数的连续性.

引理 9.1 在赋范线性空间 V 中, 加法, 数乘和范数都是距离 $d(f,g)$ 下的连续函数.

证明 设 V 中有收敛的序列 $\lim\limits_{n\to\infty} f_n = f^*$ 及 $\lim\limits_{n\to\infty} g_n = g^*$, 则有

$$\begin{aligned} d(f^*+g^*, f_n+g_n) &= \|f^*+g^*-(f_n+g_n)\| \\ &\leqslant \|f^*-f_n\| + \|g^*-g_n\| \\ &= d(f^*,f_n) + d(g^*,g_n), \end{aligned}$$

故而 $\lim\limits_{n\to\infty}(f_n+g_n)=\lim\limits_{n\to\infty}f_n+\lim\limits_{n\to\infty}g_n$ 成立.

类似地, 可以证明 $\lim\limits_{n\to\infty}\lambda f_n=\lambda\lim\limits_{n\to\infty}f_n$ 和 $\lim\limits_{n\to\infty}\|f_n\|=\|\lim\limits_{n\to\infty}f_n\|$ 成立. □

下面设 X 是赋范线性空间, M 是 X 的非空子集, 我们希望从 M 中选取元素逼近 X 中的元素, M 称为 X 的一个**逼近集合**.

定义 9.3 对于 $x\in X$, 如果有元素 $m^*\in M$ 使得

$$\|x-m^*\|=\inf_{m\in M}\|x-m\|\triangleq d(x,M),$$

则称 m^* 为子集 M 逼近 x 的**最佳逼近元**, 记为 $m^*\in\mathcal{B}_M(x)$, 其中

$$\mathcal{B}_M(x)\triangleq\Big\{m\in M:\|x-m\|=d(x,M)\Big\}$$

表示由 M 逼近 x 的最佳逼近元构成的集合, 用 $\#\mathcal{B}_M(x)$ 表示最佳逼近元的个数.

有了最佳逼近元的定义之后, 自然地会产生以下问题:

(1) 存在性, 即是否有 $\#\mathcal{B}_M(x)\geqslant 1$;

(2) 唯一性, 即是否有 $\#\mathcal{B}_M(x)\leqslant 1$;

(3) 最佳逼近元应具有什么特征;

(4) 最佳逼近元的构造及其应用.

定义 9.4 X 的一个子集 M 称为**列紧的**, 如果 M 中的每个点列都有一个收敛于 M 中一点的子序列.

定理 9.2 设 M 是 X 的列紧子集, 则对于任意的 $x\in X$, 存在最佳逼近元 $m^*\in M$.

证明 若 $x\in M$, 显然 $x=m^*\in\mathcal{B}_M(x)$. 下设 $x\in X\backslash M$, 由于

$$d(x,M)=\inf_{m\in M}\|x-m\|,$$

故存在 $m_n\in M$ 使得 $\lim\limits_{n\to\infty}\|x-m_n\|=d(x,M)$. 利用 M 的列紧性, 存在 $\{m_n\}$ 的子序列 $\{m_{n_k}\}$ 使得

$$\lim_{k\to\infty}m_{n_k}=m^*\in M$$

成立. 又因范数的三角不等式, 可知

$$\|x-m^*\|\leqslant\|x-m_{n_k}\|+\|m_{n_k}-m^*\|.$$

上述不等式两边取极限, 有 $\|x-m^*\| \leqslant d(x,M)$. 另一方面, $m^* \in M$ 有 $\|x-m^*\| \geqslant d(x,M)$. 因此, 存在 $m^* \in M$ 使得 $\|x-m^*\| = d(x,M)$ 成立. □

推论 9.3 若 M 是 X 的线性子空间, 且 $\dim(M) < +\infty$, 则对任意的 $x \in X$, 存在最佳逼近元 $m^* \in M$.

证明 下面先证明 X 的任意有限维有界闭子集 F 是紧集. 因 F 是 X 的有限维线性子空间, 故存在一个线性无关集 $\{f_1, f_2, \ldots, f_n\}$, 使得每个 $f \in F$ 都可唯一的表示成

$$f = \lambda_1 f_1 + \lambda_2 f_2 + \cdots + \lambda_n f_n, \quad \lambda = (\lambda_1, \lambda_2, \cdots, \lambda_n)^{\mathrm{T}} \in \mathbb{R}^n.$$

因此, 定义映射 $T: \mathbb{R}^n \mapsto F$, $T(\lambda) = f$. 记集合 $N = \{\lambda \in \mathbb{R}^n : T(\lambda) \in F\}$, 则 F 可视为 N 在映射 T 下的像, 且不难看出 T 在 $\mathbb{R}^n$ 的 $\|\cdot\|_\infty$ 范数下是连续映射.

设 N 中任意收敛的序列 $\lim\limits_{n\to\infty} \lambda_n = \lambda^*$, 则

$$T(\lambda^*) = T(\lim_{n\to\infty} \lambda_n) = \lim_{n\to\infty} T(\lambda_n).$$

因 F 是闭集, 故 $T(\lambda^*) \in F$. 据 N 的定义, 知 $\lambda^* \in N$. 因此, N 是闭集.

记集合 $S = \{\lambda \in \mathbb{R}^n : \|\lambda\|_\infty = 1\}$, 显然 S 是一个紧集, 且 T 在 S 上连续. 因此, $\|T(\lambda)\|$ 的下确界 α 在 S 上能取到. 由于 $f_1, f_2, \cdots, f_n$ 是线性无关的, 从而 $\alpha > 0$. 对于任意的 $\lambda(\neq 0) \in N$, 则有

$$\|T(\lambda)\| = \|T(\lambda/\|\lambda\|_\infty)\| \cdot \|\lambda\|_\infty \geqslant \alpha\|\lambda\|_\infty.$$

因 $\|T(\lambda)\|$ 在 F 上有界, 故 $\|\lambda\|_\infty$ 在 N 上有界.

综上, N 是 $\mathbb{R}^n$ 的有界闭集, 故是一个紧集. 又因为 F 是 N 在连续映射下的像, 所以 F 也是一个紧集.

记集合 $K = \{m \in M : \|m-x\| \leqslant \|x\|\}$, 显然 $0 \in K$, 故集合 K 非空. 容易看出, K 是 X 的一个有限维的有界闭子集, 利用前面的结论可知 K 是一个紧集. 考虑函数 $f: K \mapsto \mathbb{R}$,

$$f(m) = \|m-x\|, \quad \forall m \in K,$$

由引理9.1知 $f(m)$ 是 K 上的连续函数. 因此, $f(m)$ 在紧集 K 上可以取到最小值, 即存在 $m^* \in M$ 使得 $\|x-m^*\| = d(x,M)$ 成立. □

上述定理及推论回答了一般赋泛线性空间最佳逼近元的存在性问题.

接下来, 我们讨论唯一性问题. 一般情况下, 最佳逼近元是不唯一的. 容易看出, $\mathcal{B}_M(x) = M \bigcap B(x, d(x, M))$, 其中 $B(x, d(x, M))$ 表示以 x 为球心, 半径为 $d(x, M)$ 的球. 因此最佳逼近元的唯一性与 M 的性质及 X 中单位球的性质相关.

定义 9.5 设 M 是赋范线性空间 X 的非空子集, 称 M 是**凸集**, 如果对任意的 $m_1, m_2 \in M, t \in (0, 1)$, 均有 $t * m_1 + (1 - t) * m_2 \in M$ 成立. 进一步, 若 $m_1 \neq m_2$, 均有 $t * m_1 + (1 - t) * m_2 \in M^\circ$ 成立 (M° 表示集合 M 的内部), 则称 M 是**严格凸集**.

容易验证赋范线性空间的闭球 $B(x, r)$ 是凸集. 在 $\mathbb{R}^n$ 和 $L^p[a, b]$ 中, 当 $1 < p < \infty$ 时, 按范数 $\|\cdot\|_p$ 定义的闭球 $B(x, r)$ 是严格凸集; 当 $p = 1$ 或 $p = \infty$ 时, $B(x, r)$ 是凸集.

定义 9.6 如果赋范线性空间 X 按某一范数 $\|\cdot\|$ 的闭球 $B(x, r)$ 是严格凸集, 则称该范数 $\|\cdot\|$ 是**严格凸的**.

几何上, 严格凸意味者单位球在其球面上不含任何线段.

定理 9.4 设 M 是 X 的列紧子集, 且 M 是严格凸集, 则对任意的 $x \in X$, 存在唯一的最佳逼近元 $m^* \in M$.

证明 存在性由定理9.3保证. 下证唯一性. 若 $d(x, M) = 0$, 则 $\mathcal{B}_M(x) = \{x\}$, 命题成立. 不妨设 $d(x, M) > 0$, 假若存在两个不同的元素 $m_1, m_2 \in \mathcal{B}_M(x)$, 则有

$$\Big\|\frac{1}{2}(m_1 + m_2) - x\Big\| \leqslant \frac{1}{2}\Big\|m_1 - x\Big\| + \frac{1}{2}\Big\|m_2 - x\Big\| = d(x, M).$$

因 M 是凸集, 故 $\dfrac{1}{2}(m_1 + m_2) \in M \Rightarrow \dfrac{1}{2}(m_1 + m_2) \in \mathcal{B}_M(x)$. 考虑集合

$$\Big\{\lambda \in [0, 1] : \frac{1}{2}(m_1 + m_2) + \lambda\big[x - \frac{1}{2}(m_1 + m_2)\big] \in M\Big\},$$

显然它存在上确界. 又由 M 是列紧的, 知 λ 能取到最大值 $\bar{\lambda}$. 此时,

$$\Big\|\frac{1}{2}(m_1 + m_2) + \bar{\lambda}\big[x - \frac{1}{2}(m_1 + m_2)\big] - x\Big\| = (1 - \bar{\lambda})d(x, M).$$

因 M 是严格凸集, 故 $\dfrac{1}{2}(m_1 + m_2)$ 是 M 的内点, 从而有 $\bar{\lambda} > 0$. 利用上式, 可知存在

$$\bar{m} = \frac{1}{2}(m_1 + m_2) + \bar{\lambda}\big[x - \frac{1}{2}(m_1 + m_2)\big] \in M,$$

使得 $\|\bar{m} - x\| < d(x, M)$, 这与 $d(x, M)$ 的最小性矛盾. □

推论 9.5 若 M 是 X 的线性子空间, 且 $\dim(M) < +\infty$, X 的范数 $\|\cdot\|$ 是严格凸的, 则对任意的 $x \in X$, 存在唯一的最佳逼近元 $m^* \in M$.

证明 存在性由推论9.3保证. 下证唯一性. 若 $d(x, M) = 0$, 则 $\mathcal{B}_M(x) = \{x\}$, 命题成立. 不妨设 $d(x, M) > 0$. 用反证法. 假设存在两个不同的元素 $m_1, m_2 \in \mathcal{B}_M(x)$, 即 $m_1, m_2 \in B(x, d(x, M))$. 因 X 按范数 $\|\cdot\|$ 是严格凸的, 故 $\frac{1}{2}(m_1+m_2)$ 是 $B(x, d(x, M))$ 的内点. 因此, 有

$$\left\|\frac{1}{2}(m_1 + m_2) - x\right\| < d(x, M).$$

因 M 是 X 的线性子空间, 故 $\frac{1}{2}(m_1 + m_2) \in M$, 这与 $d(x, M)$ 的最小性矛盾. □

本节仅讨论了一般赋范线性空间最佳逼近元的存在唯一问题. 关于其他的几个问题, 以后各节会逐一讨论. 接下来, 针对一些具有重要理论意义和应用背景的赋范线性空间研究相应的最佳逼近问题.

9.2 内积空间的最佳逼近

在线性代数中, 我们学习过 n 维的欧几里德空间, 即装配了内积的有限维线性空间, 可以描述向量的长度、正交等几何性质. 对于无穷维的线性空间, 可以按相同的方式引入内积的定义.

定义 9.7 设集合 V 是实数域 $\mathbb{R}$ 上的线性空间, 如果 V 中任意一对元素 f, g 都按某一法则对应一个实数, 记作 (f, g), 并且满足下列条件:

(1) **对称性**: $(f, g) = (g, f), \quad \forall f, g \in V$;

(2) **线性性**: $(\lambda f + \mu g, h) = \lambda(f, h) + \mu(g, h), \quad \forall \lambda, \mu \in \mathbb{R}, \forall f, g, h \in V$;

(3) **正定性**: $(f, f) \geqslant 0, \quad \forall f \in V; (f, f) = 0 \Leftrightarrow f = 0$,

则称二元实函数 $(\cdot, \cdot)$ 是线性空间 V 上的一个**内积**. 定义了内积的线性空间 V 称为**内积空间**.

例 9.5 在 $\mathbb{R}^n$ 空间中, 任取一组标准正交基 $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n$, 则

$$(\mathbf{x}, \mathbf{y}) = x_1y_1 + x_2y_2 + \cdots + x_ny_n, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

其中 $\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n$, $\mathbf{y} = y_1\mathbf{e}_1 + y_2\mathbf{e}_2 + \cdots + y_n\mathbf{e}_n$.

例 9.6 在 $L^2[a,b]$ 空间中, 定义

$$(f,g)=\int_a^b f(x)g(x)\,\mathrm{d}x,\quad \forall f,g\in L^2[a,b]. \tag{9.2}$$

易验证 $(\cdot,\cdot)$ 满足条件 (1) $\sim$ (3). 因此, $L^2[a,b]$ 按 $(\cdot,\cdot)$ 构成一内积空间.

命题 9.6 (Cauchy-Schwarz 不等式) 设 V 是内积空间, 则有

$$|(f,g)|\leqslant\sqrt{(f,f)\cdot(g,g)},\quad \forall f,g\in V.$$

证明 设 $f,g\in V$, 因 V 是线性空间, 故 $\lambda f+g\in V(\lambda\in\mathbb{R})$. 利用内积的正定性, 知

$$(\lambda f+g,\lambda f+g)\geqslant 0\Longleftrightarrow\lambda^2(f,f)+2\lambda(f,g)+(g,g)\geqslant 0.$$

因上式对任意的 $\lambda\in\mathbb{R}$ 成立, 故由二次函数的性质知:

$$4(f,g)^2-4(f,f)\cdot(g,g)\leqslant 0\Longrightarrow|(f,g)|\leqslant\sqrt{(f,f)\cdot(g,g)}.$$

由于 f,g 是任取的, 故命题成立. □

若在内积空间 V 中定义

$$\|f\|=\sqrt{(f,f)},\quad \forall f\in V,$$

则有

$$\begin{aligned}\|f+g\|^2&=(f+g,f+g)=(f,f)+2(f,g)+(g,g)\\&\leqslant(f,f)+2\sqrt{(f,f)\cdot(g,g)}+(g,g)=(\|f\|+\|g\|)^2,\end{aligned}$$

即 $\|f+g\|\leqslant\|f\|+\|g\|$. 易验证, $\|\cdot\|$ 构成 V 的一个范数, 称 $\|\cdot\|$ 是**内积诱导的范数**. 与一般的赋范线性空间不同, 内积空间具有很好的几何性质.

命题 9.7 (平行四边形等式) 设 V 是内积空间, 则有

$$\|f+g\|^2+\|f-g\|^2=2(\|f\|^2+\|g\|^2).$$

证明 设 $f,g\in V$, 因 V 是线性空间, 故 $f+g,f-g\in V$. 经简单计算知

$$\begin{aligned}\|f+g\|^2&=(f+g,f+g)=(f,f)+2(f,g)+(g,g),\\\|f-g\|^2&=(f-g,f-g)=(f,f)-2(f,g)+(g,g),\end{aligned}$$

上面两式相加即得. □

在内积空间中, 若 f 与 g 的内积为零, 即 $(f,g)=0$, 则称 f 与 g 是**正交**的. 此时, $\|f+g\|^2=\|f\|^2+\|g\|^2$, 类似于欧式空间中的勾股定理.

下面, 在内积空间中讨论最佳逼近问题.

定义 9.8 设 V 是内积空间, $M\subset V$ 为有限维子空间. 对于 $x\in V$, 如果有元素 $m^*\in M$ 使得

$$\|x-m^*\|=\inf_{m\in M}\|x-m\|\triangleq d(x,M),$$

则称 m^* 为子集 M 逼近 x 的**最佳逼近元**, 将所有 M 中 x 的最佳逼近元构成的集合记作 $\mathcal{B}_M(x)$, 这里 $\|\cdot\|$ 是 V 内积诱导的范数.

首先, 我们证明内积空间中最佳逼近元的存在唯一性.

定理 9.8 对于任意的 $x\in V$, 存在唯一的最佳逼近元 $m^*\in M$.

证明 据定义, V 是内积空间, 按内积诱导的范数 $\|\cdot\|$ 构成赋范线性空间, M 是 V 的有限维子空间. 因此, 利用推论9.3, 存在性得证.

下证唯一性, 用反证法. 设 M 中存在两个不同的最佳逼近元素 m_1,m_2, 使得 $d(x,M)=\|x-m_1\|=\|x-m_2\|$. 因 M 是 V 的线性子空间, 故 $(m_1+m_2)/2\in M$. 容易看出

$$d(x,M)\leqslant\left\|x-\frac{m_1+m_2}{2}\right\|\leqslant\frac{1}{2}\left\|x-m_1\right\|+\frac{1}{2}\left\|x-m_2\right\|=d(x,M),$$

即 $(m_1+m_2)/2$ 亦是 x 的最佳逼近元. 此外, 利用内积的平行四边形等式可得

$$\begin{aligned}d(x,M)^2&=\left\|x-\frac{m_1+m_2}{2}\right\|^2=\left\|\frac{x-m_1}{2}+\frac{x-m_2}{2}\right\|^2\\&=\frac{1}{2}\Big(\|x-m_1\|^2+\|x-m_2\|^2\Big)-\left\|\frac{x-m_1}{2}-\frac{x-m_2}{2}\right\|^2\\&=d(x,M)^2-\frac{1}{4}\|m_1-m_2\|^2.\end{aligned}$$

因此, $m_1=m_2$, 与假设矛盾, 故原命题正确. □

其次, 我们给出刻画内积空间最佳逼近元的特征性质.

定理 9.9 对任意的 $x \in V$, 则 $m^* \in M$ 为 x 的最佳逼近元的充分必要条件是 $x - m^*$ 与 M 中的任意元素正交, 即

$$(x - m^*, m) = 0, \quad \forall m \in M.$$

证明 (必要性) 用反证法. 假设存在某一元素 $m \in M$, 使得 $(x - m^*, m) \neq 0$. 显然, m 不为零元素. 若令

$$m^\circ = \frac{m}{\|m\|}, \quad \lambda = (x - m^*, m^\circ),$$

则有

$$\begin{aligned} \|x - m^* - \lambda m^\circ\|^2 &= (x - m^* - \lambda m^\circ, x - m^* - \lambda m^\circ) \\ &= \|x - m^*\|^2 - 2\lambda(x - m^*, m^\circ) + \lambda^2 \\ &= \|x - m^*\|^2 - \lambda^2 \\ &< \|x - m^*\|^2, \end{aligned}$$

与 m^* 为 x 的最佳逼近元的定义矛盾! 故假设错误, 原命题正确.

(充分性) 若对任意的 $m \in M$, 均有 $(x - m^*, m) = 0$ 成立, 则有

$$\begin{aligned} \|x - m\|^2 - \|x - m^*\|^2 &= \|m\|^2 - \|m^*\|^2 - 2(x, m) + 2(x, m^*) \\ &= \|m - m^*\|^2 + 2(x - m^*, m^* - m) \\ &= \|m - m^*\|^2 \geqslant 0. \end{aligned}$$

因此, m^* 为 x 的最佳逼近元. □

定理9.9的几何意义: x 在 M 中的正交投影 m^* 即为 x 的最佳逼近元. 利用该定理, 最佳逼近元的距离可表示为:

$$d(x, M)^2 = (x, x) - (x, m^*). \tag{9.3}$$

下面, 假设 M 是 X 的 n 维线性子空间, M 有一组基: $\varphi_1, \varphi_2, \cdots, \varphi_n$, 那么 x 的最佳逼近元 m^* 可表示为 $m^* = \sum_{i=1}^{n} c_i^* \varphi_i$. 利用定理9.9, m 分别取 $\varphi_1, \varphi_2, \cdots, \varphi_n$, 可得

$$\sum_{i=1}^{n} (\varphi_i, \varphi_j)\, c_i^* = (x, \varphi_j), \quad j = 1, 2, \cdots, n, \tag{9.4}$$

称式 (9.4) 为最佳逼近元的**法方程组**. 若引入记号

$$G=\begin{pmatrix}(\varphi_1,\varphi_1) & (\varphi_1,\varphi_2) & \cdots & (\varphi_1,\varphi_n)\\ (\varphi_2,\varphi_1) & (\varphi_2,\varphi_2) & \cdots & (\varphi_2,\varphi_n)\\ \vdots & \vdots & \ddots & \vdots\\ (\varphi_n,\varphi_1) & (\varphi_n,\varphi_2) & \cdots & (\varphi_n,\varphi_n)\end{pmatrix},\quad \mathbf{c}^*=\begin{pmatrix}c_1^*\\ c_2^*\\ \vdots\\ c_n^*\end{pmatrix},\quad \mathbf{b}=\begin{pmatrix}(x,\varphi_1)\\ (x,\varphi_2)\\ \vdots\\ (x,\varphi_n)\end{pmatrix},$$

则法方程组可写成 $G\,\mathbf{c}^*=\mathbf{b}$. 基于 $\{\varphi_i\}_{i=1}^n$ 的线性无关性, 容易证明矩阵 G 是正定的. 因此, 法方程组的解存在且唯一.

进一步, 若 $\varphi_1,\varphi_2,\cdots,\varphi_n$ 构成 M 的一组正交基, 则 G 是一个对角矩阵, 法方程组可以直接解出, 最佳逼近元 m^* 显式地表示为

$$m^*=\sum_{i=1}^{n}\frac{(x,\varphi_i)}{(\varphi_i,\varphi_i)}\varphi_i, \tag{9.5}$$

称式 (9.5) 为 x 的**广义 Fourier 展开**, φ_i 的系数为**广义 Fourier 系数**. 利用 $\{\varphi_i\}_{i=1}^n$ 的正交性, 可知式 (9.3) 等价于

$$\|x-m^*\|^2=\|x\|^2-\sum_{i=1}^{n}(c_i^*)^2\|\varphi_i\|^2,\quad c_i^*=\frac{(x,\varphi_i)}{(\varphi_i,\varphi_i)}.$$

在上式中, 若令 $n\to\infty$, 则得 **Bessel 不等式**:

$$\sum_{i=1}^{\infty}(c_i^*)^2\|\varphi_i\|^2\leqslant\|x\|^2.$$

特别地, 若最佳逼近元序列收敛于 x, 则上述不等式变成等式, 称为**广义 Parseval 等式**.

最后, 我们讨论有限维内积空间正交基的存在性.

定理 9.10 任何 n 维内积空间 M 都存在正交基.

证明 因 M 是有限维线性空间, 故存在一组基 $\varphi_1,\varphi_2,\cdots,\varphi_n$. 接下来, 通过如下算法构造正交基 $e_1,e_2,\cdots,e_n$:

(1) $e_1=\varphi_1$;

(2) $e_i=\varphi_i-\sum_{j=1}^{i-1}\frac{(\varphi_i,e_j)}{(e_j,e_j)}e_j,\quad j=2,\cdots,n.$

先证 $e_i \neq 0$. 若不然, 则 φ_i 与 $e_1, e_2, \cdots, e_{i-1}$ 线性相关, 与 $\varphi_1, \varphi_2, \cdots, \varphi_n$ 是一组基相矛盾. 接着, 用归纳法证明 $e_1, e_2, \cdots, e_n$ 相互正交. 容易验证: e_2 与 e_1 正交. 对于一般的 e_i, 利用归纳假设, 则有

$$(e_i, e_k) = (\varphi_i, e_k) - \sum_{j=1}^{i-1} \frac{(\varphi_i, e_j)}{(e_j, e_j)}(e_j, e_k) = (\varphi_i, e_k) - \frac{(\varphi_i, e_k)}{(e_k, e_k)}(e_k, e_k) = 0,$$

其中 $k = 1, 2, \cdots, i-1$. 定理得证. □

若将 $\{e_i\}_{i=1}^n$ 进一步规范化, 则得 M 的一组标准正交基. 证明过程中用到的算法称为 **Gram-Schmidt 正交化**, 是一个非常重要的构造性算法. 下面针对一些重要内积空间的最佳逼近问题作更深入的讨论.

9.3 最小二乘法

在科学研究和工程应用领域, 经常会遇到函数 $f(x)$ 的表达式未知或者难以用初等函数表示的情况, 但是可以通过观察或测量的方式获得 $f(x)$ 的一组离散值

$$\big\{(x_i, y_i) : y_i = f(x_i)\big\}_{i=1}^n,$$

即观测数据. 那么如何根据这些数据, 来确定自变量 x 和因变量 y 之间的关系? 一种方式是使用插值的方法, 即

$$\varphi(x_i) = f(x_i), \quad i = 1, 2, \cdots, n,$$

来构造函数 $\varphi(x)$ 以逼近未知函数 $f(x)$. 但是, 该方法存在以下两个问题:

(1) 当数据有误差时, 用插值方式来构造函数不能合理地处理误差带来的影响;

(2) 当数据量很大时, 用多项式插值容易出现数值不稳定的情况.

另一种方式是使用拟合的方法, 要求 $\varphi(x)$ 与 $f(x)$ 之间的误差或距离最小, 即

$$\min_{\varphi \in \Phi} \|\varphi - f\|, \quad i = 1, 2, \cdots, n, \tag{9.6}$$

其中 $\|\cdot\|$ 表示函数空间 Φ 的某种范数. 明显地, 这是一个函数逼近问题.

在数据拟合问题 (9.6) 中, 选用函数空间 $L^p[a, b]$ 及范数 $\|\cdot\|_p$ 是合适的. 但是仍存在一个问题, 即函数 $f(x)$ 是信息不全的, 仅知道 $f(x)$ 在某些观测点的值. 因此, 我们可

选择使用式 (9.1) 的离散逼近来刻画 $f(x)$ 与 $\varphi(x)$ 之间的误差或距离, 即

$$\|\varphi - f\|_p \triangleq \left(\sum_{i=1}^{n} \left| \varphi(x_i) - f(x_i) \right|^p \right)^{\frac{1}{p}}. \tag{9.7}$$

在实际应用中, p 常取 1, 2 或 $+\infty$. 当 $p = 2$ 时, 最优化问题 (9.6) 称为**最小二乘问题**. 最小二乘法源于天文学和测地学的应用需要, 分别由德国数学家高斯 (C. F. Gauss) 和法国数学家勒让德 (A. M. Legendre) 独立提出, 如今被广泛应用于统计学、逼近论和控制论等领域. 当 $p = 1$ 或 $p = +\infty$ 时, 最优化问题 (9.6) 的求解变得复杂了, 因为目标函数是不可微的. 但是, 近年来随着压缩感知与稀疏表示技术的发展, 已经涌现了一些较好的数值方法用于求解这类优化问题, 感兴趣的读者可参见文献 [10].

在数据拟合问题 (9.6) 中, 另一个关键的问题是函数空间 Φ 的选取, 即问题的数学模型. 一般来说, 函数空间 Φ 主要靠人们的专业知识或工作经验来决定. 常用的函数空间 Φ 包括: 多项式空间 $\mathbb{P}_n[x]$、样条函数空间 $\mathcal{S}(\mathbb{P}_m[x], \mathbb{M}, \Delta)$ 或经验函数空间 $\left\{ ae^{bx} \,|\, a, b \in \mathbb{R} \right\}$ 等.

为了简化问题, 下面假设 Φ 是线性空间, 函数组 $\{\varphi_i(x)\}_{i=1}^m$ 构成它的一组基, 取 $p = 2$, 那么问题 (9.6) 可写成

$$\min_{\varphi \in \Phi} \left\| \varphi - f \right\|_2^2 = \min_{c_1, c_2, \cdots, c_m \in \mathbb{R}} \sum_{i=1}^{n} \left[\sum_{j=1}^{m} c_j \varphi_j(x_i) - f(x_i) \right]^2, \tag{9.8}$$

称为**数据的最小二乘拟合问题**. 因为 $p = 2$, 所以可用内积空间的最佳逼近理论来处理. 不难验证, 式 (9.7) 的范数 $\|\cdot\|_2$ 是离散内积

$$(f, g) \triangleq \sum_{i=1}^{n} f(x_i)\, g(x_i) \tag{9.9}$$

诱导的拟范数. 因此, 根据定理9.9, 最小二乘拟合问题可通过**法方程组**求解, 即

$$G\,\mathbf{c} = \mathbf{b}, \quad G = (g_{ij}) \in \mathbb{R}^{m \times m}, \quad \mathbf{c} = (c_i) \in \mathbb{R}^m, \quad \mathbf{b} = (b_i) \in \mathbb{R}^m, \tag{9.10}$$

其中 $g_{ij} = \sum_{l=1}^{n} \varphi_i(x_l)\varphi_j(x_l)$, $b_i = \sum_{l=1}^{n} \varphi_i(x_l) f(x_l)$.

例 9.7 澳大利亚生物学家 P. Sale 和 R. Dybdall 两年间在某处做的鱼类抽样调查如下表所示:

x_i	13	15	16	21	22	23	25	29	30	31	16
y_i	11	10	11	12	12	13	13	12	14	16	17
x_i	40	42	55	60	62	64	70	72	100	130	
y_i	13	14	22	14	21	21	24	17	23	34	

表 9.1: 鱼的数量和种类, 其中 x 为鱼的数量, y 为鱼的种类

试用线性函数拟合鱼的数量和种类之间的关系.

解 设 $\varphi(x) = c_0 + c_1 x$, 则线性函数拟合的法方程组为

$$\begin{pmatrix} m & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} x_i^2 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} x_i y_i \end{pmatrix}.$$

将表9.1中的数据代入并求解得

$$\begin{pmatrix} 21 & 956 \\ 956 & 61640 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \end{pmatrix} = \begin{pmatrix} 344 \\ 18913 \end{pmatrix} \Longrightarrow \begin{pmatrix} c_0 \\ c_1 \end{pmatrix} \approx \begin{pmatrix} 8.20841 \\ 0.17952 \end{pmatrix},$$

即 $\varphi(x) = 8.20841 + 0.17952\,x$, 输入数据和拟合函数如图9.1所示. □

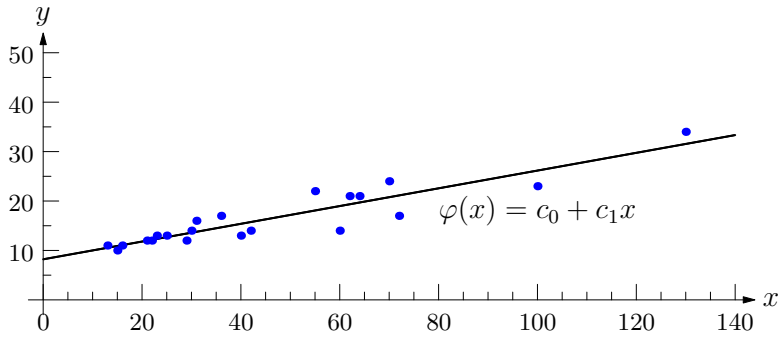


图 9.1: 线性函数拟合

更多的例子...

前面主要讨论了函数 y 依赖于单个变量的情形. 如果 y 是多个自变量 $x_1, x_2, \cdots, x_m$ 的函数, 那么可以把这些自变量本身视为一组基, 建立如下模型:

$$y \approx c_1 x_1 + c_2 x_2 + \cdots + c_m x_m,$$

来预测 y, 其中系数 $\{c_i\}_{i=1}^m$ 可利用最小二乘法求解, 即

$$\min_{c_1, c_2, \cdots, c_m \in \mathbb{R}} \sum_{j=1}^{n} \left[\sum_{i=1}^{m} c_i x_{ij} - y_j \right]^2.$$

在数理统计中, 上述方法又称为**多元线性回归**.

例 9.8 为了开拓市场, 某公司对其新产品作了一系列调查, 他们发现这一新产品的销量与下列事件关系密切: 其一是温度, 其二是上证指数, 其三是广告费, 其四是推销员数, 其五是返修率, 详细数据见下表.

记录	温度	上证指数	广告费	推销员数	返修率	销售量
1	39	567	10000	2	0.20	75
2	37	679	0	3	0.15	68
3	30	346	5000	3	0.10	105
4	25	987	5000	3	0.08	136
5	25	1101	0	4	0.07	152
6	10	1004	5000	5	0.07	191
7	15	667	0	4	0.05	148
8	5	604	10000	6	0.04	234

假设这些事件与销量量近似成线性关系, 试求这种关系的数学表达式.

解 设销售量为 y, 用 x_1, x_2, x_3, x_4, x_5 分别表示温度、上证指数、广告费、推销员数、返修率, 则需确定函数

$$y = c_1 x_1 + c_2 x_2 + c_3 x_3 + c_4 x_4 + c_5 x_5.$$

利用最小二乘法, 系数 $c_1, c_2, \cdots, c_5$ 需要满足法方程组:

$$\begin{pmatrix} 5390 & 132881 & 765000 & 594 & 21.75 \\ 132881 & 4906337 & 23395000 & 22886 & 533.67 \\ 765000 & 23395000 & 275000000 & 135000 & 3650 \\ 594 & 22886 & 135000 & 124 & 2.46 \\ 21.75 & 533.67 & 3650 & 2.46 & 0.0928 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{pmatrix} = \begin{pmatrix} 21091 \\ 858427 \\ 5250000 \\ 4636 \\ 87.35 \end{pmatrix}$$

解得:

$$y \approx 0.68772x_1 + 0.043703x_2 + 0.0048631x_3 + 29.607x_4 - 447.36x_5.$$

□

接下来, 讨论矛盾方程组的最小二乘解. 若记

$$F(c_1, c_2, \cdots, c_m) \triangleq \sum_{i=1}^{n} \left[\sum_{j=1}^{m} c_j \varphi_j(x_i) - f(x_i) \right]^2,$$

则 $F(c_1, c_2, \cdots, c_m)$ 是关于变量 $c_1, c_2, \cdots, c_m$ 的多元二次函数. 令

$$A = \begin{pmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \cdots & \varphi_m(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \cdots & \varphi_m(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(x_n) & \varphi_2(x_n) & \cdots & \varphi_m(x_n) \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \cdots \\ f(x_n) \end{pmatrix},$$

则

$$F(c_1, c_2, \cdots, c_m) = \left\| A\,\mathbf{c} - \mathbf{f} \right\|_2^2 = \mathbf{c}^T A^T A \mathbf{c} - 2\mathbf{c}^T A^T \mathbf{f} + \mathbf{f}^T \mathbf{f}.$$

一般地, 当 $n \gg m$ 时, 若线性方程组 $A\,\mathbf{c} = \mathbf{f}$ 无解, 则称它为**矛盾方程组**. 对于矛盾方程组, 称取到

$$\min_{\mathbf{c} \in \mathbb{R}^m} \left\| A\,\mathbf{c} - \mathbf{f} \right\|_2^2 \tag{9.11}$$

的 $\mathbf{c}$ 为**矛盾方程的最小二乘解**. 容易看出, 矛盾方程的最小二乘解是内积空间 $\mathbb{R}^n$ 中的最佳平方逼近问题, 即在 A 的列空间中选取元素最佳逼近 $\mathbb{R}^n$ 中的元素 $\mathbf{f}$. 此外, 数据拟合的最小二乘问题 (9.8) 可以视为矛盾方程组的最小二乘解. 从前面的讨论可知, 矛盾方程的最小二乘解可通法方程组求解, 即

$$G\,\mathbf{c} = \mathbf{b} \Longleftrightarrow A^{\mathrm{T}} A\,\mathbf{c} = A^{\mathrm{T}}\,\mathbf{f}.$$

注解 9.1 利用多元二次函数的极值理论, 亦可导出求解矛盾方程最小二乘解的法方程组. 事实上, 多元二次函数 $F(c_1, c_2, \cdots, c_m)$ 取极值的必要条件为

$$\frac{\partial F}{\partial \mathbf{c}} = \begin{pmatrix} \frac{\partial F}{\partial c_1} \\ \frac{\partial F}{\partial c_2} \\ \vdots \\ \frac{\partial F}{\partial c_m} \end{pmatrix} = 2A^T A\mathbf{c} - 2A^T\mathbf{f} = \mathbf{0} \Longleftrightarrow A^{\mathrm{T}}A\,\mathbf{c} = A^{\mathrm{T}}\,\mathbf{f},$$

即驻点条件. 又因为 $A^{\mathrm{T}}A$ 是半正定的, 所以函数 $F(c_1, c_2, \cdots, c_m)$ 在驻点取到最小值.

关于矛盾方程的最小二乘解有下面的结论.

定理 9.11 设 $A \in \mathbb{R}^{n\times m}$, $\mathbf{f} \in \mathbb{R}^n$, 则

(1) 线性方程组 $A^{\mathrm{T}}A\,\mathbf{x} = A^{\mathrm{T}}\,\mathbf{f}$ 恒有解;

(2) 函数 $\left\|A\,\mathbf{x} - \mathbf{f}\right\|_2$ 取到最小值 $\Longleftrightarrow$ $\mathbf{x}$ 满足 $A^{\mathrm{T}}A\,\mathbf{x} = A^{\mathrm{T}}\,\mathbf{f}$;

(3) 线性方程组 $A^{\mathrm{T}}A\,\mathbf{x} = A^{\mathrm{T}}\,\mathbf{f}$ 的解是唯一的 $\Longleftrightarrow$ $\operatorname{rank} A = n$.

更多的例子...

最后, 利用矩阵 A 的秩来分析最小二乘问题解集的性质:

(1) 当 $\operatorname{rank}(A) = m = n$ 时, 即 A 是可逆方阵. 最小二乘问题 (9.11) 与线性方程组 $A\mathbf{c} = \mathbf{f}$ 同解, 且解是唯一的;

(2) 当 $\operatorname{rank}(A) = m < n$ 时, 即 A 是列满秩的. 因为 $\operatorname{rank}(A^{\mathrm{T}}A) = \operatorname{rank}(A) = m$, 所以 $A^{\mathrm{T}}A$ 可逆, 最小二乘问题的解也是唯一的;

(3) 当 $\operatorname{rank}(A) < \min\{m, n\}$ 时, 即 A 是秩亏损的. 此时, 最小二乘问题有无穷多组解.

对于前两种情形, 最小二乘问题的解可以通过求解线性方程组得到, 常用的方法包括: Cholesky 分解、QR 分解等方法. 而对于最后一种情形, 必须采用特殊的求解方法, 且需要考虑确定数值秩这一难题, 一个常用的方法是使用奇异值分解 (Singular Value Decomposition, SVD), 具体的内容可参见文献 [11].

9.4 最佳平方逼近与正交多项式

本节主要讨论函数的最佳多项式逼近问题, 即连续情形的最佳平方逼近问题. 记 $L_\rho^2[a,b]$ 是区间 $[a,b]$ 上满足

$$\int_a^b \rho(x)f^2(x)\,\mathrm{d}x < +\infty$$

的 Lebesgue 可积函数 f 构成的函数类, 其中称非负函数 $\rho(x)$ 为**权函数**, 如果 $\rho(x)$ 满足:

(1) 对于非负整数 n, 积分 $\int_a^b \rho(x)x^n\,\mathrm{d}x$ 存在且有限;

(2) 对于非负连续函数 $g(x)$, 若有 $\int_a^b \rho(x)g(x)\,\mathrm{d}x = 0$, 则 $g(x)\Big|_{[a,b]} \equiv 0$.

显然, 当 $\rho(x) \equiv 1$ 时, 即为 $L^2[a,b]$. 在 $L_\rho^2[a,b]$ 中, 定义

$$(f,g) = \int_a^b \rho(x)f(x)g(x)\,\mathrm{d}x, \quad \forall f,g \in L_\rho^2[a,b],$$

$$\|f\| = \sqrt{(f,f)}, \quad \forall f \in L_\rho^2[a,b].$$

不难验证, $(\cdot,\cdot)$ 与 $\|\cdot\|$ 分别构成 $L_\rho^2[a,b]$ 的内积与范数.

令 $\mathbb{P}_n[x]$ 为所有次数不超过 n 的多项式构成的空间, 则 $\mathbb{P}_n[x]$ 是 $L_\rho^2[a,b]$ 的 $n+1$ 维子空间. 利用定理9.8知, 对任意的 $f \in L_\rho^2[a,b]$, 存在唯一的 n 次多项式

$$p(x) = \sum_{i=0}^n c_i x^i \in \mathbb{P}_n[x],$$

使得 $\|f-p\|$ 取到最小值, 即**最佳平方逼近多项式**. 多项式 $p(x)$ 的系数可由如下法方程确定:

$$\sum_{i=0}^n (x^j,x^i)\,c_i = (f,x^j), \quad j = 0,1,\cdots,n, \tag{9.12}$$

其中

$$(x^j,x^i) = \int_a^b \rho(x)x^{i+j}\,\mathrm{d}x, \quad (f,x^j) = \int_a^b \rho(x)f(x)x^j\,\mathrm{d}x.$$

例 9.9 设 $f(x) = \sin(\pi x)$, 求 $f(x)$ 在区间 $[0,1]$ 上的二次最佳平方逼近多项式.

解 取函数空间 $M=\operatorname{span}\{1,x,x^2\}$, 由定理9.8知, 二次最佳平方逼近多项式 $p_2^*(x)$ 存在且唯一. 据式 (9.12), 取权函数 $\rho(x)=1$, 容易算得

$$
\begin{gathered}
(\varphi_0,\varphi_0)=1,\quad (\varphi_0,\varphi_1)=1/2,\quad (\varphi_0,\varphi_2)=1/3,\\
(\varphi_1,\varphi_0)=1/2,\quad (\varphi_1,\varphi_1)=1/3,\quad (\varphi_1,\varphi_2)=1/4,\\
(\varphi_2,\varphi_0)=1/3,\quad (\varphi_2,\varphi_1)=1/4,\quad (\varphi_2,\varphi_2)=1/5,\\
(f,\varphi_0)=\int_0^1 \sin(\pi x)\,\mathrm{d}x=\frac{2}{\pi},\\
(f,\varphi_1)=\int_0^1 x\sin(\pi x)\,\mathrm{d}x=\frac{1}{\pi},\\
(f,\varphi_2)=\int_0^1 x^2\sin(\pi x)\,\mathrm{d}x=\frac{\pi^2-4}{\pi^3},
\end{gathered}
$$

故 $p_2^*(x)=c_0+c_1x+c_2x^2$ 满足法方程

$$
\begin{pmatrix}
1 & \frac{1}{2} & \frac{1}{3}\\
\frac{1}{2} & \frac{1}{3} & \frac{1}{4}\\
\frac{1}{3} & \frac{1}{4} & \frac{1}{5}
\end{pmatrix}
\begin{pmatrix}
c_0\\ c_1\\ c_2
\end{pmatrix}
=
\begin{pmatrix}
\frac{2}{\pi}\\ \frac{1}{\pi}\\ \frac{\pi^2-4}{\pi^3}
\end{pmatrix}
$$

求解出

$$
c_0=\frac{12\pi^2-120}{\pi^3}\approx -0.050465,\ c_1=\frac{720-60\pi^2}{\pi^3}\approx 4.12251,\ c_2=-c_1\approx -4.12251.
$$

因此, $p_2^*(x)\approx -0.050465+4.12251x-4.12251x^2$, 逼近情况如图9.2所示. □

当 $[a,b]=[0,1]$, $\rho(x)=1$ 时, 法方程组 (9.12) 的系数矩阵是

$$
H=\begin{pmatrix}
1 & \frac{1}{2} & \cdots & \frac{1}{n}\\
\frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n+1}\\
\vdots & \vdots & \ddots & \vdots\\
\frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n+1}
\end{pmatrix},
$$

H 称为希尔伯特 (Hilbert) 矩阵. 当 n 较大时, 可以证明 H 是病态的. 因此, 对于许多实际问题来说, 通过法方程来确定函数的最佳平方逼近多项式是有困难的. 然而, 通过分

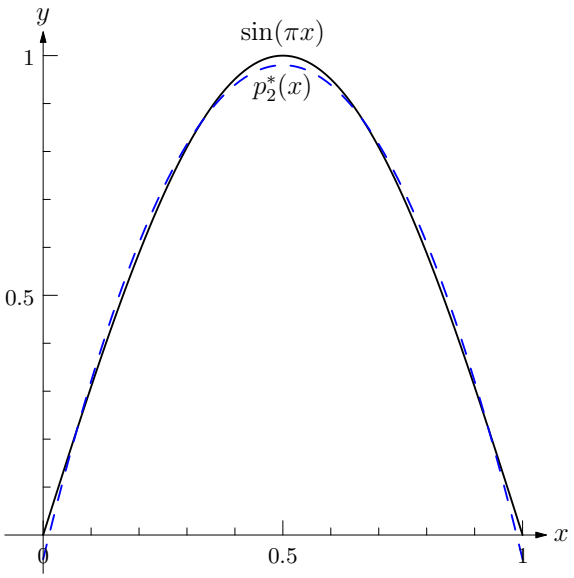


图 9.2: 函数 $\sin(\pi x)$ 在区间 $[0,1]$ 上的二次最佳平方逼近

析不难发现, 如果能找到 $\mathbb{P}_n[x]$ 的一组正交基, 即正交多项式, 那么法方程组 (9.12) 的系数矩阵就是对角矩阵, 线性方程组便可直接求解, 即式 (9.5), 从而保证了计算的可靠性. 另外, 正交多项式本身在数值积分, 微分方程, 数学物理方法, 编码理论等领域也有重要的应用.

定义 9.9 定义在 $[a,b]$ 上的函数系 $\{g_l(x)\}_{l=0}^n$ 称为 $\mathbb{P}_n[x]$ 的一组**正交多项式基**, 如果它满足以下条件

(1) $g_l(x)$ 是 l 次多项式, 即 $g_l(x) = a_l x^l + a_{l-1}x^{l-1} + \cdots + a_0$, $a_l \neq 0$;

(2) $(g_i, g_j) = 0$, $\forall i \neq j$; $\quad (g_i, g_i) > 0$, $i = 0, 1, \cdots, n$,

其中 $g_l(x)$ 称为**l 次正交多项式**. 进一步, 若有 $(g_i, g_i) = 1$, $i = 0, 1, \cdots, n$, 则称 $\{g_l(x)\}_{l=0}^n$ 是 $[a,b]$ 上 $\mathbb{P}_n[x]$ 一组**规范正交多项式基**.

利用正交性, 对于任意的 k 次多项式 $p_k(x)$, 则有

$$(p_k, g_l) = 0, \quad l > k.$$

令 $g_l^*(x) = g_l(x)/a_l$, $l = 0, 1, \cdots, n$, 称 $\{g_l^*(x)\}_{l=0}^n$ 为 $[a,b]$ 上 $\mathbb{P}_n[x]$ 一组**首项系数为 1 的正交多项式基**. 下面给出它的计算公式.

定理 9.12 正交多项式基 $\{g_l^*(x)\}_{l=0}^n$ 有递推公式:

$$\begin{cases} g_0^*(x)=1, \quad g_1^*(x)=x-\dfrac{(xg_0^*,g_0^*)}{(g_0^*,g_0^*)}, \\ g_k^*(x)=(x-a_k)g_{k-1}^*(x)-b_kg_{k-2}^*(x), \quad k=2,3,\cdots,n, \end{cases} \tag{9.13}$$

其中 $a_k=\dfrac{(xg_{k-1}^*,g_{k-1}^*)}{(g_{k-1}^*,g_{k-1}^*)}$, $b_k=\dfrac{(xg_{k-1}^*,g_{k-2}^*)}{(g_{k-2}^*,g_{k-2}^*)}$.

证明 从递推关系式 (9.13), 不难看出 $g_k^*(x)$ 是首项系数为 1 的多项式, 故均不为零. 于是在 a_k 与 b_k 公式中的分母皆非零. 下面用数学归纳法来证明 $(g_n^*,g_i^*)=0$, 对任意的整数 $i<n$ 成立.

当 $n=1$ 时, 容易验证

$$\begin{aligned}(g_1^*,g_0^*)&=(x-a_1,g_0^*)=(x,g_0^*)-(a_1,g_0^*)\\&=(x,g_0^*)-(xg_0^*,g_0^*)=0.\end{aligned}$$

假设命题对 $n-1$ 成立, 那么

$$\begin{aligned}(g_n^*,g_{n-1}^*)&=(xg_{n-1}^*-a_ng_{n-1}^*-b_ng_{n-2}^*,g_{n-1}^*)\\&=(xg_{n-1}^*,g_{n-1}^*)-a_n(g_{n-1}^*,g_{n-1}^*)=0.\end{aligned}$$

类似地有

$$\begin{aligned}(g_n^*,g_{n-2}^*)&=(xg_{n-1}^*-a_ng_{n-1}^*-b_ng_{n-2}^*,g_{n-2}^*)\\&=(xg_{n-1}^*,g_{n-2}^*)-b_n(g_{n-2}^*,g_{n-2}^*)=0.\end{aligned}$$

而对于 $i<n-2$, 我们有

$$\begin{aligned}(g_n^*,g_i^*)&=(xg_{n-1}^*-a_ng_{n-1}^*-b_ng_{n-2}^*,g_i^*)=(xg_{n-1}^*,g_i^*)\\&=(g_{n-1}^*,xg_i^*)=(g_{n-1}^*,g_{i+1}^*+a_{i+1}g_i^*+b_{i+1}g_{i-1}^*)=0.\end{aligned}$$

从而定理得证. □

下面, 我们证明一个非常有用的结论.

定理 9.13 若 $f(x)$ 为 $[a,b]$ 上任一连续函数, 与 $g_0^*(x),g_1^*(x),\cdots,g_n^*(x)$ 都正交, 则 $f(x)$ 在 (a,b) 中至少变号 $n+1$ 次或者恒等于零.

证明 因 $f \perp g_0^*$, 且 $g_0^* = 1$, 故

$$\int_a^b \rho(x) f(x)\,\mathrm{d}x = 0.$$

于是, 若 $f \neq 0$, 则 $f(x)$ 在 (a,b) 中至少变号一次.

若 $f(x)$ 在 (a,b) 中变号 k 次, 且 $k < n+1$. 令 $x_1 < x_2 < \cdots < x_k$ 为 (a,b) 中 $f(x)$ 发生变号的点, 则在每个区间 $(a,x_1),(x_1,x_2),\cdots,(x_k,b)$ 中 $f(x)$ 不变号, 但在相邻的区间内符号相反. 令 k 次多项式 $p(x) = \prod_{i=1}^{k}(x - x_i)$, 显然 $p(x)$ 亦具有此性质. 于是

$$\int_a^b \rho(x) f(x) p(x)\,\mathrm{d}x \neq 0.$$

另一方面, 因 $f(x)$ 与 $g_0^*(x), g_1^*(x), \cdots, g_n^*(x)$ 都正交, 故 $f(x)$ 与 $p(x)$ 正交, 这与上式相矛盾. □

推论 9.14 若 $f(x)$ 为 $[a,b]$ 上任一连续函数, $p(x)$ 是 $f(x)$ 的 n 次最佳平方逼近多项式, 则 $p(x)$ 在 (a,b) 中至少 $n+1$ 个点插值于 $f(x)$.

证明 因 $p(x)$ 是 $f(x)$ 的最佳平方逼近多项式, 利用定理 (9.9), 知 $(p-f)(x)$ 与 $g_0^*(x), g_1^*(x), \cdots, g_n^*(x)$ 都正交. 利用定理 (9.13), 推论得证. □

与一般的多项式不同, 正交多项式的零点具有很好的性质.

定理 9.15 区间 $[a,b]$ 上的 l 次正交多项 $g_l^*(x)$ 恰有 l 个互异的实根, 并且全部位于 $[a,b]$ 的内部.

证明 因 $g_l^*(x)$ 与 $g_0^*(x), g_1^*(x), \cdots, g_{l-1}^*(x)$ 都正交, 且 $g_l^*(x)$ 是首项系数为 1 的多项式, 利用定理 (9.13), 知 $g_l^*(x)$ 在 (a,b) 中至少变号 l 次. 又利用代数基本定理, 知 $g_l^*(x)$ 在复数域 $\mathbb{C}$ 具有 l 个根. 于是, 推论得证. □

最后, 给出几类常用的带权正交多项式.

1) 勒让德 (Legendre) 多项式

设 $\rho(x) \equiv 1$, 区间 $[-1,1]$ 上 $\mathbb{P}_n(x)$ 的正交基 $\{P_k(x)\}_{k=0}^n$, 称 $P_k(x)$ 为**勒让德多项式**. 利用正交多项式的定义, 可写出 $P_k(x)$ 的解析表达式:

$$P_k(x) = \frac{1}{2^k k!}\frac{\mathrm{d}^k}{\mathrm{d}x^k}\left[(x^2-1)^k\right], \quad k = 0,1,\cdots,n. \tag{9.14}$$

实际应用中, 因为求高阶导数较麻烦, 所以使用下面递推公式计算:

$$P_{k+1}(x)=\frac{2k+1}{k+1}xP_k(x)-\frac{k}{k+1}P_{k-1}(x),\quad k=1,2,\cdots,n. \tag{9.15}$$

利用上式, 不难证明, 当 k 为偶数时, $P_k(x)$ 为偶函数; 当 k 为奇数时, $P_k(x)$ 为奇函数.

2) 第一类切比雪夫 (Chebyshev) 多项式

设 $\rho(x)=(1-x^2)^{-1/2}$, 区间 $[-1,1]$ 上 $\mathbb{P}_n(x)$ 的正交基 $\{T_k(x)\}_{k=0}^n$, 称 $T_k(x)$ 为**第一类切比雪夫多项式**. 利用正交多项式的定义, 可写出 $T_k(x)$ 的解析表达式:

$$T_k(x)=\cos(k\arccos x),\quad k=0,1,\cdots,n. \tag{9.16}$$

由三角恒等式 $\cos k\theta+\cos(k-2)\theta=2\cos\theta\cos(k-1)\theta$, 可得 $T_k(x)$ 的递推计算公式:

$$\begin{cases}T_{k+1}(x)=2xT_k(x)-T_{k-1}(x),\quad k=1,2,\cdots,n,\\ T_0(x)=1,\quad T_1(x)=x.\end{cases} \tag{9.17}$$

利用数学归纳法, 容易证明, $T_k(x)$ 是首项系数为 2^{k-1} 的 k 次多项式, 且 T_{2k} 只含 x 的偶次幂, T_{2k-1} 只含 x 的奇次幂.

3) 第二类切比雪夫 (Chebyshev) 多项式

设 $\rho(x)=(1-x^2)^{1/2}$, 区间 $[-1,1]$ 上 $\mathbb{P}_n(x)$ 的正交基 $\{U_k(x)\}_{k=0}^n$, 称 $U_k(x)$ 为**第二类切比雪夫多项式**. 第二类切比雪夫多项式的解析表达式

$$U_k(x)=\frac{\sin[(k+1)\arccos x]}{\sqrt{1-x^2}},\quad k=0,1,\cdots,n. \tag{9.18}$$

$U_k(x)$ 的递推计算公式:

$$\begin{cases}U_{k+1}(x)=2xU_k(x)-U_{k-1}(x),\quad k=1,2,\cdots,n,\\ U_0(x)=1,\quad U_1(x)=2x.\end{cases} \tag{9.19}$$

4) 拉盖尔 (Laguerre) 多项式

设 $\rho(x)=e^{-x}$, 区间 $[0,+\infty)$ 上 $\mathbb{P}_n(x)$ 的正交基 $\{L_k(x)\}_{k=0}^n$, 称 $L_k(x)$ 为**拉盖尔多项式**. 拉盖尔多项式的解析表达式

$$L_k(x)=e^x\frac{\mathrm{d}^k}{\mathrm{d}x^k}(x^ke^{-x}),\quad k=0,1,\cdots,n. \tag{9.20}$$

$L_k(x)$ 的递推计算公式:

$$\begin{cases} L_{k+1}(x) = (2k+1-x)L_k(x) - k^2 L_{k-1}(x), \quad k = 1, 2, \cdots, n, \\ L_0(x) = 1, \quad L_1(x) = 1 - x. \end{cases} \tag{9.21}$$

5) 埃尔米特 (Hermite) 多项式

设 $\rho(x) = e^{-x^2}$, 区间 $(-\infty, +\infty)$ 上 $\mathbb{P}_n(x)$ 的正交基 $\{H_k(x)\}_{k=0}^n$, 称 $H_k(x)$ 为**埃尔米特多项式**. 埃尔米特多项式的解析表达式

$$H_k(x) = (-1)^n e^{x^2} \frac{\mathrm{d}^k}{\mathrm{d}x^k}(e^{-x^2}), \quad k = 0, 1, \cdots, n. \tag{9.22}$$

$L_k(x)$ 的递推计算公式:

$$\begin{cases} H_{k+1}(x) = 2xH_k(x) - 2kH_{k-1}(x), \quad k = 1, 2, \cdots, n, \\ L_0(x) = 1, \quad L_1(x) = 2x. \end{cases} \tag{9.23}$$

关于正交多项式及其应用的更多内容, 可参阅文献 [8].

例 9.10 设 $f(x) = \sin(\pi x)$, 利用正交多项式理论, 求 $f(x)$ 在区间 $[0, 1]$ 上的二次最佳平方逼近多项式.

解 令 $t = 2x - 1$, 则 $t \in [-1, 1]$, 函数 f 可写成关于 t 的函数, 即

$$f(t) = \sin\Big[\frac{(t+1)\pi}{2}\Big].$$

根据勒让德多项式的定义, 有

$$P_0(t) = 1, \quad P_1(t) = t, \quad P_2(t) = \frac{3t^2 - 1}{2}.$$

容易算得

$$(P_0, P_0) = 2, \quad (P_1, P_1) = 2/3, \quad (P_2, P_2) = 2/5,$$

$$(f, P_0) = \int_{-1}^{1} \sin\Big[\frac{(t+1)\pi}{2}\Big]\,\mathrm{d}t = \frac{4}{\pi},$$

$$(f, P_1) = \int_{-1}^{1} t\sin\Big[\frac{(t+1)\pi}{2}\Big]\,\mathrm{d}t = 0,$$

$$(f, P_2) = \int_{-1}^{1} \frac{3t^2 - 1}{2}\sin\Big[\frac{(t+1)\pi}{2}\Big]\,\mathrm{d}t = \frac{4(\pi^2 - 12)}{\pi^3}.$$

按式 (9.5) 有

$$\begin{aligned} p_2^*(t) &= \frac{4}{\pi} \times \frac{1}{2} \times P_0(t) + 0 \times \frac{3}{2} \times P_1(t) + \frac{4(\pi^2-12)}{\pi^3} \times \frac{5}{2} \times P_2(t) \\ &= \frac{2}{\pi} + \frac{5(\pi^2-12)(3t^2-1)}{\pi^3}. \end{aligned}$$

将 $t = 2x - 1$ 代入上式并化简得

$$p_2^*(x) = \frac{12\pi^2-120}{\pi^3} - \frac{60\pi^2-720}{\pi^3}x + \frac{60\pi^2-720}{\pi^3}x^2,$$

与例9.9的结果一致. □

通过对比, 容易看出利用正交多项式理论可以避免求解法方程组, 从而适用于法方程组的系数是病态的情形.

9.5 周期函数的最佳平方逼近与快速傅立叶变换

在科学与工程领域, 傅立叶分析作为最基本的数学工具, 被广泛应用于信号的时频域分析、谱分析、微分方程求解等. 法国数学家傅立叶 (Fourier) 出生于 1768 年, 他具有代表性的一项工作是研究热的传播和扩散现象. 在此过程中, 他发现在表示一个物体的温度分布时, 三角函数级数非常有用. 不仅如此, 他大胆断言: “任何” 周期函数都可以表示为三角函数级数的形式, 即**傅立叶级数**. 对于非周期信号, 傅立叶指出可用三角函数的加权积分来表示, 即**傅立叶变换**. 傅立叶变换与傅立叶级数有一个共同的重要性质, 即可以通过反变换恢复原来的表示.

首先, 我们讨论连续情形的傅立叶级数. 考虑区间 $[0,T)$ 上的平方可积函数空间 $L^2[0,T)$, 使用式 (9.2) 定义的内积及其诱导的范数. 若规定 $f(x) = f(x-T), \forall x \in \mathbb{R}$, 则可将 $L^2[0,T)$ 中的任一函数延拓为实数域 $\mathbb{R}$ 上周期为 T 的函数. 在此意义下, $L^2[0,T)$ 称为**周期为 T 的平方可积函数空间**.

若取 $[0,T)$ 上的逼近函数空间 $M = \text{span}\{\varphi_0, \varphi_1, \cdots, \varphi_{2n}\}$, 其中

$$\varphi_0 = \frac{1}{2}, \quad \varphi_{2k} = \cos k\omega x, \quad \varphi_{2k-1} = \sin k\omega x, \quad k = 1, 2, \cdots, n, \quad \omega = \frac{2\pi}{T}.$$

容易验证三角多项式函数系 $\{\varphi_i\}_{i=0}^{2n}$ 构成 M 的一组正交基. 利用定理9.8和9.9知, 对于任意的函数 $f(x) \in L^2[0,T)$, 存在唯一的最佳平方逼近元 $f_n(x) \in M$, 且有

$$f_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n} \big(a_k \cos k\omega x + b_k \sin k\omega x\big),$$

其中

$$\begin{aligned} a_k &= \frac{(f,\varphi_{2k})}{(\varphi_{2k},\varphi_{2k})} = \frac{2}{T}\int_0^T f(x)\cos k\omega x \,\mathrm{d}x, \quad k=0,1,\cdots,n,\\ b_k &= \frac{(f,\varphi_{2k-1})}{(\varphi_{2k-1},\varphi_{2k-1})} = \frac{2}{T}\int_0^T f(x)\sin k\omega x \,\mathrm{d}x, \quad k=0,1,\cdots,n. \end{aligned} \tag{9.24}$$

由数学分析中的 Fourier 级数理论知, 最佳平方逼近三角多项式 $f_n(x)$ 恰好是 $f(x)$ 的 Fourier 级数的部分和, 而 a_k, b_k 为 Fourier 系数. 此外, 利用 Fourier 级数的收敛性定理可得: $\lim\limits_{n\to\infty}\|f-f_n\|=0$, 即 $f_n(x)$ 平方收敛到 $f(x)$.

利用 Euler 公式 $e^{\mathrm{i}\theta}=\cos\theta+\mathrm{i}\sin\theta$, Fourier 级数可以写成复数形式

$$f(x)=\sum_{k=-\infty}^{+\infty} c_k e^{\mathrm{i}k\omega x},$$

其中

$$c_k=\frac{1}{T}\int_0^T f(x)e^{-\mathrm{i}k\omega x}\,\mathrm{d}x.$$

例 9.11 设 $f(x)=|x|$, 求 $f(x)$ 在区间 $[-\pi,\pi]$ 上的 n 次最佳平方逼近三角多项式.

解 设 $f(x)$ 在区间 $[-\pi,\pi]$ 上的 n 次最佳平方逼近三角多项式为

$$f_n(x)=\frac{a_0}{2}+\sum_{k=1}^{n}\big(a_k\cos k\omega x+b_k\sin k\omega x\big), \quad \omega=\frac{2\pi}{T}=1.$$

按式 (9.24), 展开系数

$$\begin{aligned} a_0 &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\,\mathrm{d}x = \frac{2}{\pi}\int_0^{\pi} x\,\mathrm{d}x = \pi,\\ a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos kx\,\mathrm{d}x = \frac{2}{\pi}\int_0^{\pi} x\cos kx\,\mathrm{d}x = \frac{2}{\pi k^2}[(-1)^k-1],\\ b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin kx\,\mathrm{d}x = \frac{2}{\pi}\int_0^{\pi} (x-x)\sin kx\,\mathrm{d}x = 0, \end{aligned}$$

其中 $k=1,2,\cdots,n$. 因此, 有

$$f_n(x)=\frac{\pi}{2}+\frac{2}{\pi}\sum_{k=1}^{n}\frac{[(-1)^k-1]}{k^2}\cos kx.$$

当 $n=0,1,3$ 时, 函数图像如图9.3所示. □

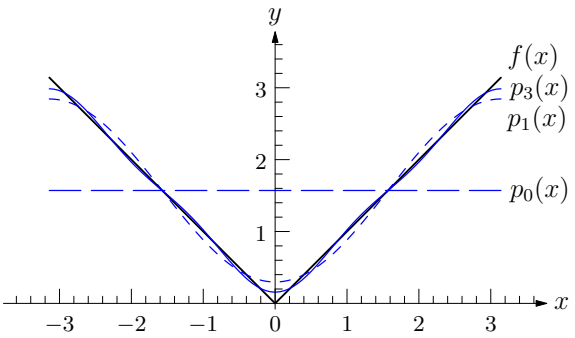


图 9.3: 函数 $f(x)=|x|$ 在区间 $[-\pi,\pi]$ 上的最佳平方逼近三角多项式

下面, 我们讨论离散情形的傅立叶级数. 在许多实际问题中, 譬如信号处理, 函数 $f(x)$ 是未知的, 但是 $f(x)$ 在某些时刻的值可以通过测量的方式获得. 数学上, 设 $f(x)\in L^2[0,T)$, 已知 $f(x)$ 在一系列等距离散点上的值, 即

$$f_k=f(x_k),\quad x_k=\frac{kT}{n},\quad k=0,1,2,\cdots,n-1.$$

明显地, $(f_0,f_1,\cdots,f_{n-1})^{\mathrm{T}}\in\mathbb{C}^n$. 若在 $L^2[0,T)$ 上定义离散的内积和诱导的拟范数

$$(f,g)=\sum_{k=0}^{n-1}f_k\cdot\overline{g}_k,\quad \forall f,g\in L^2[0,T),$$
$$\|f\|=\sqrt{(f,f)},\quad \forall f\in L^2[0,T),$$

利用公式 (9.5) 及 $\{1,e^{\mathrm{i}\omega x},e^{\mathrm{i}2\omega x},\cdots,e^{\mathrm{i}(m-1)\omega x}\}$ 的正交性, 即

$$(e^{\mathrm{i}j\omega x},e^{\mathrm{i}l\omega x})=\sum_{k=0}^{n-1}e^{\mathrm{i}(j-l)\omega x_k}=\begin{cases}0, & j\neq l,\\ n, & j=l,\end{cases}$$

则 $f(x)$ 的离散数据 $\{(x_k,f_k)\}_{k=0}^{n-1}$ 在空间 $M=\mathrm{span}\{1,e^{\mathrm{i}\omega x},e^{\mathrm{i}2\omega x},\cdots,e^{\mathrm{i}(m-1)\omega x}\}$ 的最佳平方逼近元为

$$s_m(x)=\sum_{l=0}^{m-1}g_le^{\mathrm{i}l\omega x},\quad g_l=\frac{(f,e^{\mathrm{i}l\omega x})}{n}=\frac{1}{n}\sum_{k=0}^{n-1}f_ke^{-\mathrm{i}l\omega x_k}.$$

特别地, 当 $m=n$ 时, 插值条件 $s_n(x_k)=f(x_k),k=0,1,\cdots,n-1$ 成立, 称 $s_n(x)$ 为 $f(x)$ 的$n-1$ **次插值三角多项式**. 利用插值三角多项式的系数与样本值之间的关系, 来定义离散傅立叶变换及其逆变换.

定义 9.10 设有向量 $\mathbf{f}=(f_0,f_1,\cdots,f_{n-1})^{\mathrm{T}}\in\mathbb{C}^n$, 称向量

$$\mathbf{g}=(g_0,g_1,\cdots,g_{n-1})^{\mathrm{T}}\in\mathbb{C}^n$$

为向量 $\mathbf{f}$ 的**离散傅立叶变换** (Discrete Fourier Transform, DFT), 其中

$$g_l=\frac{1}{n}\sum_{k=0}^{n-1}f_k e^{-\mathrm{i}2\pi lk/n},\quad l=0,1,\cdots,n-1.$$

反之, 称向量 $\mathbf{f}$ 为向量 $\mathbf{g}$ 的**离散傅立叶逆变换**, 即

$$f_j=\sum_{k=0}^{n-1}g_l e^{\mathrm{i}2\pi jk/n},\quad j=0,1,\cdots,n-1.$$

利用线性代数的语言, 离散傅立叶变换可写成

$$\begin{pmatrix}g_0\\g_1\\g_2\\\vdots\\g_{n-1}\end{pmatrix}=\frac{1}{n}\begin{pmatrix}\rho^0&\rho^0&\rho^0&\cdots&\rho^0\\\rho^0&\rho^1&\rho^2&\cdots&\rho^{n-1}\\\rho^0&\rho^2&\rho^4&\cdots&\rho^{2(n-1)}\\\vdots&\vdots&\vdots&\ddots&\vdots\\\rho^0&\rho^{n-1}&\rho^{2(n-1)}&\cdots&\rho^{(n-1)^2}\end{pmatrix}\begin{pmatrix}f_0\\f_1\\f_2\\\vdots\\f_{n-1}\end{pmatrix}\Longleftrightarrow\mathbf{g}=F_n\mathbf{f},\tag{9.25}$$

其中 $\rho=e^{-\mathrm{i}2\pi/n}$, 称 F_n 为**傅立叶变换矩阵**. 容易看出, F_n 是对称矩阵, 且除了第一行 (列) 外, 矩阵的每一行 (列) 元素之和为零. 此外, F_n 满足 $F_n\overline{F_n^{\mathrm{T}}}=I/n$, 即 $F_n/\sqrt{n}$ 是酉阵. 因此, 傅立叶变换矩阵 F_n 的逆为

$$F_n^{-1}=n\overline{F_n}=\begin{pmatrix}\rho^0&\rho^0&\rho^0&\cdots&\rho^0\\\rho^0&\rho^{-1}&\rho^{-2}&\cdots&\rho^{-(n-1)}\\\rho^0&\rho^{-2}&\rho^{-4}&\cdots&\rho^{-2(n-1)}\\\vdots&\vdots&\vdots&\ddots&\vdots\\\rho^0&\rho^{-(n-1)}&\rho^{-2(n-1)}&\cdots&\rho^{-(n-1)^2}\end{pmatrix},$$

即离散傅立叶逆变换的矩阵表示.

在许多工程领域, 利用计算机进行 Fourier 分析的主要方法是离散傅立叶变换. 从式 (9.25) 不难看出, 计算离散傅立叶变换需要 n^2 次乘法, $n(n-1)$ 次加法及 n 次除法, 故时间复杂度为 $O(n^2)$. 类似地, 离散傅立叶逆变换的时间复杂度亦为 $O(n^2)$. 然而, 当 N 很大时, 运算量会变得相当大, 即使用高速计算机, 也需要花费大量的时间. 正因为

如此, 在相当长的时间内, 用数值方法进行 Fourier 分析没有得到广泛的应用. 直到上世纪六十年代, 出现了快速傅立叶变换 (Fast Fourier Transform, FFT) 算法, 使得算法时间复杂度降为 $O(n\log n)$, 该问题才得以彻底的解决. 此后, 快速傅立叶变换在数字信号处理, 光谱分析和科学工程计算等众多领域得到广泛的应用, 被评为二十世纪的十大算法之一.

快速傅立叶变换采用分而治之 (Divide and conquer) 策略, 下面介绍一种常用的逐次分半算法. 若 $n=2m$ 为偶数, 记 $\omega_n=e^{-\mathrm{i}2\pi/n}$, 多项式 $p(z)=\dfrac{1}{n}\displaystyle\sum_{k=0}^{n-1}f_kz^k$, 则

$$g_l=p(\omega_n^l),\quad l=0,1,\cdots,n-1,$$

即计算向量 $\mathbf{f}$ 的离散傅立叶变换等价于求多项式 $p(z)$ 在 n 个点 $\{1,\omega_n,\omega_n^2,\cdots,\omega_n^{n-1}\}$ 处的值. 将 $p(x)$ 的系数按偶次项和奇数项分开, 构造多项式

$$p_0(z)=\frac{1}{m}\sum_{k=0}^{m-1}f_{2k}z^k,\quad p_1(z)=\frac{1}{m}\sum_{k=0}^{m-1}f_{2k+1}z^k,$$

则

$$p(z)=\frac{p_0(z^2)+z\,p_1(z^2)}{2}.$$

从而将问题转化为求多项式 $p_0(z)$ 和 $p_1(z)$ 在 $\{1,\omega_n^2,\omega_n^4,\cdots,\omega_n^{2(n-1)}\}$ 的值. 注意到, 利用单位根的性质有

$$\omega_n^{2k}=e^{-\mathrm{i}2\pi(2k)/n}=e^{-\mathrm{i}2\pi k/m}=\omega_m^k,\quad k=0,1,\cdots,n-1,$$

故前面的集合仅有 m 个不同的值, 即 $\{1,\omega_m,\omega_m^2,\cdots,\omega_m^{m-1}\}$. 此外, 对于 $k=0,1,\cdots,m-1$, 有

$$\begin{aligned}g_k&=p(\omega_n^k)=\frac{p_0(\omega_n^{2k})+\omega_n^kp_1(\omega_n^{2k})}{2}=\frac{p_0(\omega_m^k)+\omega_n^kp_1(\omega_m^k)}{2},\\ g_{k+m}&=p(\omega_n^{k+m})=\frac{p_0(\omega_n^{2k+n})+\omega_n^{k+m}p_1(\omega_n^{2k+n})}{2}=\frac{p_0(\omega_m^k)-\omega_n^kp_1(\omega_m^k)}{2}.\end{aligned}\tag{9.26}$$

因此, 多项式 $p(z)$ 的求值问题可划分成两个子问题解决, 即 $p_0(z)$ 和 $p_1(z)$ 的求值问题. 进一步, 假设 n 是 2 的幂次方, 则反复的应用上述策略, 如算法9.1所示, 这就是一种快速傅立叶算法.

事实上, 假设 $n=2^h$, 记 $C[n]$ 是计算具有 n 个分量的向量 $\mathbf{f}$ 的离散傅立叶变换所需的运算次数, 由式 (9.26) 可知:

$$C[n]\leqslant n+2C[n/2]+3\times n=2C[n/2]+4n,\tag{9.27}$$

Algorithm 9.1 Fast Fourier Transform Algorithm

function FFT(**f**)
  $n \leftarrow length[\mathbf{f}]$; ▷ n is a power of 2
  if $n = 1$ **then return f**;
  $\omega_n \leftarrow e^{-\mathrm{i}2\pi/n}$;
  $\omega \leftarrow 1$;
  $\mathbf{f}^0 \leftarrow (f_0, f_2, \cdots, f_{n-2})$;
  $\mathbf{f}^1 \leftarrow (f_1, f_3, \cdots, f_{n-1})$;
  $\mathbf{g}^0 \leftarrow \mathrm{FFT}(\mathbf{f}^0)$; ▷ Apply FFT to even coefficients
  $\mathbf{g}^1 \leftarrow \mathrm{FFT}(\mathbf{f}^1)$; ▷ Apply FFT to odd coefficients
  for $k \leftarrow 0$ to $n/2 - 1$ **do**
    $g_k \leftarrow (\mathbf{g}_k^0 + \omega \mathbf{g}_k^1)/2$; ▷ Synthesize coefficients using Eq. (9.26)
    $g_{k+n/2} \leftarrow (\mathbf{g}_k^0 - \omega \mathbf{g}_k^1)/2$;
    $\omega \leftarrow \omega\omega_n$;
  end for
  return g;
end function

其中 n 是用于计算 $\{1, \omega_n, \omega_n^2, \cdots, \omega_n^{n-1}\}$ 的运算次数, $3n$ 是利用 $p_0(z)$ 和 $p_1(z)$ 计算 $p(z)$ 的运算次数. 接下来, 用数学归纳法证明 $C[n] \leqslant 4 \cdot 2^h h = 4n \log_2(n)$.

当 $h = 1$ 时, 显然成立.

当 $h = 2$ 时, 容易看出

$$g_0 = \frac{f_0 + f_1}{2}, \quad g_1 = \frac{f_0 + (-1)f_1}{2},$$

故 $C[2] = 5 \leqslant 8$, 命题成立.

假设命题对 $n = 2^{h-1}$ 均成立, 则有 $C[2^{h-1}] \leqslant 4 \cdot 2^{h-1}(h-1)$. 现在考虑 $n = 2^h$ 情形, 利用式 (9.27) 有

$$C[n] = C[2^h] \leqslant 2 \times 4 \times 2^{h-1}(h-1) + 4 \times 2^h = 4 \cdot 2^h h = 4n \log_2(n).$$

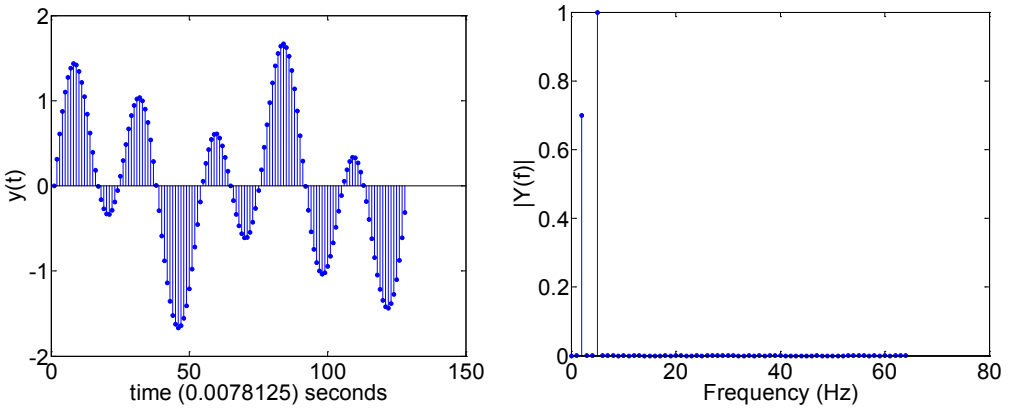
命题证毕.

从上面的讨论知, 快速傅立叶变换算法的时间复杂度为 $O(n \log n)$. 不难看出, 离散傅立叶逆变换亦有类似的快速算法. 而对于一般情形的 n, 也有类似的快速傅立叶变换算法, 限于篇幅, 就不作详细介绍了, 可参阅文献 [6].

例 9.12 设函数 $f(t) = 0.7\sin(2\pi \times 2t) + \sin(2\pi \times 5t)$, 对 $f(t)$ 在时间域 $[0, 1)$ 上进行采样, 间隔为 $\Delta t = 1/128 = 0.0078125$ 秒, 得到一个向量 $\mathbf{f} = (f_0, f_1, \cdots, f_{127})^{\mathrm{T}}$, 如图9.4(a)所示. 对 $\mathbf{f}$ 施行快速傅立叶变换得向量 $\mathbf{g}$, 取该向量的一半, 并计算每个向量的模长并乘以 2, 如图9.4(b)所示. 注意 $\mathbf{g}$ 是复数域上的向量, 且当输入向量 $\mathbf{f}$ 是实向量时, 满足共轭对称性, 即 $g_{128-k} = \overline{g_k}$. 因此, 图9.4(b)显示了频率域上能量的分布, 从该图中可以明显看出该信号在 2Hz 和 5Hz 的频率下, 振幅分别为 0.7 和 1, 与输入的函数信息完全一致.

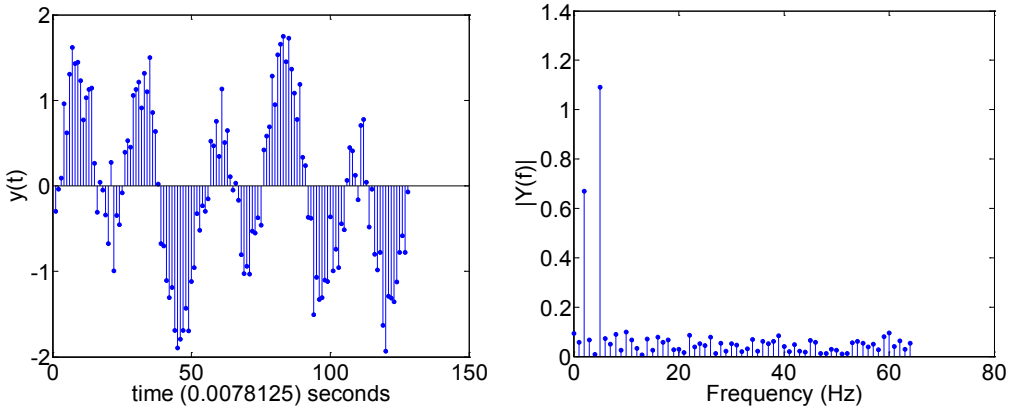
在实际问题中, 采样过程中往往会因各种原因引入误差, 导致数据含有噪音, 如图9.5(a)所示. 对于该信号, 重复上述的方法得图9.5(b), 从该图中仍可以看出: 在 2Hz 和 5Hz 的频率下, 信号的振幅较大, 而其他频率的信号振幅较小. 通过这个例子, 可以看出傅立叶分析对带有噪音的信号也是适用的.

例 9.13 现有一段通过麦克风采样得到的音频数据, 其前 256 个分量如图9.6(a)所示. 对该信号施行快速傅立叶变换, 其频率域上能量的分布见图9.6(b). 若取频率域的前 25% 的系数, 即低频部分的系数, 其他系数置为 0, 进行离散傅立叶逆变换, 得到时间域上的重建信号, 如图所示9.6(c). 明显地, 重建的信号比原来的信号光滑. 因此, 利用快速傅立叶变换, 就可以实现信号的去噪或光滑的目的. 此外, 若取频率域的前 12.5% 的



(a) 时间域上的输入信号 (b) 频率域上的能量分布

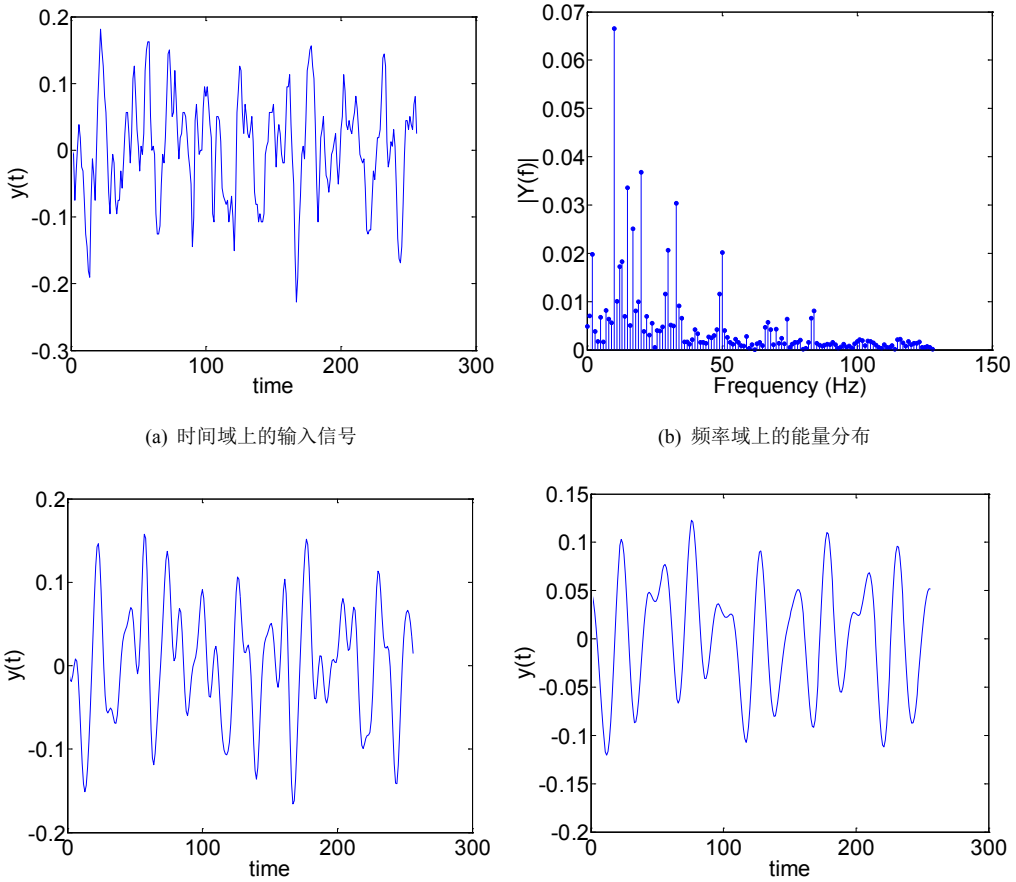
图 9.4: 信号的时频域变换



(a) 时间域上的输入信号 (b) 频率域上的能量分布

图 9.5: 带噪音信号的时频域变换

系数, 其他系数置为 0, 重复上述过程, 得到时间域上的重建信号, 见图9.6(d). 注意到, 在重建信号的过程中, 我们只用到了 12.5% 的系数, 高频系数被丢弃. 虽然引入了一些误差, 但是有效的降低了数据的存储或传输量, 在本例中实现了 1 : 8 的压缩比. 因此, 在实际应用中, 譬如文件存储、网络传输等, 人们利用快速傅立叶变换, 只存储或传输低频系数, 而丢弃高频系数, 便可实现信号的有损压缩存储或传输.



(a) 时间域上的输入信号

(b) 频率域上的能量分布

(c) 时间域上的重建信号 (使用前 25% 的系数)

(d) 时间域上的重建信号 (使用前 12.5% 的系数)

图 9.6: 信号的去噪与压缩

由于离散信号的傅立叶变换可以通过 FFT 算法计算, 人们可以快速得将时间域的信号变换到频率域, 进而开展数据的各种分析与处理, 这是 FFT 算法被广泛使用的原

因之一.

9.6 最佳一致逼近多项式

本节主要讨论函数的最佳一致逼近问题, 即连续情形的最佳一致逼近问题. 在函数空间 $C[a,b]$ 上, 引入一致范数

$$\|f\|_\infty = \max_{x\in[a,b]} \big|f(x)\big|, \quad \forall f \in C[a,b],$$

逼近集合 M 取 $n+1$ 维多项式空间 $\mathbb{P}_n[x]$. 对于任意的 $f \in C[a,b]$, 如果存在 $p^* \in \mathbb{P}_n[x]$ 使得

$$\|f-p^*\|_\infty = \inf_{p\in\mathbb{P}_n[x]} \|f-p\|_\infty \triangleq d(f,\mathbb{P}_n[x]),$$

则称 p^* 为函数 f 的**最佳一致逼近多项式**. 由推论 (9.3) 知, 最佳一致多项式 p^* 是存在的.

下面讨论逼近的其他问题. 首先研究最佳一致逼近多项式具有的特征, 即 Chebyshev 特征定理. 假设 $p^* \in \mathbb{P}_n[x]$ 是函数 f 的任意一个逼近, 考虑误差函数

$$e^*(x) = f(x) - p^*(x), \quad x \in [a,b].$$

定义 $e^*(x)$ 的极值点集, 即

$$\mathcal{L}_M = \big\{x \in [a,b] : |e^*(x)| = \|e^*\|_\infty\big\}.$$

若多项式 $(p^* + \lambda p)$ 是函数 f 的最佳一致逼近多项式, 其中 $\lambda \in \mathbb{R}$, $p \in \mathbb{P}_n[x]$, 则

$$|e^*(x) - \lambda p(x)| < |e^*(x)|, \quad \forall x \in \mathcal{L}_M.$$

不失一般性, 假设 λ 是正实数, 从上式不难看出, 对于任意的 $x \in \mathcal{L}_M$, 函数值 $e^*(x)$ 的符号与 $p(x)$ 的符号必然相同. 因此, p^* 是 f 的最佳一致逼近多项式的充分条件是: 不存在多项式 $p \in \mathbb{P}_n[x]$ 使得

$$[f(x) - p^*(x)]p(x) > 0, \qquad \forall x \in \mathcal{L}_M. \tag{9.28}$$

相反地, 可以证明条件 (9.28) 亦为 p^* 是 f 的最佳一致逼近多项式的必要条件. 为便于后续的讨论, 我们考虑更为一般的最佳一致逼近问题, 即

$$\arg\min_{p\in\mathbb{P}_n[x]} \Big\{ \max_{x\in\mathcal{L}} \big|f(x) - p(x)\big| \Big\}, \tag{9.29}$$

其中 $\mathcal{L}$ 是区间 $[a,b]$ 的任意闭子集.

定理 9.16 对任意的 $f \in C[a,b]$, $p^* \in \mathbb{P}_n[x]$, 记 $e^*(x) = f(x) - p^*(x)$, $\mathcal{L}_M = \{x \in \mathcal{L} : |e^*(x)| = \|e^*\|_\infty\}$, 则 p^* 是式 (9.29) 的解的充分必要条件是不存在多项式 $p \in \mathbb{P}_n[x]$ 使得条件 (9.28) 成立.

证明 (充分性) 当 $\mathcal{L} = [a,b]$ 时, 充分性已证. 另外, 不难验证, 当 $\mathcal{L}$ 是区间 $[a,b]$ 的任意闭子集, 命题仍然成立.

(必要性) 等价于证明: 若存在多项式 $p \in \mathbb{P}_n[x]$ 使得条件 (9.28) 成立, 即有 $\lambda \in \mathbb{R}$, 使得

$$\max_{x\in\mathcal{L}} |e^*(x) - \lambda p(x)| < \max_{x\in\mathcal{L}} |e^*(x)| \tag{9.30}$$

成立, 则 p^* 不是式 (9.29) 的解.

不失一般性, 假定 λ 是正实数, 多项式 p 满足

$$|p(x)| \leqslant 1, \quad \forall x \in [a,b].$$

记集合

$$\mathcal{L}_0 = \{x \in \mathcal{L} : e^*(x)p(x) \leqslant 0\},$$

因 $p(x)$ 和 $e^*(x)$ 均为连续函数, 故 $\mathcal{L}_0$ 是闭集, 从而 $|e^*(x)|$ 在 $\mathcal{L}_0$ 上能到取最大值

$$d = \max_{x\in\mathcal{L}_0} |e^*(x)|.$$

又因 $\mathcal{L}_0 \bigcap \mathcal{L}_M = \emptyset$, 故

$$d < \max_{x\in\mathcal{L}} |e^*(x)|.$$

若 $\mathcal{L}_0 = \emptyset$, 则令 $d = 0$. 我们取

$$\lambda = \frac{1}{2}\big[\max_{x\in\mathcal{L}} |e^*(x)| - d\big],$$

下面证明 λ 满足式 (9.30).

因 $\mathcal{L}$ 是闭集, 故存在 $\xi \in \mathcal{L}$ 使得

$$|e^*(\xi) - \lambda p(\xi)| = \max_{x\in\mathcal{L}} |e^*(x) - \lambda p(x)|$$

成立. 分情形讨论:

(1) 若 $\xi \in \mathcal{L}_0$, 则有

$$\max_{x\in\mathcal{L}} |e^*(x) - \lambda p(x)| = |e^*(\xi)| + |\lambda p(\xi)| \leqslant d + \lambda = \frac{1}{2}\big[\max_{x\in\mathcal{L}} |e^*(x)| + d\big] < \max_{x\in\mathcal{L}} |e^*(x)|.$$

(2) 若 $\xi \notin \mathcal{L}_0$, 则有

$$|e^*(\xi) - \lambda p(\xi)| < \max\left\{|e^*(\xi)|, |\lambda p(\xi)|\right\} \leqslant \max\left\{|e^*(\xi)|, \lambda\right\} \leqslant \max_{x\in\mathcal{L}} |e^*(x)|.$$

因此, 命题成立. □

通过例子, 不难发现, 最佳逼近元的误差应在整个逼近区间上均匀分布, 即误差函数的最大值与最小值大小相等, 符号相反, 且交错分布.

定义 9.11 设 $g \in C[a,b]$, 称满足 $a \leqslant x_0 < x_1 < \cdots < x_k \leqslant b$ 的点集 $\{x_i\}_{i=0}^k$ 为 $g(x)$ 在 $[a,b]$ 上的**交错点组**, 如果它满足

$$g(x_i) = (-1)^i \sigma \|g\|_\infty, \quad i = 0, 1, \cdots, k,$$

其中 $\sigma = 1$ 或 $\sigma = -1$, 并称 x_i 为交错点.

定理 9.17 (Chebyshev 交错定理) 设函数 $f \in C[a,b]$ 且 $f \notin \mathbb{P}_n[x]$, 则 p^* 是 f 的 n 次最佳一致逼近多项式的充分必要条件是 $f - p^*$ 在 $[a,b]$ 上存在有 $n+2$ 个点组成的交错点组, 即有 $a \leqslant x_0 < x_1 < \cdots < x_{n+1} \leqslant b$ 使得

$$f(x_i) - p^*(x_i) = (-1)^i \sigma \|f - p^*\|_\infty, \quad i = 0, 1, \cdots, n+1.$$

证明 (充分性) 设 $f - p^*$ 在 $[a,b]$ 上存在有 $n+2$ 个点组成的交错点组 $\{x_i\}_{i=0}^k \subset \mathcal{L}_M$. 用反证法. 如果存在多项式 $p \in \mathbb{P}_n[x]$ 使得条件 (9.28) 成立, 那么 p 显然至少有 $n+1$ 个零点. 另一方面, 利用代数基本定理知, 次数不超过 n 次的非零多项式 p 至多有 n 零点, 矛盾! 因此, 由定理 (9.16) 知, p^* 是 f 的最佳一致逼近多项式.

(必要性) 设 p^* 是 f 的最佳一致逼近多项式, 记 $E_n(f) = \|f - p^*\|_\infty$, 因 $f \notin \mathbb{P}_n[x]$, 故 $E_n(f) > 0$. 用反证法. 设 $f - p^*$ 在 $[a,b]$ 上的任意交错点组为 $\{x_i\}_{i=0}^k$, 均有 $k \leqslant n$. 因此, 存在区间 $[a,b]$ 的一个分割 $\{t_i\}_{i=0}^{k+1}$, 其中 $t_0 = a, t_{k+1} = b$, 使得

$$S_i = \left\{x \in [t_{i-1}, t_i] : \left|(f - p^*)(x)\right| = E_n(f)\right\}, \quad i = 1, 2, \cdots, k+1,$$

非空, 且函数 $f - p^*$ 在 S_i 上保持符号不变. 由于 $f - p^*$ 在相邻 S_i 和 S_{i+1} 上交替取 $\pm E_n(f)$, 故 $[t_{i-1}, t_{i+1}]$ 上至少存在一个 $f - p^*$ 的零点, 不失一般性, 可以调整 t_i 使得

$$(f - p^*)(t_i) = 0, \quad i = 1, 2, \cdots, k.$$

由于 $k \leqslant n$, 可构造如下 k 次多项式

$$p(x) = \sigma \prod_{i=1}^{k} (x - t_i) \in \mathbb{P}_n[x],$$

使得 p 满足条件 (9.28). 而这与定理 (9.16) 矛盾! 因此, 假设错误, 原命题成立. □

例 9.14 试在线性函数空间中求 e^x 在区间 $[-1, 1]$ 上的最佳一致逼近多项式.

解 设 e^x 在 $[-1, 1]$ 上的一次最佳一致逼近多项式为 $p_1^*(x) = c_0 + c_1 x$. 利用 Chebyshev 交错定理及函数图像知: 误差函数 $\varepsilon(x) = e^x - p_1^*(x)$ 的极值在下面三个点上正负相间地取到, 即

$$\varepsilon(-1) = \rho, \quad \varepsilon(x_3) = -\rho, \quad \varepsilon(1) = \rho,$$

其中 $\rho = \max\limits_{x \in [-1,1]} |\varepsilon(x)|, x_1 < x_3 < x_2$, x_1 和 x_2 是 $p_1^*(x)$ 与 e^x 在 $[-1, 1]$ 上的交点.

因 x_3 是 $\varepsilon(x)$ 的极值点, 故满足

$$\varepsilon'(x_3) = 0.$$

结合前面的条件, 代入并求解得

$$\begin{cases} e^{-1} - (c_0 - c_1) = \rho, \\ e^{x_3} - (c_0 + c_1 x_3) = -\rho, \\ e - (c_0 + c_1) = \rho, \\ e^{x_3} - c_1 = 0. \end{cases} \Longrightarrow \begin{cases} c_1 = \dfrac{e - e^{-1}}{2} \approx 1.1752, \\ x_3 = \ln(c_1) \approx 0.1614, \\ \rho = \dfrac{e^{-1} + c_1 x_3}{2} \approx 0.2788, \\ c_0 = \rho + (1 - x_3) c_1 \approx 1.2643. \end{cases}$$

综上, 所求的一次最佳逼近多项式 $p_1^*(x) \approx 1.2643 + 1.1752x$, 函数图像见图9.7(a), 误差分布见图9.7(b).

□

需要指出, 交错点组往往是不唯一的, 且计算交错点组是一个困难的问题. 但是, 在一些特殊的情况下, 交错点组中的个别点可以被很快确定.

推论 9.18 设 p^* 是 f 的 n 次最佳一致逼近多项式, 如果 f 在区间 $[a, b]$ 上有 $n+1$ 阶导数, 且 f^{n+1} 在 (a, b) 上不变号 (恒正或恒负), 则 $f - p^*$ 的交错点组有且仅有 $n+2$ 个交错点, 且区间 $[a, b]$ 的端点属于 $f - p^*$ 的交错点组.

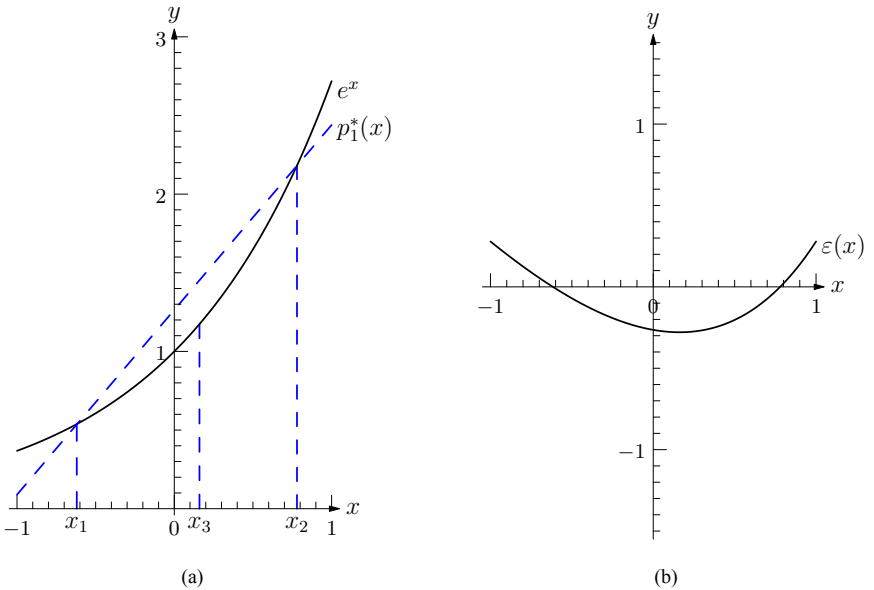


图 9.7: e^x 在 $[-1,1]$ 的最佳一致逼近

证明 用反证法. 设 $f-p^*$ 的交错点组的个数超过 $n+2$ 个, 或者 a 和 b 中有一个不属于 $f-p^*$ 的交错点组, 则至少有 $n+1$ 个交错点 $\{\xi_i\}_{i=0}^{n+1}$ 落在 (a,b) 内. 因这些点是 $f-p^*$ 的极值点, 故有

$$f'(\xi_i)-p^{*'}(\xi_i)=0, \quad i=0,1,\cdots,n.$$

反复利用 Rolle 定理, 知存在 $\eta\in(a,b)$ 使得

$$f^{(n+1)}(\eta)-p^{*(n+1)}(\eta)=f^{(n+1)}(\eta)=0,$$

而这与 $f^{(n+1)}$ 在 (a,b) 上不变号矛盾! 原命题成立. □

例 9.15 设 $f(x)=\sqrt{x}$, 求 $f(x)$ 在区间 $[1/4,1]$ 上的一次最佳一致逼近多项式.

解 设 $\sqrt{x}$ 在 $[1/4,1]$ 上的一次最佳一致逼近多项式为 $p_1^*(x)=c_0+c_1x$, 并记 $\varepsilon(x)=f(x)-p_1^*(x)$. 因 $f''(x)=-x^{-3/2}/4$ 在 $(1/4,1)$ 上不变号, 利用推论9.18 知, $x_0=1/4, x_2=1$ 均为 $\varepsilon(x)$ 的交错点. 而另一个交错点 $x_1\in(1/4,1)$ 应满足

$$\varepsilon'(x_1)=\frac{1}{2\sqrt{x_1}}-c_1=0.$$

利用交错点的定义, 有

$$f\Big(\frac{1}{4}\Big)-\Big(c_0+\frac{c_1}{4}\Big)=f(1)-(c_0+c_1)\Longrightarrow c_1=\frac{2}{3}.$$

代入 x_1 需满足的条件, 得 $x_1=9/16$. 最后, 利用交错点的定义

$$f(1)-(c_0+c_1)=-[f(x_1)-(c_0+c_1x_1)],$$

可求得 $c_0=17/48$. 综上, 所求的一次最佳一致逼近多项式

$$p_1^*(x)=\frac{17}{48}+\frac{2}{3}x.$$

□

自然地, 我们会问: 最佳一致逼近多项式是否是唯一的? 回答是肯定的.

定理 9.19 (唯一性定理) 设函数 $f\in C[a,b]$, 则 f 用空间 $\mathbb{P}_n[x]$ 的元素所做的最佳一致逼近是唯一的.

证明 用反证法. 设 $p_1,p_2\in\mathbb{P}_n[x]$ 均为 f 的 n 次最佳一致逼近多项式, 令

$$p^*=\frac{p_1+p_2}{2}\in\mathbb{P}_n[x],$$

则

$$\|f-p^*\|\leqslant\frac{1}{2}\big(\|f-p_1\|+\|f-p_2\|\big)=d(f,\mathbb{P}_n[x]).$$

因此, p^* 亦为 f 的 n 次最佳一致逼近多项式. 利用定理 (9.17) 知, 存在交错点组 $a\leqslant x_0<x_1<\cdots<x_{n+1}\leqslant b$, 使得

$$d(f,\mathbb{P}_n[x])=|f(x_i)-p^*(x_i)|=\Big|\frac{f(x_i)-p_1(x_i)}{2}+\frac{f(x_i)-p_2(x_i)}{2}\Big|.$$

但 $|f(x_i)-p_1(x_i)|$ 和 $|f(x_i)-p_2(x_i)|$ 均不大于 $d(f,\mathbb{P}_n[x])$, 于是有

$$f(x_i)-p_1(x_i)=f(x_i)-p_2(x_i)=\sigma d(f,\mathbb{P}_n[x]),\quad \sigma=1\text{ or }-1.$$

从而 $p_1(x_i)=p_2(x_i)$, 其中 $i=0,1,\cdots,n+1$. 又因 p_1 和 p_2 均为次数不超过 n 的多项式, 故 $p_1\equiv p_2$, 命题成立. □

注解 9.2 因空间 $C[a,b]$ 在一致范数下是凸的, 但不是严格凸的, 故不能直接应用推论 (9.5).

其次, 我们研究 $d(f,\mathbb{P}_n[x])$ 的下界, 有如下定理.

定理 9.20 (de la Vallée-Poussin 定理) 设函数 $f\in C[a,b]$, 若存在多项式 $p\in\mathbb{P}_n[x]$, 使得 $f-p$ 在 $[a,b]$ 上至少 $n+2$ 个点 $x_0,x_1,\cdots,x_{n+1}$ 处的取值正负相间, 则

$$d(f,\mathbb{P}_n[x])\geqslant\delta=\min_{0\leqslant i\leqslant n+1}|f(x_i)-p(x_i)|.$$

证明 用反证法. 如果 $p^*\in\mathbb{P}_n[x]$ 为 f 的 n 次最佳一致逼近多项式, 且 $d(f,\mathbb{P}_n[x])<\delta$, 即 $\|f-p^*\|_\infty<\delta$. 记多项式

$$q(x)=p^*(x)-p(x)=f(x)-p(x)-(f(x)-p^*(x))\in\mathbb{P}_n[x],$$

则 q 在 x_i 处的符号完全由 $f(x_i)-p(x_i)$ 决定, 其中 $i=0,1,\cdots,n+1$. 利用介值定理知, q 在 $[a,b]$ 上至少有 $n+1$ 个零点. 因此, 必有 $q\equiv0\Rightarrow p^*\equiv p$, 与假设矛盾! 命题成立. □

最后, 我们讨论最佳一致逼近多项式的求解问题. 虽然 Chebyshev 交错定理从理论上给出了最佳一致逼近的特征性质, 但在一般情况下, 求解最佳一致逼近多项式是很困难的, 通常只能是近似计算, 一种常用的方法是 Remez 算法.

设函数 $f(x)$ 的 n 次最佳一致逼近多项式为 $p^*(x)$, 利用切比雪夫交错定理知, $f-p^*$ 在 $[a,b]$ 上存在 $n+2$ 个交错点 $\{x_i\}_{i=0}^{n+1}$, 使得

$$p^*(x_i)-f(x_i)=(-1)^i\mu,\quad i=0,1,\cdots,n+1,\tag{9.31}$$

其中 $p^*(x)=\sum\limits_{i=0}^{n}c_i^*x^i,\mu=\sigma d(f,\mathbb{P}_n[x])$. 不难看出, 如果交错点 $\{x_i\}_{i=0}^{n+1}$ 一旦确定, 那么通过线性方程组 (9.31) 可求出 p^* 的系数 $c_0^*,c_1^*,\cdots,c_n^*$ 和最佳逼近值 $d(f,\mathbb{P}_n[x])$. 然而, 寻找交错点 $\{x_i\}_{i=0}^{n+1}$ 并非一件容易的事, 为此, Remez 采用逐次逼近策略提出一种近似算法, 具体步骤如下:

(1) 设定精度 $\varepsilon>0$, 在 $[a,b]$ 上任选 $n+2$ 个初始点 $a\leqslant x_0^0<x_1^0<\cdots<x_{n+1}^0\leqslant b$ 作为初始交错点, 代入式 (9.31), 求得 $p^0(x)=\sum\limits_{i=0}^{n}c_i^0x^i$ 及 μ^0.

(2) 设第 l 步的交错点为 $\{x_i^l\}_{i=0}^{n+1}$, 逼近多项式 $p^l(x)=\sum\limits_{i=0}^{n}c_i^lx^i$, 逼近误差为 μ^l, 记

$$\eta^l\triangleq\max_{x\in[a,b]}|f(x)-p^l(x)|=|f(\hat{x})-p^l(\hat{x})|,$$

如果 $\eta^l - |\mu^l| < \varepsilon$, 则算法结束, $p^l(x)$ 作为最佳一致逼近多项式 p^* 的近似; 否则, 利用 $\hat{x}$ 替换交错点 $\{x_i^l\}_{i=0}^{n+1}$ 中的某一点, 得到新的交错点 $\{x_i^{l+1}\}_{i=0}^{n+1}$, 规则如下:

(a) 当 $\hat{x} \in (x_i^l, x_{i+1}^l)$ 时, 若 $f(\hat{x}) - p^l(\hat{x})$ 与 $f(x_i^l) - p^l(x_i^l)$ 同号, 则用 $\hat{x}$ 替换 x_i^l; 否则用 $\hat{x}$ 替换 x_{i+1}^l.

(b) 当 $\hat{x} < x_0^l$ 时, 若 $f(\hat{x}) - p^l(\hat{x})$ 与 $f(x_0^l) - p^l(x_0^l)$ 同号, 则用 $\hat{x}$ 替换 x_0^l; 否则新的交错点为 $\{\hat{x}, x_0^l, \cdots, x_n^l\}$.

(c) 当 $\hat{x} > x_{n+1}^l$ 时, 若 $f(\hat{x}) - p^l(\hat{x})$ 与 $f(x_{n+1}^l) - p^l(x_{n+1}^l)$ 同号, 则用 $\hat{x}$ 替换 x_{n+1}^l; 否则新的交错点为 $\{x_1^l, \cdots, x_{n+1}^l, \hat{x}\}$.

(3) 将新的交错点 $\{x_i^{l+1}\}_{i=0}^{n+1}$ 代入式 (9.31), 求得 $p^{l+1}(x) = \sum_{i=0}^{n} c_i^{l+1} x^i$ 及 μ^{l+1}. 令 $l \leftarrow l+1$, 返回步骤 (2).

注解 9.3 在步骤 (2) 中, 最大偏差点 $\hat{x}$ 一般只能近似求出; 另外, 算法的结束条件可以不仅限于条件 $\eta^l - |\mu^l| < \varepsilon$.

注解 9.4 在步骤 (2) 中, 每次只改变一个交错点的方法称为单一交换法; 若用最大偏差点集中的多个点按前述规则替换 $\{x_i^l\}_{i=0}^{n+1}$, 则称为同时交换法. 容易看出, 同时交换法比单一交换法具有更高的效率.

注解 9.5 在 Remez 算法中, 每次迭代都需要确定最大偏差点, 但由于没有很好的方法确定一般函数在有界闭区间上的最大值, 故 Remez 算法的实施存在一定的困难. 一个可行的解决方法是采用局部最优策略代替整体最优策略, 譬如, 若 $f(x)$ 具有连续的二阶导数时, 可利用如下 Newton 迭代

$$x^{k+1} = x^k - \frac{f'(x^k) - p_n'(x^k)}{f''(x^k) - p_n''(x^k)}$$

来计算最大偏差点, 迭代的初始值可取当前交错点组中的某个点; 若采用同时交换法, 则初始值可取当前交错点组中的多个点, 得到多个近似最大偏差点.

可以证明, Remez 算法是收敛的, 且对于许多函数, 收敛速度甚至是二次的. 另外, Remez 算法对于初值的选取也不太敏感. 尽管 Remez 算法有这些好的性质, 但是它的计算量仍是比较大的. 所以, 在实际计算中, 常使用一些其他近似的方法, 其中有一类重要的方法是使用所谓的切比雪夫多项式, 将在下一节作深入的讨论.

9.7 切比雪夫多项式

在9.4节中, 我们已经给出了切比雪夫多项式 $T_k(x)$ 的定义及递推公式, 指出它们是区间 $[-1,1]$ 上的多项式空间 $\mathbb{P}_n(x)$ 关于权函数 $\rho(x)=(1-x^2)^{-1/2}$ 的一组正交基, 即

$$T_k(x)=\cos(k\arccos x),\quad x\in[-1,1],\qquad \int_{-1}^{1}\frac{T_m(x)T_n(x)}{\sqrt{1-x^2}}\,\mathrm{d}x=\begin{cases}0, & m\neq n,\\ \pi/2, & m=n\neq 0,\\ \pi, & m=n=0.\end{cases}$$

容易看出, $T_k(x)$ 在 $[-1,1]$ 恰有 k 个不同的实根

$$x_i=\cos\frac{(2i-1)\pi}{2k},\quad i=1,2,\cdots,k.$$

规定 $T_0(x)=1$, 利用递推关系可算出 $T_k(x)$ 的具体表示, 见表9.2, 相应的形状如图9.8所示.

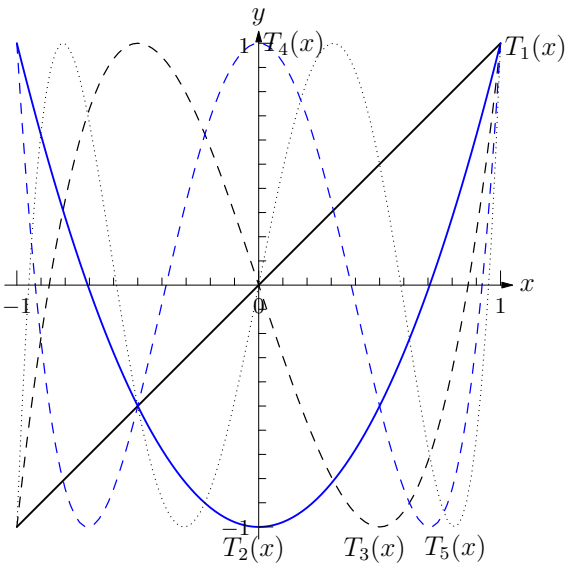


图 9.8: 切比雪夫多项式

本节将讨论如何用切比雪夫多项式来解决一些逼近问题. 首先, 我们考虑区间

$[-1,1]$ 上空间 $\mathbb{P}_{n-1}[x]$ 对 $f(x)=x^n$ 的最佳一致逼近问题, 即求 $p_{n-1}^*\in\mathbb{P}_{n-1}[x]$, 使得

$$\|f-p_{n-1}^*\|=\min_{p_{n-1}\in\mathbb{P}_{n-1}[x]}\|x^n-p_{n-1}\|,$$

其中 $\|f\|=\max\limits_{x\in[-1,1]}|f(x)|$. 若记 $\mathbb{P}_n^1[x]$ 为所有首项系数为 1 的 n 次多项式的全体, 则前述问题等价于求 $p_n^*\in\mathbb{P}_n^1[x]$, 使得

$$\|p_n^*-0\|=\min_{p_n\in\mathbb{P}_n^1[x]}\|p_n-0\|.$$

因此, 该问题亦称为**最小零偏差问题**. 利用切比雪夫交错定理知, 当且仅当误差函数

$$x^n-p_{n-1}^*(x)=x^n-\sum_{i=0}^{n-1}c_ix^i=p_n^*(x)$$

在 $[-1,1]$ 上的 $n+1$ 个点处符号交错并取到最大或最小值. 另一方面, 注意到切比雪夫多项式 $T_n(x)$ 在 $n+1$ 个点

$$x_i=\cos\frac{i\pi}{n},\quad i=0,1,\cdots,n,$$

处符号交错并取到最大值 1 或最小值 -1, 且 $T_n(x)$ 是首项系数为 2^{n-1} 的多项式. 因 $p_n^*(x)$ 是首项系数为 1 的 n 次多项式, 故可大胆猜测 $p_n^*(x)=2^{1-n}T_n(x)$.

定理 9.21 对于任意的 $p_n\in\mathbb{P}_n^1[x]$, 有

$$\|p_n\|=\max_{x\in[-1,1]}|p_n(x)|\geqslant\|2^{1-n}T_n(x)\|=\frac{1}{2^{n-1}}.$$

当且仅当 $p_n(x)=2^{1-n}T_n(x)$ 时, 等号成立.

证明 用反证法. 假设命题不成立, 则存在多项式 $p_n\in\mathbb{P}_n^1[x]$, 使得

$$|p_n(x)|<\frac{1}{2^{n-1}},\quad\forall x\in[-1,1].$$

若记 $q_{n-1}(x)=2^{1-n}T_n(x)-p_n(x)\neq 0$, 则 $q_{n-1}(x)$ 是一个多项式, 次数至多为 $n-1$. 在 $T_n(x)$ 的极值点 $\{x_i\}_{i=0}^n$ 处, 我们有

$$q_{n-1}(x_i)=2^{1-n}T_n(x_i)-p_n(x_i)=\frac{(-1)^i}{2^{n-1}}-p_n(x_i).$$

利用前面的假设条件, 知 $q_{n-1}(x_i)$ 与 $T_n(x_i)$ 同号. 所以, $q_{n-1}(x)$ 在 $[-1,1]$ 上至少发生 n 次变号, 利用介值定理知, $q_{n-1}(x)$ 至少有 n 个零点. 而另一方面, $q_{n-1}(x)$ 的次数不超过 $n-1$, 至多有 $n-1$ 个零点, 产生矛盾!

容易看出, 当且仅当 $q_{n-1}(x)\equiv 0 \Longleftrightarrow p_n(x)=2^{1-n}T_n(x)$ 时, 等号成立. □

下面, 我们讨论多项式插值余项的极小化问题. 设函数 $f(x)\in C^{n+1}[-1,1]$, 若取 $n+1$ 个互不相同的节点 $-1<x_0<x_1<\cdots<x_n<+1$, 构造 $f(x)$ 的 n 次插值多项式 $q_n(x)$, 则有

$$\|f-q_n\|\leqslant \frac{\|f^{(n+1)}\|}{(n+1)!}\max_{x\in[-1,1]}|(x-x_0)(x-x_1)\cdots(x-x_n)|.$$

一个自然的问题是: 如何选取节点 x_i, $i=0,1,\cdots,n$ 使得插值余项尽可能的小, 等价于寻找多项式 $p_{n+1}(x)=(x-x_0)(x-x_1)\cdots(x-x_n)\in\mathbb{P}^1_{n+1}[x]$, 使其在区间 $[-1,1]$ 的零偏差最小.

利用定理9.21知, 当节点取 $n+1$ 次切比雪夫多项式的零点时, 即

$$x_{i-1}=\cos\frac{(2i-1)\pi}{2n+2},\quad i=1,2,\cdots,n+1,$$

$\|p_{n+1}(x)\|$ 取到最小值. 此时, 多项式插值余项为

$$\|f-q_n\|\leqslant\frac{\|f^{(n+1)}\|}{(n+1)!}\cdot 2^{-n}\|T_{n+1}\|=\frac{\|f^{(n+1)}\|}{2^n(n+1)!}.$$

注解 9.6 不难看出, 当 $f^{(n+1)}(x)$ 在 $[-1,1]$ 上变化不大时, 取切比雪夫多项式的零点作为插值节点, 此时 $q_n(x)$ 可视为 $f(x)$ 的近似最佳逼近多项式. 另外, 如果插值区间为 $[a,b]$ 时, 利用仿射变换 $t=[a+b+(b-a)x]/2$, 插值节点可取为

$$x_{i-1}=\frac{a+b}{2}+\frac{b-a}{2}\cos\frac{(2i-1)\pi}{2n+2},\quad i=1,2,\cdots,n+1.$$

最后, 我们研究切比雪夫逼近问题, 即用切比雪夫多项式级数来逼近函数. 对于任意的 $f\in C[-1,1]$, 取权函数 $\rho(x)=(1-x^2)^{-1/2}$, 可按空间 $L^2_\rho[-1,1]$ 的内积构造 n 次最佳平方逼近多项式

$$S_n(x)=c_0\frac{T_0(x)}{2}+\sum_{i=1}^{n}c_iT_i(x),$$

其中

$$c_i=\frac{(f,T_i)}{(T_i,T_i)}=\frac{2}{\pi}\int_{-1}^{1}\frac{f(x)T_i(x)}{\sqrt{1-x^2}}\,\mathrm{d}x,\quad i=0,1,\cdots,n,$$

称 $S_n(x)$ 为函数 $f(x)$ 按切比雪夫多项式展开的部分和. 可以证明:

$$\lim_{n\to\infty}\|f-S_n\|_2^2=\lim_{n\to\infty}\int_{-1}^{1}\frac{[f(x)-S_n(x)]^2}{\sqrt{1-x^2}}\,\mathrm{d}x=0.$$

此外, 如果 $f \in C^1[-1,1]$, 那么

$$\lim_{n\to\infty} \|f - S_n\|_\infty = \lim_{n\to\infty} \max_{x\in[-1,1]} |f(x) - S_n(x)| = 0,$$

即 $S_n(x)$ 一致收敛于 $f(x)$. 进一步, 如果 $f \in C^r[-1,1]$, 那么存在由 f 和 r 决定的常数 c, 使得

$$|c_i| \leqslant \frac{c}{i^r}, \quad i = 1, 2, \cdots, n.$$

因此, 对于足够光滑的函数 f, 随着 i 的增大, 它的切比雪夫多项式展开系数 c_i 迅速趋于零. 对于截断的展开式 $S_n(x)$, 如果 $c_{n+1} \neq 0$ 且系数 c_i 迅速趋于零, 那么

$$f(x) - S_n(x) = \sum_{i=n+1}^{\infty} c_i T_i(x) \approx c_{n+1} T_{n+1}(x),$$

而 $T_{n+1}(x)$ 恰好有 $n+2$ 个相等的极大和极小值, 故 $S_n(x)$ 相似于最佳一致逼近多项式. 下面来看一个例子.

例 9.16 设 $f(x) = e^x$, 求 $f(x)$ 在区间 $[-1,1]$ 上的三次最佳平方逼近切比雪夫多项式.

解 按切比雪夫多项式展开公式, 知

$$S_3(x) = c_0 \frac{T_0(x)}{2} + \sum_{i=1}^{3} c_i T_i(x),$$

其中

$$c_i = \frac{2}{\pi} \int_{-1}^{1} \frac{f(x) T_i(x)}{\sqrt{1-x^2}}\, \mathrm{d}x = \frac{2}{\pi} \int_0^{\pi} e^{\cos\theta} \cos(j\theta)\, \mathrm{d}\theta, \quad j = 0, 1, 2, 3.$$

利用数值积分可求得

$$c_0 \approx 2.5321318, \quad c_1 \approx 1.1303182, \quad c_2 \approx 0.2714953, \quad c_3 \approx 0.0443369.$$

综上, 所求的三次最佳平方逼近切比雪夫多项式

$$S_3(x) \approx 1.2660659 + 1.1303182 T_1(x) + 0.2714953 T_2(x) + 0.0443369 T_3(x),$$

其形状如图9.9(a)所示, 逼近误差见图9.9(b).

□

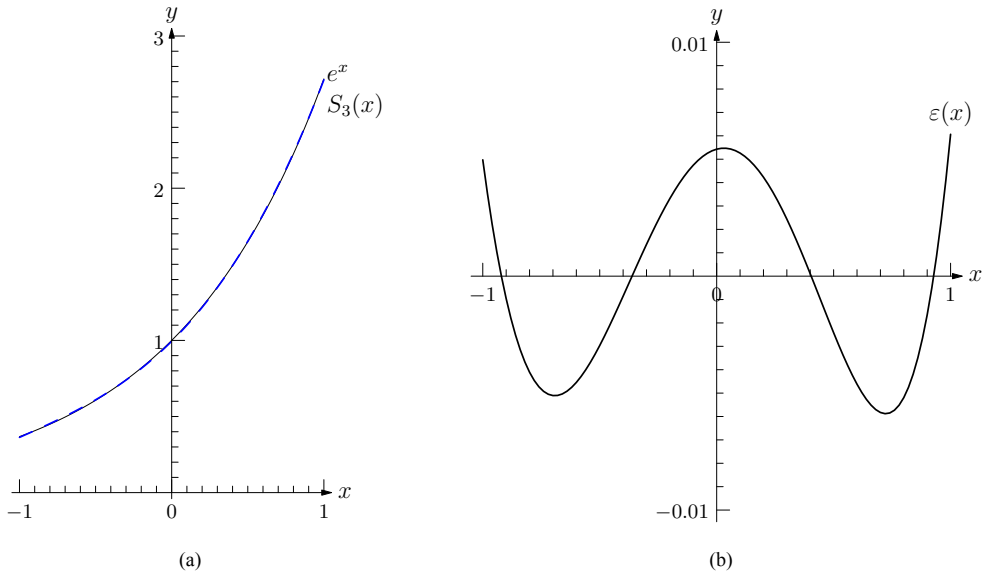


图 9.9: e^x 在 $[-1,1]$ 的三次最佳平方逼近切比雪夫多项式

从逼近误差的分布情况来看, 切比雪夫多项式级数 $S_n(x)$ 与最佳一致逼近具有非常相似的特征, 即误差函数是均匀分布的. 鉴于最佳一致逼近多项式往往难于求得, 而最佳平方逼近多项式相对易于计算, 故用切比雪夫多项式 $S_n(x)$ 来近似求解最佳一致逼近问题是一个常用的方法.

注解 9.7 基于类似的想法, 我们可以处理高次多项式的低次逼近问题. 在许多实际应用中, 多项式作为一类常用的函数被广泛的使用, 具有简单、高效等优点. 然而, 当多项式的次数较高时, 譬如 $n \geqslant 20$, 这时多项式往往会出现出不稳定, 强震荡等现象, 计算效率也降低了, 制约了它的使用范围. 因此, 一个自然的想法是: 用低次的多项式去逼近高次的多项式. 由于切比雪夫多项式 $\{T_i(x)\}_{i=0}^n$ 和幂函数 $\{x^i\}_{i=0}^n$ 分别构成函数空间 $\mathbb{P}_n[x]$ 的一组基, 故它们直接可以相互线性表示, 见表9.2.

表 9.2: $\{T_i(x)\}_{i=0}^n$ 与 $\{x^i\}_{i=0}^n$ 之间的相互线性表示

$T_0 = 1$	$1 = T_0$
$T_1 = x$	$x = T_1$
$T_2 = 2x^2 - 1$	$x^2 = (T_0 + T_2)/2$
$T_3 = 4x^3 - 3x$	$x^3 = (3T_1 + T_3)/4$
$T_4 = 8x^4 - 8x^2 + 1$	$x^4 = (3T_0 + 4T_2 + T_4)/8$
$T_5 = 16x^5 - 20x^3 + 5x$	$x^5 = (10T_1 + 5T_3 + T_5)/16$
$T_6 = 32x^6 - 48x^4 + 18x^2 - 1$	$x^6 = (10T_0 + 15T_2 + 6T_4 + T_6)/32$
$T_7 = 64x^7 - 112x^5 + 56x^3 - 7x$	$x^7 = (35T_1 + 21T_3 + 7T_5 + T_7)/64$
$T_8 = 128x^8 - 256x^6 + 160x^4 - 32x^2 + 1$	$x^8 = (35T_0 + 56T_2 + 28T_4 + 8T_6 + T_8)/128$
$T_9 = 256x^9 - 576x^7 + 432x^5 - 120x^3 + 9x$	$x^9 = (126T_1 + 64T_3 + 36T_5 + 9T_7 + T_9)/256$
$\vdots$	$\vdots$

利用切比雪夫多项式级数的性质, 多项式 $p_n(x) = \sum_{i=0}^{n} a_i x^i$ 的低次逼近可按以下步骤计算:

(1) 利用表9.2, 将多项式 $p_n(x)$ 写成切比雪夫多项式级数的形式, 即

$$p_n(x) = \sum_{i=0}^{n} c_i T_i(x);$$

(2) 取切比雪夫多项式级数的前 m 项, 记为

$$q_m(x) = \sum_{i=0}^{m} c_i T_i(x);$$

(3) 再利用表9.2, 将多项式 $q_n(x)$ 写成幂级数的形式, 即

$$q_m(x) = \sum_{i=0}^{m} \hat{a}_i x^i.$$

例 9.17 设 $f(x) = e^{-x}$, 其 Taylor 展开式的前 10 项为

$$e^{-x} \approx p_9(x) \triangleq 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots + \frac{x^9}{9!},$$

利用表9.2, 多项式 $p_9(x)$ 可以写成 Chebyshev 多项式级数的形式, 即

$$\begin{aligned} p_9(x) \approx\ & 1.2661 \times T_0(x) - 1.1303 \times T_1(x) + 0.2715 \times T_2(x) - 0.0443 \times T_3(x) \\ & + 0.005474 \times T_4(x) - 0.000543 \times T_5(x) + 0.000045 \times T_6(x) \\ & - 0.000003198 \times T_7(x) + 0.0000001992 \times T_8(x) - 0.00000001104 \times T_9(x). \end{aligned}$$

容易看出, 随着 k 增大, $T_k(x)$ 的系数迅速变小, 又因 $|T_k(x)| \leqslant 1$, 故可略去次数较高的 $T_k(x)$ 项. 此时, 逼近多项式的次数显著降低了, 从而大大节省了计算工作量.

若要求 $[-1, 1]$ 上逼近 e^{-x} 的绝对误差不超过 0.00005 的多项式, 可只取 $T_5(x)$ 以前的项作近似, 即

$$\begin{aligned} S_5(x) \approx\ & 1.2661 \times T_0(x) - 1.1303 \times T_1(x) + 0.2715 \times T_2(x) - 0.0443 \times T_3(x) \\ & + 0.005474 \times T_4(x) - 0.000543 \times T_5(x) \\ =\ & 1.000045 - 1.000022x + 0.499199x^2 - 0.166488x^3 + 0.043794x^4 - 0.008687x^5. \end{aligned}$$

而若直接按 Taylor 展开式截断到含有 x^5 的项, 则有

$$p_5(x) \triangleq 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} - \frac{x^5}{5!},$$

此时, $|e^{-x} - p_5(x)| \leqslant 1/6! + 1/7! + \cdots \approx 0.0016$, 约为前者绝对误差的 33 倍.

多项式 $S_5(x)$ 的形状见图9.10(a), 函数 e^{-x} 的两种多项式逼近的误差见图9.10(b), 其中 $\varepsilon_1(x) = e^{-x} - S_5(x)$, $\varepsilon_2(x) = e^{-x} - p_5(x)$. 可以看出, 当 x 在原点附件时, Taylor 逼近误差非常小, 但越偏离原点, 其误差就越大. 而用 Chebyshev 多项式级数构造的逼近函数 $S_5(x)$, 整体误差更小且分布更均匀, 因此可作为 e^{-x} 的近似最佳一致逼近多项式.

9.8 函数逼近的若干重要定理

函数逼近论是现代数学的一个重要分支, 起源于 1885 年维尔斯特拉斯 (Weierstrass) 所建立的关于连续函数可以用多项式逼近的著名定理和 1859 年切比雪夫 (Chebyshev) 提出的最佳逼近的特征定理. 自上个世纪以来, 经 Jackson, Bernstein 以及前苏联学派的一系列深刻工作所推动, 它成为一门独立的学科并得以蓬勃发展, 其研究目标明确为用简单的可计算函数对一般函数的逼近, 进而分析这种逼近的程度和被逼近函数本身的特性. 近些年来, 人们对利用插值多项式, 有理函数, Müntz 多项式, 样条函数, 小波, 神经网络等作为逼近工具的问题进行了深入的研究. 函数逼近论不仅与泛

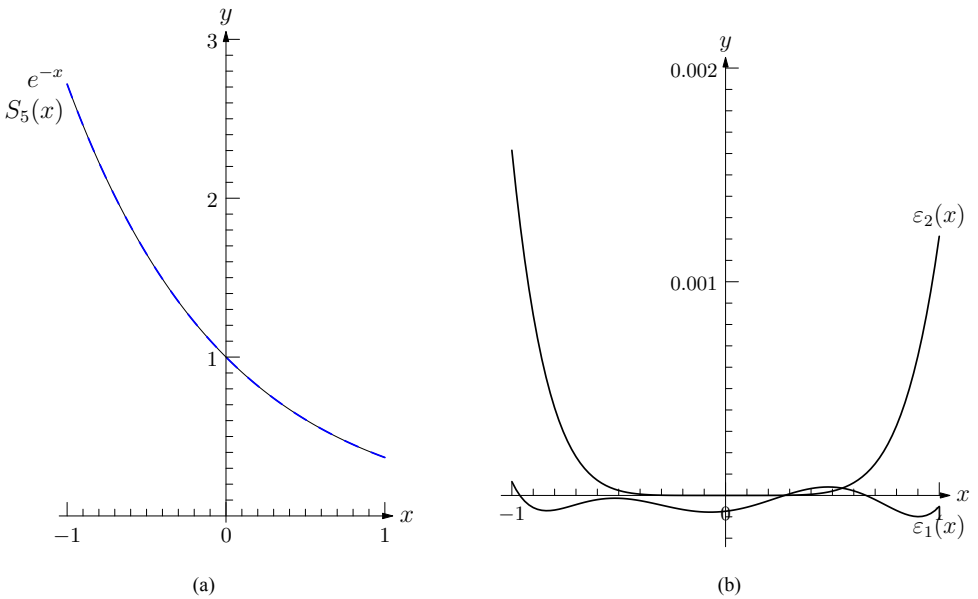


图 9.10: e^{-x} 在 $[-1,1]$ 的多项式逼近

函分析, 调和分析, 微分方程, 代数等学科密切相关, 而且已成为计算数学与应用数学的理论基础. 本小节简要介绍函数逼近论中的一些重要定理.

定理 9.22 (Weierstrass 第一定理) 设 $f(x)$ 是区间 $[a,b]$ 上的连续函数, 则对任意 $\varepsilon > 0$, 存在多项式 $p(x)$, 使得

$$\|f - p\| = \max_{x\in[a,b]} |f(x) - p(x)| < \varepsilon$$

成立.

该定理告诉我们可以用多项式按预先指定的精度来一致逼近连续函数, 即多项式函数空间是连续函数空间的稠密子集. Weierstrass 第一定理有许多不同的证明方法, 其中一个构造性的证明是伯恩斯坦 (Bernstein) 给出的, 他引进了如下的伯恩斯坦多项式:

$$B_n(f;x) = \sum_{i=0}^{n} f(\frac{i}{n})\binom{n}{i}x^i(1-x)^{n-i}, \quad x \in [0,1].$$

明显地, $B_n(f;x)$ 是区间 $[0,1]$ 上的 n 次多项式. 可以证明: 对于任意的函数 $f(x) \in C[0,1]$, 当 $n \to \infty$, 伯恩斯坦多项式 $B_n(f;x)$ 一致收敛于 $f(x)$. 而一般的区间 $[a,b]$,

可通过仿射变换 $y = (x-a)/(b-a)$ 转化为区间 $[0,1]$ 来证明. 进一步, 如果函数 $f(x) \in C^r[0,1]$, 那么

$$\lim_{n\to\infty} \|f^{(i)} - B_n^{(i)}(f)\| = \lim_{n\to\infty} \max_{x\in[0,1]} |f^{(i)}(x) - B_n^{(i)}(f;x)| = 0, \quad i = 1, 2, \cdots, r,$$

即伯恩斯坦多项式 $B_n(f;x)$ 的导函数亦一致收敛于 $f(x)$ 的导函数.

注解 9.8 虽然伯恩斯坦多项式 $B_n(f;x)$ 一致收敛于 $f(x)$, 但是它的收敛速度是相当慢的, 且随着 n 的不断增大, $B_n(f;x)$ 会变成次数很高的多项式, 其数值上是不稳定的. 因此, 在实际应用中, 伯恩斯坦多项式并不适于构造 $f(x)$ 的一致逼近多项式.

对于连续的周期函数, 有类似的结论. 不失一般性, 设周期为 2π, 记所有以 2π 为周期的连续函数全体为 $C_{2\pi}$, 定义

$$\|f\| = \max_{x\in(-\infty,+\infty)} |f(x)|,$$

可以证明 $\|\cdot\|$ 构成 $C_{2\pi}$ 的范数.

定理 9.23 (Weierstrass 第二定理) 设 $f(x)$ 是 $(-\infty,+\infty)$ 上以 2π 为周期的连续函数, 则对任意 $\varepsilon > 0$, 存在三角多项式 $t(x)$, 使得

$$\|f - t\| = \max_{x\in(-\infty,+\infty)} |f(x) - t(x)| < \varepsilon$$

成立.

可以证明: Weierstrass 第一定理与第二定理是相互等价的. Weierstrass 定理作为分析学的基本定理之一, 是逼近论的基础性定理, 有很多形式的推广, 譬如一般的正线性算子序列 (Korovkin 定理), 紧距离空间的子代数 (Stone 定理) 和赋范线性空间的基本集 (Müntz 定理), 详细内容可参见文献 [2, 5]. 甚至对于多变量连续函数, 也有类似的定理, 可参见文献 [13].

接下来, 我们介绍刻画逼近程度和被逼近函数本身特性的若干定理. 在数学分析中, 常用函数的可微分次数, 连续与间断等来描述函数的性质, 但是, 这些描述是不够精细的. 为了描述收敛速度的快慢, 需要引入连续模和光滑模的概念.

定义 9.12 设函数 $f(x)$ 在区间 I 上有定义, 其中 I 是有限或无限, 开或不开均可以. 对于 $h > 0$, 称

$$\omega(f;h) \triangleq \sup_{x, x+t\in I, |t|<h} |f(x+t) - f(x)|$$

为 $f(x)$ 在区间 I 上的**连续模**.

连续性模的几何意义是: 当 x_1, x_2 的距离小于 h 时, $f(x)$ 在 x_1, x_2 的值相差不会超过 $\omega(f;h)$. 明显地, $\omega(f;h)$ 是 $[0,+\infty)$ 上的非负单调增函数. 函数 $f(x)$ 在 I 上一致连续的充分必要条件是 $\lim\limits_{h\to 0^+}\omega(f;h)=0$.

若记 $\Delta_h f(x)=f(x+h)-f(x)$ 是函数 f 在点 x 的步长为 h 的一阶向前差分, 那么

$$\Delta_h^r f(x)\triangleq\Delta_h(\Delta_h^{r-1}f(x))=\sum_{i=0}^{r}(-1)^{r-i}\binom{r}{i}f(x+ih)$$

称为 f 在点 x 步长为 h 的 r 阶向前差分.

定义 9.13 设函数 $f(x)$ 在区间 $[a,b]$ 上的连续函数, 称

$$\omega_r(f;h)\triangleq\sup_{x,x+rt\in[a,b],|t|<h}|\Delta_t^r f(x)|$$

为 $f(x)$ 在区间 I 上的r **阶光滑模**. 当 $r=1$ 时, r 阶光滑模就是连续模.

容易看出, $\omega_r(f;h)$ 是 h 的增函数, 连续函数, 且 $\omega_r(f;0)=0$. 不难证明:

$$\omega_r(f;h)\leqslant 2^{r-1}\omega(f;h)\leqslant 2^r\|f\|_\infty.$$

进一步, 当 $f(x)\in C^r[a,b]$ 时, 有

$$\omega_r(f;h)\leqslant h^r\|f^{(r)}\|_\infty,$$

以及 $\lim\limits_{h\to 0}\omega_r(f;h)/h^r=0$ 的充分必要条件是 $f(x)\in\mathbb{P}_r[x]$.

对于 $f(x)\in C[a,b]$, 函数 f 的 n 次最佳一致多项式逼近误差记为

$$E_n(f)=\min_{p_n\in\mathbb{P}_n[x]}\|f-p_n\|_\infty.$$

而对于 $f(x)\in C_{2\pi}$, 函数 f 的 n 次最佳一致三角多项式逼近误差记为

$$E_n^*(f)=\min_{t_n\in\mathbb{T}_n[x]}\|f-t_n\|_\infty,$$

其中

$$\mathbb{T}_n[x]\triangleq\left\{\frac{a_0}{2}+\sum_{k=1}^{n}(a_k\cos kx+b_k\sin kx)\;\middle|\;a_0,a_k,b_k\in\mathbb{R}\right\}.$$

根据 Weierstrass 第一和第二定理知, 数列 $\{E_n(f)\}_{n=0}^\infty$ 和 $\{E_n^*(f)\}_{n=0}^\infty$ 均收敛于 0.

定理 9.24 (Jackson 定理) 设 $f(x)\in C_{2\pi}$, 则存在常数 C 使得

$$E_n^*(f)\leqslant C\omega\Big(f;\frac{1}{n}\Big),\quad n=1,2,\cdots,$$

以及

$$E_n^*(f)\leqslant C\omega_2\Big(f;\frac{1}{n}\Big),$$

成立.

若函数 f 有更高阶的连续导数, 则 $E_n^*(f)$ 有更快的收敛速度.

定理 9.25 设 $f(x)\in C_{2\pi}$, 且具有 r 阶的连续导数, 则存在常数 C_r 使得

$$E_n^*(f)\leqslant \frac{C_r}{n^r}\omega\Big(f^{(r)},\frac{1}{n}\Big),\quad n=1,2,\cdots,$$

成立.

另一方面, 若已知 $E_n^*(f)$ 的收敛速度, 则可反过来估计函数的连续模和光滑模.

定义 9.14 设 $f(x)$ 是区间 I 上的函数, 若 $f(x)$ 满足

$$|f(x)-f(y)|\leqslant C|x-y|^\alpha,\quad \forall x,y\in I,$$

其中 $C,\alpha\in\mathbb{R}$ 是正常数, 则称 $f(x)$ 满足**李普希茨条件**, 记作 $f\in \mathrm{Lip}_C\alpha$ 或 $f\in \mathrm{Lip}\alpha$.

定理 9.26 设 $f(x)\in C_{2\pi}$, 则存在常数 $C_r(r=1,2,\cdots)$ 使得

$$\omega_r(f,h)\leqslant C_r h^r\sum_{0\leqslant n\leqslant h^{-1}}(n+1)^{r-1}E_n^*(f),\quad h>0,$$

成立.

定理 9.27 (Bernstein 逆定理) 设 $f(x)\in C_{2\pi}$, 则 $f\in \mathrm{Lip}\alpha\ (0<\alpha<1)$ 的充分必要条件是

$$E_n^*(f)=O(n^{-\alpha}).$$

类似地, 基于代数多项式的连续函数类的最佳逼近有类似的定量理论, 但与基于三角多项式的最佳逼近理论有不同的性态, 详细的内容参见文献 [2, 5].

在函数逼近论中, 还有许多非常重要的课题, 譬如宽度, 熵, 最优恢复, 样条函数的逼近理论等, 具体可参见专著 [14, 16, 19]. 至于多变量的函数逼近理论, 目前虽然有了一些结果, 但是还很不完善, 可参见专著 [13, 15]. 在应用方面, 近些年来流行的小波分析 [9], 压缩感知 [10], 深度学习 [12] 等都大量用到了函数逼近理论, 使得它成为一门经久不衰的学科.

参考文献

[1] Kendall Atkinston. An Introduction to Numerical Analysis. John Wiley & Sons, 1978.

[2] Ward Cheney, David Kincaid. An Introduction to Approximation Theory. AMS Cheslea Publishing, 1982.

[3] Philip J. Davis. Interpolation and Approximation. Dover Publications Inc, 1975.

[4] Michael J. D. Powell. Approximation Theory and Methods. Cambridge University Press, 1981.

[5] George G. Lorentz. Approximation of Functions. 2nd Ed. AMS Chelsea Publishing, 1986.

[6] Charles F. Van Loan. Computational Frameworks for the Fast Fourier Transform. Society for Industrial and Applied Mathematics, 1992.

[7] David Kincaid, Ward Cheney. Numerical Analysis: Mathematics of Scientific Computing. 3rd ed. American Mathematical Society, 2002.

[8] Walter Gautschi. Orthogonal Polynomials: computation and approximation. Oxford Science Publications, 2004.

[9] Stéphane Mallat. A Wavelet Tour of Signal Processing: The Sparse Way. 3rd Ed. Elsevier Inc, 2009.

[10] Michael Elad. Sparse and Redundant Representations From Theory to Applications in Signal and Image Processing. Springer, 2010.

[11] Gene H.Golub, Charles F.Van Loan. 矩阵计算. 袁亚湘等译. 第 3 版. 人民邮电出版社, 2011.

[12] Ian Goodfellow, Yoshua Bengio, Aaron Courville. Deep Learning. MIT Press, 2016.

[13] 王仁宏, 梁学章. 多元函数逼近. 科学出版社, 1988.

[14] 孙永生, 房艮孙. 函数逼近论 (上下册). 北京师范大学出版社, 1989.

[15] 沈燮昌. 复变函数逼近论. 科学出版社, 1992.

[16] 谢庭藩, 周颂平. 实函数逼近论. 杭州大学出版社, 1998.

[17] 蒋尔雄, 赵风光, 苏仰锋. 数值逼近. 复旦大学出版社, 2007.

[18] 黄云清, 舒适, 陈艳萍, 金继承, 文立平. 数值计算方法. 科学出版社, 2009.

[19] 冯玉瑜, 曾芳玲, 邓建松. 样条函数与逼近论. 中国科学技术大学出版社, 2012.

习 题

1. 验证 $\|f\| = \max\limits_{x\in[a,b]} |f(x)|$ 构成线性空间 $C[a,b]$ 的一个范数, 并证明:

$$\|fg\| \leqslant \|f\|\|g\|, \quad \forall f, g \in C[a,b].$$

2. 判断: 当 $0 < p < 1$ 时,

$$\|\mathbf{x}\|_p = \left(|x_1|^p + |x_2|^p + \cdots + |x_n|^p\right)^{1/p}, \quad \forall \mathbf{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}} \in \mathbb{R}^n$$

是否构成线性空间 $\mathbb{R}^n$ 的一个范数? 如果是请给出证明, 否则指出原因.

3. 下列公式中:

(a) $d(x,y) = \log(1 + |x-y|)$;

(b) $d(x,y) = e^{|x-y|} - 1$;

(c) $d(x,y) = |x-y|^2$,

哪一个是定义在实轴上的距离函数? 并给出证明.

4. 设 $C[a,b]$ 中有一函数列 $\{f_n\}_{n=0}^{\infty}$, 证明: f_n 按一致范数收敛于 g 的充分必要条件是 f_n 一致收敛于 g.

5. 验证 $\int_a^b f(x)g(x)\,\mathrm{d}x$ 构成线性空间 $C[a,b]$ 的一个内积.

6. 设 $f(x), g(x) \in C^1[a,b]$, 定义

$$(f,g) = \int_a^b f'(x)g'(x)\,\mathrm{d}x,$$

问 $(\cdot,\cdot)$ 是否为 $C^1[a,b]$ 的一个内积? 记空间

$$C_0^1[a,b] = \{f(x) \mid f(a) = 0, f(x) \in C^1[a,b]\},$$

问 $(\cdot,\cdot)$ 是否为 $C_0^1[a,b]$ 的一个内积? 如果是请给出证明, 否则指出原因.

7. 试利用 Gram-Schmidt 正交化算法, 求 $[0,1]$ 上的三次多项式空间关于内积

$$(f,g) = \int_0^1 \sqrt{x} f(x)g(x)\,\mathrm{d}x$$

的一组正交基.

8. 设 V 是内积空间, $M \subset V$ 为有限维子空间. 对任意的 $x \in V$, 记 $P(x)$ 为子集 M 逼近 x 的最佳逼近元, 使用 V 内积诱导的范数. 证明: $P(x)$ 是 $V \to M$ 的线性算子, 即投影算子.

9. 下面的资料于 1970 年 3 月提交美国参议院反托拉斯委员会, 说明各种等级汽车在碰撞时严重损伤的比率, 求此数据的最小二乘拟合直线 (1 磅 $\approx$ 0.45 千克).

车型	车重 (磅)	严重损伤的比率 (%)
美国制豪华型	4800	3.1
美国制中级型	3700	4.0
美国制经济型	3400	5.2
美国制轻便型	2800	6.4
外国制轻便型	1900	9.6

10. 试分别用一次、二次多项式拟合下面的数据, 并给出最小平方误差.

x_i	-1.00	-0.50	0.00	0.25	0.75
y_i	0.22	0.80	2.00	2.50	3.80

11. 对下列数据用最小二乘法求形如 $\varphi(x) = a\cos x + b\sin x$ 的经验公式.

x_i	0.20	0.25	0.30	0.50
y_i	1.36	1.20	1.02	0.32

12. 对下列数据用最小二乘法求形如 $\varphi(x) = ae^{bx}$ 的经验公式.

x_i	0.25	0.50	0.70	0.75
y_i	2.57	1.21	0.99	4.23

13. 对下列数据用最小二乘法求形如 $\varphi(x) = \dfrac{x}{a + bx}$ 的拟合函数.

x_i	2.10	2.50	2.80	3.20
y_i	0.6087	0.6849	0.7368	0.8111

14. 用最小二乘法求解下列矛盾方程组:

$$(1)\begin{cases} x_1+2x_2=5,\\ 2x_1+x_2=6,\\ x_1+x_2=4; \end{cases} \qquad (2)\begin{cases} x_1-2x_2=1,\\ x_1+5x_2=13.1,\\ 2x_1+x_2=7.9,\\ x_1+x_2=5.1. \end{cases}$$

15. 试确定常数 $c_0,c_1\in\mathbb{R}$ 使得 $\int_0^1 \left|e^x-c_0-c_1x\right|^2 \mathrm{d}x$ 达到极小, 并求出极小值.

16. 求函数 $f(x)=\cos x$ 在区间 $[0,1]$ 上关于权函数 $\rho(x)=\sqrt{x}$ 的三次最佳平方逼近多项式.

17. 设区间 $[a,b]$ 上 $\mathbb{P}_n(x)$ 的一组首项系数为 1 的正交多项式基 $\{g_l^*(x)\}_{l=0}^n$, 证明: 对于 $l\geqslant 1$, 多项式 $g_l^*(x)$ 与 $g_{l+1}^*(x)$ 的零点必交错, 即

$$a<x_1^{l+1}<x_1^l<x_2^{l+1}<\cdots<x_l^{l+1}<x_l^l<x_{l+1}^{l+1}<b,$$

其中 x_i^l, x_i^{l+1} 分别为 $g_l^*(x), g_{l+1}^*(x)$ 的零点. (提示: 从递推关系式 (9.13) 出发, 先证 $a<a_k<b, 0<b_k\leqslant \max\{a^2,b^2\}$, 再用数学归纳法.)

18. 设 $f(x)=x^2$, 求 $f(x)$ 在区间 $[-\pi,\pi]$ 上的二次最佳平方逼近三角多项式.

19. 设 $f(x)=x$, 求 $f(x)$ 在区间 $[-1,1]$ 上的 n 次最佳平方逼近三角多项式.

20. 实现 FFT 算法, 输入向量 $\mathbf{f}=(-2,4,7,3,1,5,-3,4)$, 计算 $\mathbf{f}$ 的离散傅立叶变换 $\mathbf{g}=(g_0,g_1,\cdots,g_{n-1})$, 其中

$$g_l=\frac{1}{n}\sum_{k=0}^{n-1} f_k\omega^{kl},\quad \omega=e^{-\mathrm{i}2\pi/n},\quad n=8.$$

21. 设 $n=3^m$, 其中 $m\in\mathbb{N}$, 试建立快速离散傅立叶变换的递推公式.

22. 试确定常数 $c\in\mathbb{R}$ 使得 $\max\limits_{x\in[0,1]}|x^2-cx|$ 达到极小, 并判断该解是否唯一.

23. 设 $f(x)\in C[a,b]$, 求 $f(x)$ 的零次最佳一致逼近多项式.

24. 设 $f(x)\in C[-a,a]$, 记 $p_n^*(x)$ 是 $f(x)$ 的 n 次最佳一致逼近多项式, 证明: 当 $f(x)$ 是偶 (奇) 函数时, $p_n^*(x)$ 亦是偶 (奇) 函数.

25. 设 $f(x)\in C^2[a,b]$, 且 $f''(x)\neq 0$. 设 $f(x)$ 在 $[a,b]$ 上的一次最佳一致逼近多项式为 $p_1^*(x)=c_0+c_1x$.

(a) 证明: $c_1 = f'(c) = \dfrac{f(b)-f(a)}{b-a}$, $c_0 = \dfrac{f(a)+f(c)}{2} - \dfrac{f(b)-f(a)}{b-a}\cdot\dfrac{a+c}{2}$;

(b) 求 $f(x)=\cos x$ 在 $[0,\pi/2]$ 上的一次最佳一致逼近多项式.

26. 求多项式 $p(x)=6x^3+3x^2+x+4$ 在 $[-1,1]$ 上的二次最佳一致逼近多项式.

27. 求函数 $f(x)=\cos\dfrac{\pi}{2}x$ 在 $[-1,1]$ 上关于权函数 $\rho(x)=(1-x^2)^{-1/2}$ 的三次最佳平方逼近多项式.

28. 求函数 $f(x)=\arctan x$ 在 $[-1,1]$ 上的近似最佳一致逼近多项式, 要求绝对误差不超过 10^{-3}.

29. 证明: 切比雪夫多项式 $T_n(x)$ 满足关系式

$$T_{n+m}(x)+T_{n-m}(x)=2T_n(x)T_m(x),\quad n\geqslant m.$$

30. 设 $f(x)\in C[a,b]$, $M_n=\max\limits_{x\in[a,b]}|f^n(x)|$, 若取

$$x_{k-1}=\frac{a+b}{2}+\frac{b-a}{2}\cos\frac{2k-1}{2n+2}\pi,\quad k=1,2,\cdots,n+1,$$

作为插值节点构造 n 次多项式插值函数 $p_n(x)$. 证明: $f(x)$ 的 n 次多项式插值余项 $R(x)=f(x)-p_n(x)$ 满足

$$\max_{x\in[a,b]}|R(x)|\leqslant\frac{M_n}{n!}\frac{(b-a)^n}{2^{2n-1}}.$$

第十章　最优化方法

最优化方法是一门研究求解数学问题最优解的学科, 即对于给定的实际问题, 从众多的解中选取最优的解. 在日常生活中, 我们经常遇到这类问题: 上下班如何规划乘车路线, 才能快速又经济地到达公司; 旅游中如何选择航班和宾馆, 既省钱又能玩得开心. 从数学意义上说, 最优化方法是一种求函数极值的方法, 即在一组条件为等式或不等式的约束下, 使选定的目标函数达到极值, 即最大值或最小值. 研究内容包括最优化模型的建立、分析、求解及应用, 譬如最优解的条件或标准, 求解的算法, 以及收敛性、时间复杂度分析等. 随着计算机的快速发展和普及, 最优化方法在经济规划、工程设计、生产管理、交通运输、国防安全等领域得到了广泛的应用, 发挥着越来越重要的作用, 自身也得到了迅速的发展.

最优化问题可以追溯到十分古老的极值问题, 例如阿基米德 (Archimedes) 证明：给定周长, 圆所包围的面积为最大, 这是欧洲古代城堡几乎都建成圆形的原因之一. 然而, 最优化方法成为一门独立的学科是在第二次世界大战前后, 由于军事上的需要以及科学技术和生产的迅速发展, 许多实际的最优化问题已经无法用古典方法来解决, 从而促进了近代最优化方法的产生. 具有代表性的工作有: 以美国的丹齐格 (Dantzig) 和苏联的康托罗维奇 (Kantorovich) 为代表的线性规划; 以美国的库恩 (Kuhn) 和塔克尔 (Tucker) 为代表的非线性规划；以美国的贝尔曼 (Bellman) 为代表的动态规划；以苏联的庞特里亚金 (Pontryagin) 为代表的极大值原理等, 这些方法后来都形成体系，成为很活跃的领域. 此外, 近些年来在实际应用应用的驱动下, 譬如信号处理, 机器学习, 推荐系统, 自动驾驶等, 凸优化, 非光滑优化, 整数规划等得到了深入的研究.

目前, 最优化方法已经成为一门独立且应用广泛的学科, 内容非常丰富, 限于篇幅, 下面重点介绍最优化方法中求解线性规划的单纯形法与求解无约束非线性优化问题的梯度下降法.

10.1 线性规划问题

在生产管理和经营活动中, 经常会遇到一类问题: 如何合理利用有限的人力、物力、财力等资源, 以取到最佳的经济效益.

例 10.1 某工厂在计划期内要安排生产 I、II 两种产品, 已知生产每种单位产品所需的设备台数及 A、B 两种原材料的消耗, 如表10.1所示.

产品	产品 I	产品 II	现有资源
设备	4 台时/件	2 台时/件	18 台时
原材料 A	4 kg/件	1 kg/件	16 kg
原材料 B	1 kg/件	3 kg/件	12 kg

表 10.1: 产品的资源需求

该工厂每生产一件 I 产品可获利 2 元, 每生产一件 II 产品可获利 5 元, 问应该如何安排计划使该工厂获利最多?

解 设 x_1, x_2 分别表示在计划期内生产 I、II 产品的数量. 因为设备的总台数是 8, 这是一个限制性条件, 所以在优化 x_1, x_2 时, 需满足不等式约束:

$$4x_1 + 2x_2 \leqslant 18.$$

同理, 因原材料 A、B 的总量限制, 得到以下不等式约束:

$$4x_1 + x_2 \leqslant 16,$$
$$x_1 + 3x_2 \leqslant 12.$$

此外, 产量 x_1, x_2 明显是非负的, 满足 $x_1, x_2 \geqslant 0$.

若用 z 表示利润, 则 $z = 2x_1 + 5x_2$. 该工厂的目标是在不超过现有资源限量的条件下, 确定产量 x_1, x_2 使得利润 z 最大化, 用数学模型可表示为

$$\text{目标函数} \qquad \max z = 2x_1 + 5x_2$$

$$\text{约束条件} \qquad \begin{cases} 4x_1 + 2x_2 \leqslant 18, \\ 4x_1 + x_2 \leqslant 16, \\ x_1 + 3x_2 \leqslant 12, \\ x_1, x_2 \geqslant 0. \end{cases}$$

□

例 10.2 生产某汽车需要用 I、II、III 三种规格的轴各一根, 长度规格分别为 1.5、1、0.7 米, 它们需要用一种圆钢来制作, 圆钢的长度为 4 米. 现在要制造 1000 辆这种类型的汽车, 问至少需要多少根圆钢以满足生产需求?

解 圆钢的长度是 4 米, I、II、III 轴的长度分别为 1.5、1、0.7 米, 因此, 在不浪费材料的前提下, 可有以下不同的切法:

切法	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
I	2	2	1	1	1	0	0	0	0	0
II	1	0	2	1	0	4	3	2	1	0
III	0	1	0	2	3	0	1	2	4	5

若将这些不同的切法 $\{x_i\}_{i=1}^{10}$ 的数量设为变量, 用 z 表示圆钢的总数, 则 $z = x_1 + x_2 + \cdots + x_{10}$, 问题的数学模型可表示为

$$\text{目标函数} \qquad \min z = x_1 + x_2 + \cdots + x_{10}$$

$$\text{约束条件} \qquad \begin{cases} 2x_1 + 2x_2 + x_3 + x_4 + x_5 \geqslant 1000, \\ x_1 + 2x_3 + x_4 + 4x_6 + 3x_7 + 2x_8 + x_9 \geqslant 1000, \\ x_2 + 2x_4 + 3x_5 + x_7 + 2x_8 + 4x_9 + 5x_{10} \geqslant 1000, \\ x_1, x_2, \cdots, x_{10} \geqslant 0. \end{cases}$$

□

通过以上两个例子, 可以看出最优化问题的数学模型由三个要素组成:

(1) **变量**或称**决策变量**, 即问题中要优化的未知量, 用于描述问题中用数量表示的方案、措施等, 其值可由决策者控制和确定;

(2) **目标函数**, 即决策变量的函数, 按优化目标在这个函数前加上 max 或 min, 表示目标函数欲取最大值或最小值;

(3) **约束条件**, 即刻画决策变量取值时受到的各种资源条件的限制, 通常表示为含决策变量函数的等式或不等式.

如果在优化问题的数学模型中, 决策变量的取值是连续的, 目标函数是决策变量的线性函数, 约束条件是含决策变量的线性等式或不等式, 则称它为**线性规划问题**, 其一般的形式为

$$\begin{aligned}
&\max(\min)\, z = c_1x_1 + c_2x_2 + \cdots + c_nx_n,\\
&\text{s. t.}\begin{cases}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \leqslant (=, \geqslant) b_1,\\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \leqslant (=, \geqslant) b_2,\\
\cdots\\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \leqslant (=, \geqslant) b_m,\\
x_1, x_2, \cdots, x_n \geqslant 0.
\end{cases}
\end{aligned}$$

或简写为

$$\begin{aligned}
&\max(\min)\, z = \sum_{i=1}^{n} c_ix_i,\\
&\text{s. t.}\begin{cases}
\displaystyle\sum_{j=1}^{n} a_{ij}x_j \leqslant (=, \geqslant) b_i, \quad i = 1, 2, \cdots m,\\
x_i \geqslant 0, \quad j = 1, 2, \cdots n.
\end{cases}
\end{aligned}$$

在实际问题中, 线性规划模型是建立在以下假设基础之上的:

(1) 比例性, 指每个决策变量 x_j 在约束条件中和在目标函数中数值变化时, 按 x_j 对应的系数 a_{ij} 与价值系数 c_j 严格的成比例变化, 如生产某产品对资源的消耗和可获取的利润, 同其生产数量成比例;

(2) 可加性, 指目标函数的总值是各项组成部分值 c_ix_i 之和; 第 i 个约束关系式中各组成部分值之和就是第 i 项资源需求总量. 如生产多种产品时, 可获取的总利润是各项产品的利润之和, 对某项资源的消耗量等于各产品对该资源的消耗量之和. 决策变量是相互独立的, 不发生关联, 且不允许有交叉;

(3) 可分性, 即模型中的变量可以取小数、分数或某一实数;

(4) 确定性, 即模型中的参数均为确定的常数.

然而, 有些实际问题不符合上述条件, 例如每件产品售价 3 元, 但批量采购的话, 可以打七折. 对于这类不符合线性的条件, 在一些合理的假设下, 有时可看作近似满足线性条件, 也可用线性规划来建模.

注解 10.1 在前面的例子中, 产品的数量和不同切法的数量实际上应该是整数, 但是这类约束是非线性的, 甚至是不光滑的, 求解起来比较困难. 为此, 我们对问题作了简化, 允许它们取连续的实数. 如果优化问题中含有整数约束的变量, 则称它为**整数规划问题**, 相应的求解算法可参见 [9, 10].

10.2 线性规划问题的几何意义

当线性规划模型中只含 2 个变量时, 问题具有明显的几何意义, 可通过在平面上作图的方法求解, 称为**图解法**. 在介绍图解法之前, 我们先明确线性规划问题解的概念.

定义 10.1 称线性规划问题**有解**, 是指能找到一组值 $(x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$ 以满足所有约束条件; 否则, 称该问题**无解**. 任意一个满足约束条件的解 $(x_1, x_2, \cdots, x_n)$ 都称为该线性规划问题的一个**可行解**. 全部可行解构成的集合称为**可行域**. 在可行域中使目标函数值达到最优的可行解称为**最优解**.

图解法的基本步骤如下: 在平面上建立直角坐标系; 图示约束条件, 判别是否存在可行域, 如果存在, 则找出可行域; 图示目标函数, 寻找最优解. 下面通过例10.1来作具体说明. 先将其数学模型重写如下:

$$\max z = 2x_1 + 5x_2,$$

$$\text{s. t.} \begin{cases} 4x_1 + 2x_2 \leqslant 18, & \text{(10.1a)} \\ 4x_1 + x_2 \leqslant 16, & \text{(10.1b)} \\ x_1 + 3x_2 \leqslant 12, & \text{(10.1c)} \\ x_1, x_2 \geqslant 0. & \text{(10.1d)} \end{cases}$$

1. 以 x_1 为横坐标轴, x_2 为纵坐标轴, 适当选取坐标长度的单位, 建立平面直角坐标系. 由变量的非负约束条件(10.1d)知, 满足该条件的解均在第 I 象限内.
2. 依据约束条件, 找出可行域. 约束条件(10.1a)表示含直线 $4x_1 + 2x_2 = 18$ 上的点及其左下方的半平面, 约束条件(10.1b)、(10.1c)的含义类似. 同时满足(10.1a)-(10.1d)的点构成该线性规划问题的可行域, 如图10.1所示, 即凸多边形 $OP_1P_2P_3P_4$.
3. 图示目标函数. 随着 z 的变化, 方程 $z = 2x_1 + 5x_2$ 表示斜率为 $-2/5$ 的一族平行直线, 见图10.1, 其中向量 $\mathbf{v}$ 表示目标函数值 z 的增大方向.

4. 确定最优解. 因为最优解是可行域中使目标函数值 z 达到最大值的点, 从图10.1中可看出, 当代表目标函数的那条直线由原点开始向右上方移动时, z 的值逐渐增大, 一直移动目标函数的直线, 直到与约束条件确定的凸多边形相切时停止, 切点所在的位置即为最优解.

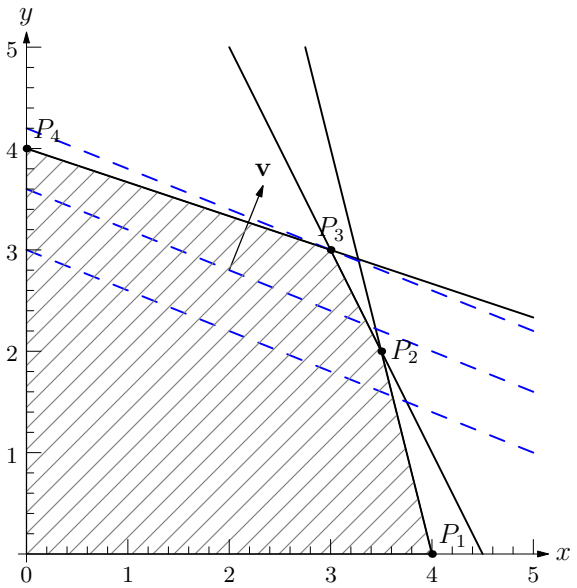


图 10.1: 图解法

在本例中, 最优解位于切点 P_3, 其坐标由直线方程 $4x_1+2x_2=18$ 与 $x_1+3x_2=12$ 确定, 即 $(x_1,x_2)=(3,3)$, 代入目标函数 $z=2x_1+5x_2=21$, 代表该工厂生产 3 件产品 I 和 3 件产品 II 时, 能获得最大利润 21 元. 不难看出, 此问题的最优解是唯一. 但是, 对于一般的线性规划问题来说, 还可能出现下列的情况:

(1) 无穷多最优解. 若将目标函数换成 $z=x_1+3x_2$, 则目标函数的直线族与约束条件(10.1c)平行, 此时线段 P_3P_4 上所有的点都是最优解, 即有无穷多最优解.

(2) 无界解. 若将约束条件(10.1a)-(10.1c)换成 $4x_1\leqslant 18$, 则可行域是无穷区域, 即变量 x_2 的取值可无限增大, 此时目标函数值也可无限增大, 即最优解无界.

(3) 无解, 或无可行解. 若将约束条件(10.1a)换成 $4x_1+2x_2\geqslant 24$, 则不存在满足所有约束条件的公共区域, 可行域是空集, 即无解.

当求解结果出现无界解或无解情况时, 一般来说是线性规划问题的数学模型有错误, 前者缺乏必要的约束条件, 后者存在矛盾的约束条件, 需重新建模.

虽然图解法仅能用于求解具有两个变量的线性规划问题, 但是它为求解一般线性规划问题提供了一些启示:

(1) 若线性规划问题的可行域存在, 则可行域是凸集.

(2) 若线性规划问题的最优解存在, 则最优解或最优解之一 (有无穷多解时) 是可行域的某个顶点.

因此, 我们可以先找出可行域的任一顶点, 计算在该顶点处的目标函数值, 检查周围相邻顶点的目标函数值是否比这个值大, 如果是否, 则它就是最优解的点或最优解的点之一; 否则, 转到比这个点的目标函数值更大的另一顶点. 重复上述过程, 直至找到使目标函数值达到最大的顶点为止, 这就是单纯形法的基本思路.

因目标函数和约束条件内容和形式上的差别, 例如目标函数有的要求 max, 有的要求 min; 决策变量一般是非负约束, 但有的问题也允许决策变量取负值, 即无约束, 故线性规划问题的表示形式是多种多样的. 因此, 为便于后续方法的描述, 有必要规定线性规划问题的**标准形式**:

$$\max\ z = \sum_{i=1}^{n} c_i x_i, \tag{10.2}$$

$$\text{s. t.} \begin{cases} \displaystyle\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = 1, 2, \cdots m, & \text{(10.3a)} \\ x_j \geqslant 0, \quad j = 1, 2, \cdots n. & \text{(10.3b)} \end{cases}$$

在标准形式中, 目标函数为求极大值, 约束条件全为等式, 约束条件右端常数项 b_i 全为非负值, 变量 x_j 均取非负值.

对于不符合标准形式的线性规划问题, 可通过下列变换进行转化:

(1) 若目标函数为求极小值, 即 $\min z = \sum_{i=1}^{n} c_i x_i$, 可令 $\tilde{z} = -z$, 则原问题可化为 $\max \tilde{z} = \Big(-\sum_{i=1}^{n} c_i x_i \Big)$, 这就与标准形式中的目标函数一致了;

(2) 若约束条件是不等式, 有两种情况: 一种是约束方程为 "$\leqslant$" 不等式, 可在不等式的左端加入一个非负变量; 另一种是约束方程为 "$\geqslant$" 不等式, 可在不等式的左端减去一个非负变量, 新加入的变量称为**松弛变量**, 则可将不等式约束化为等式约

束. 例如: $4x_1 \leqslant 16$, 可令 $x_2 = 16 - 4x_1$, 得 $4x_1 + x_2 = 16$, 显然 $x_2 \geqslant 0$, 是松弛变量; $2x_1 + 2x_2 + x_3 + x_4 + x_5 \geqslant 1000$, 可令 $x_6 = 2x_1 + 2x_2 + x_3 + x_4 + x_5 - 1000$, 得 $2x_1 + 2x_2 + x_3 + x_4 + x_5 - x_6 = 1000$, 显然 $x_6 \geqslant 0$, 是松弛变量;

(3) 若约束条件的右端项 $b_i < 0$, 可将等式两端乘以"-1", 则等式右端项必大于零;

(4) 若约束条件 $x_k \leqslant 0$, 可令 $x_{k+1} = -x_k$, 则化为约束条件 $x_{k+1} \geqslant 0$;

(5) 若存在取值无约束的变量 x_k, 可令 $x_k = x_{k+1} - x_{k+2}$, 则化为约束条件 $x_{k+1}, x_{k+2} \geqslant 0$.

不难证明, 任何形式的线性规划问题都可以化为标准形式.

例 10.3 将下列线性规划问题化为标准形式

$$\min\ z = x_1 - 4x_2 + x_3,$$

$$\text{s. t.} \begin{cases} -2x_1 - x_2 + 2x_3 \geqslant 7, \\ 3x_1 + 2x_2 - x_3 \leqslant 2, \\ x_1 + 2x_2 - x_3 = -1, \\ x_1 \leqslant 0, x_2 \geqslant 0, x_3 \text{无约束}. \end{cases}$$

解 按下列步骤进行变换:

(1) 用 $-x_1$ 代替 x_1;

(2) 用 $x_4 - x_5$ 替换 x_3, 其中 $x_4, x_5 \geqslant 0$;

(3) 在第一个约束不等式 $\geqslant$ 号的左端减去松弛变量 x_6;

(4) 在第二个约束不等式 $\leqslant$ 号的左端加上松弛变量 x_7;

(5) 在第三个约束的两端都乘以 -1;

(6) 令 $\tilde{z} = -z$, 将求 $\max z$ 化为求 $\min \tilde{z}$,

得到该线性规划问题的标准形式

$$\max\ \tilde{z} = x_1 + 4x_2 - (x_4 - x_5) + 0x_6 + 0x_7,$$

$$\text{s. t.} \begin{cases} 2x_1 - x_2 + 2(x_4 - x_5) - x_6 = 7, \\ -3x_1 + 2x_2 - (x_4 - x_5) + x_7 = 2, \\ x_1 - 2x_2 + (x_4 - x_5) = 1, \\ x_1, x_2, x_4, x_5, x_6, x_7 \geqslant 0. \end{cases}$$

□

接下来, 对于标准形式的线性规划问题, 进一步引入有关基的几个概念.

定义 10.2 设 $A=(a_{ij})_{m\times n}$ 为等式约束条件(10.3a)的系数矩阵, 其中 $m<n$, $\mathrm{rank}(A)=m$. 若 B 是 A 的一个 $m\times m$ 满秩子矩阵, 则称 B 是线性规划问题的一组**基**. 不失一般性, 可设

$$B=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix}=(\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_m),$$

其中 B 中的每一个列向量 $\mathbf{p}_j$ 称为**基向量**, 与基向量 $\mathbf{p}_j$ 对于的变量 x_j 称为**基变量**. 线性规划问题中除基变量以外的变量称为**非基变量**.

定义 10.3 在等式约束条件(10.3a)中, 若令所有非基变量 $x_{m+1}=x_{m+2}=\cdots=x_n=0$, 又因 $\det(B)\neq 0$, 知由 m 个约束方程可解出 m 个基变量的唯一解 $\mathbf{x}_B=(x_1,x_2,\cdots,x_m)^{\mathrm{T}}$. 将这个解加上非基变量取 0 的值, 有 $\mathbf{x}=(x_1,x_2,\cdots,x_m,0,\cdots,0)^{\mathrm{T}}$, 称 $\mathbf{x}$ 为线性规划问题的**基解**.

注意, 在线性规划问题求解过程中, 基是可以变化的, 此时基向量, 基变量, 基解等也随之变化. 容易看出, 在基解中变量取非零值的个数不大于方程个数 m, 故基解的总数不超过 $\binom{n}{m}$.

定义 10.4 满足非负约束条件(10.3b)的基解称为**基可行解**.

定义 10.5 与基可行解对应的基称为**可行基**.

最后, 我们讨论一般线性规划问题的几何意义.

定义 10.6 设 C 是 n 维欧式空间的一个点集, 若对任意的 $\mathbf{x}_1,\mathbf{x}_2\in C$, 均有

$$\lambda\mathbf{x}_1+(1-\lambda)\mathbf{x}_2\in C,\quad \forall\lambda\in[0,1],$$

即 $\mathbf{x}_1$ 与 $\mathbf{x}_2$ 连线上的所有点也都在 C 中, 则称 C 是**凸集**.

从直观上看, 凸集没有凹入部分, 其内部没有空洞. 实心圆, 实心球体, 实心立方体等都是凸集, 而圆环就不是凸集. 容易证明: 任何两个凸集的交集是凸集.

定义 10.7 设 $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k$ 是 n 维欧式空间中的点, 若存在实数 $0 \leqslant \lambda_1, \lambda_2, \cdots, \lambda_k \leqslant 1$, 且 $\lambda_1 + \lambda_2 + \cdots + \lambda_k = 1$, 使

$$\mathbf{x} = \lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 + \cdots + \lambda_k \mathbf{x}_k,$$

则称 $\mathbf{x}$ 是 $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k$ 的**凸组合**. 当 $0 < \lambda_i < 1\,(i = 1, 2, \cdots, k)$ 时, 称为**严格凸组合**.

定义 10.8 设 C 是 n 维欧式空间的一个凸集, 点 $\mathbf{x} \in C$, 若不存在不同的两点 $\mathbf{x}_1 \in C$ 和 $\mathbf{x}_2 \in C$ 使得

$$\mathbf{x} = \lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2, \quad 0 < \lambda < 1,$$

成立, 则称 $\mathbf{x}$ 是 C 的一个**顶点**或**极点**.

定理 10.1 若线性规划问题的可行域存在, 则它是凸集.

证明 设线性规划问题存在可行域, 即约束条件(10.3)有解, 记

$$D = \left\{ \mathbf{x} \in \mathbb{R}^n \,\middle|\, \sum_{i=1}^{n} \mathbf{p}_i x_i = \mathbf{b}, \quad x_i \geqslant 0 \right\}.$$

下证 D 是凸集. 设 $\mathbf{x}_1 = (x_{11}, x_{12}, \cdots, x_{1n})^{\mathrm{T}} \in D, \mathbf{x}_2 = (x_{21}, x_{22}, \cdots, x_{2n})^{\mathrm{T}} \in D$, 则

$$\sum_{i=1}^{n} \mathbf{p}_i x_{1i} = \mathbf{b}, \quad \sum_{i=1}^{n} \mathbf{p}_i x_{2i} = \mathbf{b}, \quad , x_{1i} \geqslant 0, \quad x_{2i} \geqslant 0.$$

对任意的 $\lambda \in [0, 1]$, 记 $\mathbf{x} = \lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2 = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$, 则

$$\sum_{i=1}^{n} \mathbf{p}_i x_i = \sum_{i=1}^{n} \mathbf{p}_i [\lambda x_{1i} + (1 - \lambda) x_{2i}] = \lambda \sum_{i=1}^{n} \mathbf{p}_i x_{1i} + (1 - \lambda) \sum_{i=1}^{n} \mathbf{p}_i x_{2i} = \mathbf{b},$$

$$x_i = \lambda x_{1i} + (1 - \lambda) x_{2i} \geqslant 0.$$

故 $\mathbf{x} \in D$. 又因为 $\mathbf{x}_1, \mathbf{x}_2$ 是任取的, 所以 D 是凸集. □

引理 10.2 线性规划问题的可行解 $\mathbf{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$ 为基可行解的充要条件是 $\mathbf{x}$ 的正分量所对应的系数列向量是线性无关的.

证明 (必要性) 由基可行解的定义可知.

(充分性) 不失一般性, 设 $\mathbf{x}$ 的正分量为 $x_1, x_2, \cdots, x_k$, 相应的系数列向量 $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_k$ 线性无关, 则必有 $k \leqslant m$. 当 $k = m$ 时, 向量 $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_k$ 恰好构成一组基, 从而

$\mathbf{x}=(x_1,x_2,\cdots,x_k,0,\cdots,0)^{\mathrm{T}}$ 是一个基可行解; 当 $k<m$ 时, 则一定可以从其余列向量中选取 $m-k$ 个与 $\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_k$ 一起构成一组基, 相应的解恰为 $\mathbf{x}$, 故也是一个基可行解. □

定理 10.3 线性规划问题的基可行解对应于可行域的顶点.

证明 命题等价于: $\mathbf{x}$ 不是基可行解 $\Longleftrightarrow$ $\mathbf{x}$ 不是可行域的顶点. 下面用反证法, 分两步证明.

(1) 若 $\mathbf{x}$ 不是基可行解, 则 $\mathbf{x}$ 一定不是可行域的顶点. 不失一般性, 设 $\mathbf{x}$ 的前 k 个分量为正, 则有

$$\sum_{i=1}^{k}\mathbf{p}_i x_i=\mathbf{b},$$

由引理10.2知, 向量 $\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_k$ 线性相关, 即存在一组不全为零的 $\mu_1,\mu_2,\cdots,\mu_k\in\mathbb{R}$, 使得

$$\sum_{i=1}^{k}\mathbf{p}_i\mu_i=\mathbf{0}.$$

对任意的 $\lambda>0$, 都有

$$\begin{aligned}(x_1-\lambda\mu_1)\mathbf{p}_1+(x_2-\lambda\mu_2)\mathbf{p}_2+\cdots+(x_k-\lambda\mu_k)\mathbf{p}_k&=\mathbf{b},\\(x_1+\lambda\mu_1)\mathbf{p}_1+(x_2+\lambda\mu_2)\mathbf{p}_2+\cdots+(x_k+\lambda\mu_k)\mathbf{p}_k&=\mathbf{b}.\end{aligned}$$

令 $\mathbf{x}_1=(x_1-\lambda\mu_1,x_2-\lambda\mu_2,\cdots,x_k-\lambda\mu_k,0,\cdots,0)^{\mathrm{T}}$, $\mathbf{x}_2=(x_1+\lambda\mu_1,x_2+\lambda\mu_2,\cdots,x_k+\lambda\mu_k,0,\cdots,0)^{\mathrm{T}}$, 则

$$\mathbf{x}=\frac{1}{2}\mathbf{x}_1+\frac{1}{2}\mathbf{x}_2.$$

另一方面, 当 λ 充分小时, 有

$$x_i\pm\lambda\mu_i\geqslant 0,\quad i=1,2,\cdots,k,$$

即 $\mathbf{x}_1,\mathbf{x}_2$ 是可行解. 因此, $\mathbf{x}$ 不是可行域的顶点.

(2) 若 $\mathbf{x}$ 不是可行域的顶点, 则 $\mathbf{x}$ 一定不是基可行解. 不失一般性, 设

$$\mathbf{x}=(x_1,x_2,\cdots,x_k,0,\cdots,0)^{\mathrm{T}},\quad x_i\geqslant 0,\quad i=1,2,\cdots,n$$

不是可行域的顶点, 故存在可行域内不同的两个点 $\mathbf{x}_1$ 和 $\mathbf{x}_2$, 使得

$$\mathbf{x} = \lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \quad \lambda \in (0,1),$$

其中 $\mathbf{x}_1 = (x_{11}, x_{12}, \cdots, x_{1n})^{\mathrm{T}}, \mathbf{x}_2 = (x_{21}, x_{22}, \cdots, x_{2n})^{\mathrm{T}}$. 将其写成分量的形式

$$x_i = \lambda x_{1i} + (1-\lambda)x_{2i}, \quad x_{1i} \geqslant 0, \quad x_{2i} \geqslant 0, \quad i = 1, 2, \cdots, n,$$

又因 $\lambda > 0, 1-\lambda > 0$, 故当 $x_i = 0$ 时, 必有 $x_{1i} = x_{2i} = 0$. 由于

$$\sum_{i=1}^{n} \mathbf{p}_i x_i = \sum_{i=1}^{k} \mathbf{p}_i x_i = \mathbf{b},$$

则有

$$\sum_{i=1}^{n} \mathbf{p}_i x_{1i} = \sum_{i=1}^{k} \mathbf{p}_i x_{1i} = \mathbf{b}, \quad \sum_{i=1}^{n} \mathbf{p}_i x_{2i} = \sum_{i=1}^{k} \mathbf{p}_i x_{2i} = \mathbf{b},$$

两式相减可得

$$\sum_{i=1}^{k} \mathbf{p}_i (x_{1i} - x_{2i}) = \mathbf{0}.$$

因系数 $\{x_{1i} - x_{2i}\}_{i=1}^{k}$ 不全为零, 故向量 $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_k$ 线性相关. 利用引理10.2, 知 $\mathbf{x}$ 不是基可行解.

□

定理 10.4 若线性规划问题的可行域有界, 则必存在一个基可行解是最优解.

证明 设线性规划问题的可行域为 D, 其顶点为 $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k$, 目标函数 $z = \sum_{i=1}^{n} c_i x_i$ 在 $\mathbf{x}^* \in D$ 取到最大值. 若 $\mathbf{x}^*$ 是 D 的一个顶点, 则命题成立. 下面, 不妨设 $\mathbf{x}^*$ 不是 D 的顶点, 又因 D 是凸集, 故存在 $0 \leqslant \lambda_1, \lambda_2, \cdots, \lambda_k \leqslant 1$, 且 $\lambda_1 + \lambda_2 + \cdots + \lambda_k = 1$, 使得

$$\mathbf{x}^* = \sum_{i=1}^{k} \lambda_i \mathbf{x}_i.$$

若记 $\mathbf{c} = (c_1, c_2, \cdots, c_n)^{\mathrm{T}}$, 则

$$\mathbf{c}^{\mathrm{T}} \mathbf{x}^* = \mathbf{c}^{\mathrm{T}} \Big(\sum_{i=1}^{k} \lambda_i \mathbf{x}_i \Big) = \sum_{i=1}^{k} \lambda_i \Big(\mathbf{c}^{\mathrm{T}} \mathbf{x}_i \Big).$$

因 D 的顶点个数是有限的, 故存在某一顶点 $\mathbf{x}_r$, 使得

$$\mathbf{c}^{\mathrm{T}}\mathbf{x}_r \geqslant \mathbf{c}^{\mathrm{T}}\mathbf{x}_i, \quad i = 1, 2, \cdots, k.$$

从而

$$\mathbf{c}^{\mathrm{T}}\mathbf{x}_r = \mathbf{c}^{\mathrm{T}}\Big(\sum_{i=1}^{k}\lambda_i\mathbf{x}_r\Big) = \sum_{i=1}^{k}\lambda_i\Big(\mathbf{c}^{\mathrm{T}}\mathbf{x}_r\Big) \geqslant \sum_{i=1}^{k}\lambda_i\Big(\mathbf{c}^{\mathrm{T}}\mathbf{x}_i\Big) = \mathbf{c}^{\mathrm{T}}\mathbf{x}^*,$$

又因 $\mathbf{c}^{\mathrm{T}}\mathbf{x}^*$ 是最大值, 故 $\mathbf{c}^{\mathrm{T}}\mathbf{x}_r = \mathbf{c}^{\mathrm{T}}\mathbf{x}^*$, 即目标函数在顶点 $\mathbf{x}_r$ 也取到最大值. □

从证明过程中可以看出, 如果目标函数在多个顶点达到最大值, 那么在这些顶点的凸组合上也达到最大值, 此时线性规划问题有无穷多最优解.

另外, 若可行域无界, 则可能无最优解, 也可能有最优解. 若有最优解, 则也必定在某顶点上取到.

因此, 若线性规划问题有最优解, 则必可在某顶点上取到. 又因可行域的顶点数是有限的, 若采用枚举法找出所有基可行解, 总数不超过 $\binom{n}{m}$, 则通过比较可找到最优解. 但是, 当 n, m 的值较大时, 该方法的时间复杂度迅速增加, 不能满足实际应用的需求. 在下一节, 我们介绍求解线性规划问题的一个经典方法: 单纯形法.

10.3 单纯形法

一般地, 在线性规划问题的等式约束条件(10.3a)中, 方程组的变量数大于方程的个数, 即 $n > m$, 方程组有无穷多解. 结合非负约束条件(10.3b), 线性规划问题的可行域常构成 $\mathbb{R}^n$ 中的多面体, 它由一系列单纯形组合而成. 单纯形法采用迭代的策略, 基本思路为: 先找出一个基可行解, 判断其是否为最优解, 若为否, 则转换到相邻的基可行解, 并使目标函数值不断增大, 一直找到最优解为止.

注解 10.2 在 n 维欧式空间中, 零维的单纯形是点, 一维的单纯形是线段, 二维的单纯形是三角形, 三维的单纯形是四面体, k 维的单纯形是有 $k+1$ 个顶点的多面体.

1. 确定初始基可行解.

对标准形式的线性规划问题, 在等式约束条件(10.3a)中, 不失一般性, 总存在一个

由单位阵构成的基, 即

$$(\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_m) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

当线性规划的约束条件均为 $\leqslant$ 时, 相应松弛变量的系数矩阵即为单位阵. 而对约束条件为 $\geqslant$ 或 $=$ 的情形, 为便于寻找初始基可行解, 可引入人工变量, 产生一个单位矩阵. 令等式约束条件(10.3a)中所有非基变量等于零, 可找到一个解

$$\mathbf{x} = (x_1, x_2, \cdots, x_m, 0, \cdots, 0)^{\mathrm{T}} = (b_1, b_2, \cdots, b_m, 0, \cdots, 0)^{\mathrm{T}}.$$

又因 $b_i \geqslant 0$, 故 $\mathbf{x}$ 满足不等式约束条件(10.3b), 是一个基可行解.

2. 从一个基可行解转换为相邻的基可行解.

设初始基可行解中的前 m 个为基变量, 即 $\mathbf{x}^0 = (x_1^0, x_2^0, \cdots, x_m^0, 0, \cdots, 0)^{\mathrm{T}}$, 其中 $x_i^0 \geqslant 0, i = 1, 2, \cdots, m$. 等式约束条件(10.3a) 可写成增广矩阵的形式:

$$\left(\begin{array}{cccccccc|c} 1 & 0 & \cdots & 0 & a_{1,m+1} & \cdots & a_{1j} & \cdots & a_{1n} & b_1 \\ 0 & 1 & \cdots & 0 & a_{2,m+1} & \cdots & a_{2j} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & a_{m,m+1} & \cdots & a_{mj} & \cdots & a_{mn} & b_m \\ \uparrow & \uparrow & & \uparrow & \uparrow & & \uparrow & & \uparrow & \uparrow \\ \mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_m & \mathbf{p}_{m+1} & \cdots & \mathbf{p}_j & \cdots & \mathbf{p}_n & \mathbf{b} \end{array}\right)$$

由基可行解的定义, 知

$$\sum_{i=1}^{m} x_i^0 \mathbf{p}_i = \mathbf{b}.$$

又因 $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_m$ 构成一组基, 故向量 $\mathbf{p}_j$ 可表示为

$$\mathbf{p}_j = \sum_{i=1}^{m} a_{ij} \mathbf{p}_i.$$

对任意的 $\lambda > 0$, 从而有

$$\sum_{i=1}^{m} x_i^0 \mathbf{p}_i + \lambda\Big(\mathbf{p}_j - \sum_{i=1}^{m} a_{ij} \mathbf{p}_i\Big) = \mathbf{b} \Longleftrightarrow \sum_{i=1}^{m} (x_i^0 - \lambda a_{ij}) \mathbf{p}_i + \lambda \mathbf{p}_j = \mathbf{b}.$$

若记 $\mathbf{x}^1=(x_1^0-\lambda a_{1j},x_2^0-\lambda a_{2j},\cdots,x_m^0-\lambda a_{mj},0,\cdots,\lambda,\cdots,0)^{\mathrm{T}}$, 显然 $\mathbf{x}^1$ 也满足等式约束条件(10.3a). 要使 $\mathbf{x}^1$ 成为一个基可行解, 则 λ 需满足

$$x_i^0-\lambda a_{ij}\geqslant 0,\quad i=1,2,\cdots,m,$$

且其中至少有一个不等式取到等号. 显然, 当 $a_{ij}\leqslant 0$ 时, 上述不等式对 $\lambda>0$ 总成立. 若取

$$\lambda=\min_i\left\{\frac{x_i^0}{a_{ij}}\,\middle|\,a_{ij}>0\right\}=\frac{x_l^0}{a_{lj}},$$

容易验证 $\mathbf{x}^1$ 是一个基可行解. 将变量 $x_1,\cdots,x_{l-1},x_j,x_{l+1},\cdots,x_m$ 对应的向量和 $\mathbf{b}$ 写成增广矩阵的形式:

$$\left(\begin{array}{cccccccc|c}
1 & 0 & \cdots & 0 & a_{1j} & 0 & \cdots & 0 & b_1\\
0 & 1 & \cdots & 0 & a_{2j} & 0 & \cdots & 0 & b_2\\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots\\
0 & 0 & \cdots & 1 & a_{l-1,j} & 0 & \cdots & 0 & b_{l-1}\\
0 & 0 & \cdots & 0 & a_{l,j} & 0 & \cdots & 0 & b_l\\
0 & 0 & \cdots & 0 & a_{l+1,j} & 1 & \cdots & 0 & b_{l+1}\\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots\\
0 & 0 & \cdots & 0 & a_{m,j} & 0 & \cdots & 1 & b_m\\
\hline
\uparrow & \uparrow & & \uparrow & \uparrow & \uparrow & & \uparrow & \uparrow\\
\mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_{l-1} & \mathbf{p}_j & \mathbf{p}_{l+1} & \cdots & \mathbf{p}_m & \mathbf{b}
\end{array}\right)$$

因 $a_{lj}>0$, 故 $\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_{l-1},\mathbf{p}_j,\mathbf{p}_{l+1},\cdots,\mathbf{p}_m$ 构成一组基. 对上述增广矩阵施行初等行变换: 第 l 行乘以 $(1/a_{lj})$, 再分别乘以 $(-a_{ij})$ 加到第 i 行上去, $i=1,\cdots,l-1,l+1,\cdots,m$, 得变换后的向量

$$\mathbf{b}=(b_1-\lambda a_{1j},\cdots,b_{l-1}-\lambda a_{l-1,j},\lambda,b_{l+1}-\lambda a_{l+1,j},\cdots,b_m-\lambda a_{mj})^{\mathrm{T}}.$$

明显地, $\mathbf{x}^1$ 是与 $\mathbf{x}^0$ 相邻的一个基可行解, 且相应基向量组成的矩阵仍是单位阵.

3. 最优性检验和解的判断.

将基可行解 $\mathbf{x}^0$ 和 $\mathbf{x}^1$ 分别代入目标函数(10.3)得

$$z^0=\sum_{i=1}^m c_i x_i^0,$$

$$z^1=\sum_{i=1}^m c_i(x_i^0-\lambda a_{ij})+\lambda c_j=z^0+\lambda\Big(c_j-\sum_{i=1}^m c_i a_{ij}\Big).$$

又因 $\lambda > 0$, 故当 $c_j - \sum_{i=1}^{m} c_i a_{ij} > 0$ 时, 就有 $z^1 > z^0$. 若记 $\sigma_j = c_j - \sum_{i=1}^{m} c_i a_{ij}$, 则 σ_j 可用作线性规划问题的解进行最优性检验的指标, 具体的判别规则如下:

(a) 当所有的 $\sigma_j \leqslant 0$ 时, 现有基可行解的目标函数值比起相邻各基可行解的目标函数值都大, 又因可行域是凸集, 故该基可行解是最优解. 进一步，若存在某个非基变量 x_k, 使得 $\sigma_k = 0$, 这意味着可以找到另外一个基可行解使目标函数值取到最大值, 从而它们之间连线上的点也都取到最大值, 故线性规划问题存在无穷多解. 反之, 当所有非基变量的 $\sigma_j < 0$ 时, 线性规划问题具有唯一的最优解;

(b) 当存在某个 $\sigma_j > 0$, 且 $\mathbf{p}_j \leqslant \mathbf{0}$, 即 $\mathbf{p}_j$ 的每个分量都小于等于零，这意味着对任意 $\lambda > 0$, 均有 $x_i^0 - \lambda a_{ij} \geqslant 0$, 而 λ 的取值可无限增大, z^1 的值也可无限增大, 故线性规划问题有无界解.

对求解结果为无可行解的判别将在后面讨论.

对标准形式的线性规划问题, 由于总可以设法使等式约束方程(10.3a)的系数矩阵包含一个单位阵 $I = (\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_m)$, 以它作为一组基, 便可求出一个初始基可行解. 为检验一个基可行解是否最优, 需要将其目标函数值与相邻基可行解的目标函数值进行比较. 为了书写规范和便于计算, 下面引入一种称为**单纯形表**的表格:

$c_j \rightarrow$			c_1	$\cdots$	c_m	$\cdots$	c_j	$\cdots$	c_n
$\mathbf{c_B}$	$\mathbf{x_B}$	$\mathbf{b}$	x_1	$\cdots$	x_m	$\cdots$	x_j	$\cdots$	x_n
c_1	x_1	b_1	1	$\cdots$	0	$\cdots$	a_{1j}	$\cdots$	a_{1n}
c_2	x_2	b_2	0	$\cdots$	0	$\cdots$	a_{2j}	$\cdots$	a_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$
c_m	x_m	b_m	0	$\cdots$	1	$\cdots$	a_{mj}	$\cdots$	a_{mn}
σ_j			0	$\cdots$	0	$\cdots$	$c_j - \sum_{i=1}^{m} c_i a_{ij}$	$\cdots$	$c_n - \sum_{i=1}^{m} c_i a_{in}$

表 10.2: 单纯形表

其中

- $\mathbf{x_B}$ 列中填入基变量, 这里是 $x_1, x_2, \cdots, x_m$;
- $\mathbf{c_B}$ 列中填入目标函数中基变量的相应系数, 这里是 $c_1, c_2, \cdots, c_m$;
- $\mathbf{b}$ 列中填入等式约束方程组右端的常数;
- x_j 列中填入等式约束中变量 x_j 对应的系数向量;
- c_j 行中填入目标函数中变量 x_j 的相应系数;
- σ_j 行中填入非基变量 x_j 的检验指标 $c_j - \sum_{i=1}^{m} c_i a_{ij}$.

在迭代计算过程中, 每找出一个新的基可行解时, 就重新建立一张单纯形表.

接下来, 给出**单纯形法**的计算步骤:

1. 根据线性规划模型的标准形式确定初始可行基和初始基可行解, 建立单纯形表;
2. 计算各非基变量 x_k 的检验指标
$$\sigma_k = c_k - \sum_{i=1}^{m} c_i a_{ik}, \quad k = m+1, m+2, \cdots, n.$$
若 $\sigma_k \leqslant 0$ 对所有 $k = m+1, m+2, \cdots, n$ 成立, 则已得到最优解, 算法终止. 否则, 进入下一步.
3. 在 $\sigma_k > 0\,(m+1 \leqslant k \leqslant n)$ 中, 若有某个 σ_j 对应的 x_j 的系数向量 $\mathbf{p}_j \leqslant \mathbf{0}$, 则该线性规划问题有无界解, 算法终止. 否则, 进入下一步.
4. 根据 $\max\{\sigma_k\} = \sigma_j$, 确定 x_j 为换入变量, 按公式
$$\lambda = \min_i \left\{ \frac{x_i^0}{a_{ij}} \,\middle|\, a_{ij} > 0 \right\} = \frac{x_l^0}{a_{lj}}$$
可确定 x_l 为换出变量;
5. 以 a_{lj} 为主元素, 对等式约束条件(10.3a)的增广矩阵施行初等行变换, 将变量 x_j 对应的列向量
$$\mathbf{p}_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{lj} \\ \vdots \\ a_{mj} \end{pmatrix} \Longrightarrow \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \leftarrow \text{第 } l \text{ 行},$$

并将 $\mathbf{x_B}$ 列中的 x_l 换为 x_j, 建立新的单纯形表.

6. 重复上述步骤 2-5, 直至算法终止.

下面, 用例10.1来演示单纯形法的计算步骤.

例 10.4 用单纯形法解线性规划问题

$$\max z = 2x_1 + 5x_2$$

$$\text{s. t.} \begin{cases} 4x_1 + 2x_2 \leqslant 18, \\ 4x_1 + x_2 \leqslant 16, \\ x_1 + 3x_2 \leqslant 12, \\ x_1, x_2 \geqslant 0. \end{cases}$$

解 先将线性规划问题转换化为标准形式:

$$\max z = 2x_1 + 5x_2 + 0x_3 + 0x_4 + 0x_5$$

$$\text{s. t.} \begin{cases} 4x_1 + 2x_2 + x_3 = 18, \\ 4x_1 + x_2 + x_4 = 16, \\ x_1 + 3x_2 + x_5 = 12, \\ x_1, x_2, x_3, x_4, x_5 \geqslant 0. \end{cases}$$

将等式约束条件写成增广矩阵形式:

$$\left(\begin{array}{ccccc|c} 4 & 2 & 1 & 0 & 0 & 18 \\ 4 & 1 & 0 & 1 & 0 & 16 \\ 1 & 3 & 0 & 0 & 1 & 12 \end{array}\right)$$

$$\begin{array}{cccccc} \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ \mathbf{p}_1 & \mathbf{p}_2 & \mathbf{p}_3 & \mathbf{p}_4 & \mathbf{p}_5 & \mathbf{b} \end{array}$$

显然 $(\mathbf{p}_3, \mathbf{p}_4, \mathbf{p}_5) = I$ 构成一组基, 相应的变量 x_3, x_4, x_5 是基变量. 令非基变量 x_1, x_2 等于零, 即可找到初始基可行解

$$\mathbf{x}^0 = (0, 0, 18, 16, 12)^{\mathrm{T}},$$

以此建立初始的单纯形表, 见表10.3.

因表10.3中有大于零的检验指标, 故表中的基可行解不是最优解. 又因 $\sigma_1 < \sigma_2$, 故确定 x_2 为换入变量. 将 $\mathbf{b}$ 除以 $\mathbf{p}_1$ 的同行系数得

$$\lambda = \min\{\frac{18}{2}, \frac{16}{1}, \frac{12}{3}\} = \frac{12}{3} = 4,$$

$c_j \rightarrow$			2	5	0	0	0
$\mathbf{c_B}$	$\mathbf{x_B}$	$\mathbf{b}$	x_1	x_2	x_3	x_4	x_5
0	x_3	18	4	2	1	0	0
0	x_4	16	4	1	0	1	0
0	x_5	12	1	[3]	0	0	1
σ_j			2	5	0	0	0

表 10.3: 初始单纯形表

由此确定 3 为主元素, 表中用 [*] 标记, 主元素所在行的基变量 x_5 为换出变量. 用变量 x_2 替换出变量 x_5, 得到一组新的基 $\mathbf{p}_3, \mathbf{p}_4, \mathbf{p}_2$, 按基可行解转换方法可以找到一个新的基可行解, 并以此建立新的单纯形表, 见表10.4.

$c_j \rightarrow$			2	5	0	0	0
$\mathbf{c_B}$	$\mathbf{x_B}$	$\mathbf{b}$	x_1	x_2	x_3	x_4	x_5
0	x_3	10	[10/3]	0	1	0	-2/3
0	x_4	12	11/3	0	0	1	-1/3
5	x_2	4	1/3	1	0	0	1/3
σ_j			1/3	0	0	0	-5/3

表 10.4: 第一次迭代的单纯形表

因表10.4中存在大于零的检验指标 $\sigma_1 > 0$, 故重复上述步骤, 得表10.5.

因表10.5中所有检验指标 $\sigma_j \leqslant 0$, 故表中的基可行解 $\mathbf{x} = (3, 3, 0, 1, 0)^{\mathrm{T}}$ 是最优解, 代入目标函数得 $z = 2x_1 + 5x_2 = 27$. □

最后, 我们讨论单纯形法计算中的几个问题

1. 初始基可行解的寻找. 在前面, 曾假设在等式约束条件(10.3a)中存在一个单位阵构成的基, 以此求初始基可行解和建立初始单纯形表十分方便. 但是, 有化为标准形式后等式约束条件的系数矩阵中不存在子单位矩阵的例子. 对于这种情形, 可以通过添加人工变量, 采用大 M 法或两阶段法来处理, 具体的做法是:

$c_j \rightarrow$			2	5	0	0	0
$\mathbf{c_B}$	$\mathbf{x_B}$	$\mathbf{b}$	x_1	x_2	x_3	x_4	x_5
2	x_1	3	1	0	3/10	0	-1/5
0	x_4	1	0	0	-11/10	1	2/5
5	x_2	3	0	1	-1/10	0	2/5
σ_j			0	0	-1/10	0	-8/5

表 10.5: 第二次迭代的单纯形表

(a) 对于含有 "$\geqslant$" 不等式约束或等式约束的情况, 若不存在子单位矩阵, 则可添加**人工变量**, 即对 "$\geqslant$" 不等式, 在不等式的左端减去一个非负的松弛变量, 再加上一个非负的人工变量; 对等式约束, 再加上一个非负的人工变量, 总能得到一个子单位矩阵;

(b) 由于约束条件在添加人工变量前已是等式, 为使它们仍得到满足, 故在最优解中人工变量取值必须为零. 为此, 可在目标函数中令人工变量的系数为任意大的负值, 用 "$-M$" 表示, 称为罚因子, 即只要人工变量取值大于零, 目标函数就不可能实现最大化, 称这种方法为**大 M 法**;

(c) 当在计算机上用大 M 法求解时, 需要用一个很大的数字来表示 M, 常取机器的最大字长. 但是, 如果线性规划问题中的 a_{ij}, b_i, c_i 等系数的值与表示 M 的值相近或远小于, 因舍入误差的影响, 数值计算结果可能会出现问题. 为克服这个困难, 可将线性规划问题分成两个阶段来计算, 称为**两阶段法**. 第一阶段先求解一个目标函数中只包含人工变量的线性规划问题, 即令目标函数中其他变量的系数为零, 人工变量的系数取某个正常数 (一般取为 1), 在保持原问题约束条件不变的情形下, 求这个目标函数极小化时的解. 显然, 在第一阶段中, 当人工变量的取值均为零时, 目标函数的值也为零, 此时, 最优解是原线性规划问题的一个基可行解; 否则, 若目标函数的极小值不为零, 那么表明原线性规划问题无可行解, 算法终止. 第二阶段是在原问题中去除人工变量, 以第一阶段的最优解为初始可行解, 继续寻找原问题的最优解.

2. 解的退化. 在按最小比值 λ 来确定换出的基变量时, 可能会出现两个或以上相同的最小比值, 从而使下一个表的基可行解中出现一个或多个基变量等于零的退化

解. 该现象出现的原因是模型中存在多余的约束, 使得多个基可行解对应于同一顶点. 当退化解出现时, 就可能出现迭代计算的无限循环, 尽管可能性极其微小. 为避免出现计算的循环, 可采用勃兰特 (Bland) 规则: (a) 当存在多个 $\sigma_j > 0$ 时, 始终选取下标最小的变量作为换入变量; (b) 当计算 λ 值出现两个或以上相同的最小值时, 始终选取下标最小的变量作为换出变量, 详细讨论可参见 [9, 10].

3. 无可行解的判别. 当线性规划问题中添加人工变量后, 初始单纯形表中的解因含人工变量, 故实质上是非可行的. 当求解结果出现所有 $\sigma_j \leqslant 0$ 时, 若基变量中仍含有非零的人工变量, 则表明该线性规划问题无可行解.

例 10.5 试用大 M 法求解线性规划问题

$$\max z = 2x_1 - x_2 - x_3$$
$$\text{s. t.} \begin{cases} x_1 - 2x_2 + x_3 \leqslant 11, \\ -4x_1 + x_2 + 2x_3 \geqslant 3, \\ -2x_1 + x_3 = 1, \\ x_1, x_2, x_3 \geqslant 0. \end{cases}$$

解 通过增加松弛变量 x_4, x_5 将原线性规划问题转换化为标准形式, 并添加人工变量 x_6, x_7 可得

$$\max z = 2x_1 - x_2 - x_3 + 0x_4 + 0x_5 - Mx_6 - Mx_7$$
$$\text{s. t.} \begin{cases} x_1 - 2x_2 + x_3 + x_4 = 11, \\ -4x_1 + x_2 + 2x_3 - x_5 + x_6 = 3, \\ -2x_1 + x_3 + x_7 = 1, \\ x_1, x_2, x_3, x_4, x_5, x_6, x_7 \geqslant 0. \end{cases}$$

以此建立初始的单纯形表, 见表10.6. 然后, 用单纯形法进行迭代计算, 见表10.7-10.8. 由于表10.8中所有检验指标 $\sigma_j \leqslant 0$ 且基变量中不含非零的人工变量, 故基可行解 $\mathbf{x} = (0, 1, 1, 10, 0, 0, 0)^{\mathrm{T}}$ 是最优解, 代入目标函数得 $z = 2x_1 - x_2 - x_3 = -2$.

□

$c_j \rightarrow$			2	-1	-1	0	0	-M	-M
$\mathbf{c_B}$	$\mathbf{x_B}$	**b**	x_1	x_2	x_3	x_4	x_5	x_6	x_7
0	x_4	11	1	-2	1	1	0	0	0
-M	x_6	3	-4	1	2	0	-5	1	0
-M	x_7	1	-2	0	[1]	0	0	0	1
σ_j			2-6M	-1+M	-1+3M	0	-5M	0	0

表 10.6: 初始单纯形表

$c_j \rightarrow$			2	-1	-1	0	0	-M	-M
$\mathbf{c_B}$	$\mathbf{x_B}$	**b**	x_1	x_2	x_3	x_4	x_5	x_6	x_7
0	x_4	10	3	0	0	1	0	2	-1
-M	x_6	1	0	[1]	0	0	-5	1	-2
-1	x_3	1	-2	0	1	0	0	0	1
σ_j			0	-1+M	0	0	-5M	0	1-3M

表 10.7: 第一次迭代的单纯形表

$c_j \rightarrow$			2	-1	-1	0	0	-M	-M
$\mathbf{c_B}$	$\mathbf{x_B}$	**b**	x_1	x_2	x_3	x_4	x_5	x_6	x_7
0	x_4	10	3	0	0	1	0	2	-1
-1	x_2	1	0	1	0	0	-5	1	-2
-1	x_3	1	-2	0	1	0	0	0	1
σ_j			0	0	0	0	-5	1-M	-1-M

表 10.8: 第二次迭代的单纯形表

例 10.6 试用两阶段法求解线性规划问题

$$\max z = 2x_1 - x_2 - x_3$$

$$\text{s. t.} \begin{cases} x_1 - 2x_2 + x_3 \leqslant 11, \\ -4x_1 + x_2 + 2x_3 \geqslant 3, \\ -2x_1 + x_3 = 1, \\ x_1, x_2, x_3 \geqslant 0. \end{cases}$$

解 通过增加松弛变量 x_4, x_5 将原线性规划问题转换化为标准形式, 并添加人工变量 x_6, x_7 可得第一阶段的线性规划问题

$$\min z = x_6 + x_7 \Longleftrightarrow \max \tilde{z} = -x_6 - x_7$$

$$\text{s. t.} \begin{cases} x_1 - 2x_2 + x_3 + x_4 = 11, \\ -4x_1 + x_2 + 2x_3 - x_5 + x_6 = 3, \\ -2x_1 + x_3 + x_7 = 1, \\ x_1, x_2, x_3, x_4, x_5, x_6, x_7 \geqslant 0. \end{cases}$$

用单纯形法求解, 见表10.9-10.11. 容易看出, 第一阶段的最优解是 $\mathbf{x} = (0, 1, 1, 10, 0, 0, 0)^{\mathrm{T}}$, 目标函数值 $z = x_6 + x_7 = 0$. 因人工变量 $x_6 = x_7 = 0$, 故 $\mathbf{x} = (0, 1, 1, 10, 0, 0, 0)^{\mathrm{T}}$ 是原线性规划问题的基可行解. 将第一阶段最终表中的人工变量 x_6, x_7 所在的列删除, 并在变量 x_1, x_2, x_3 的 c_j 处填入原问题目标函数的系数, 构建第二阶段的初始单纯形表, 见表10.12. 由于所有检验指标 $\sigma_j \leqslant 0$, 故 $\mathbf{x} = ((0, 1, 1, 10, 0)^{\mathrm{T}}$ 是原线性规划问题的最优解, 代入目标函数得 $z = 2x_1 - x_2 - x_3 = -2$.

$c_j \rightarrow$			0	0	0	0	0	-1	-1
$\mathbf{c_B}$	$\mathbf{x_B}$	$\mathbf{b}$	x_1	x_2	x_3	x_4	x_5	x_6	x_7
0	x_4	11	1	-2	1	1	0	0	0
-1	x_6	3	-4	1	2	0	-5	1	0
-1	x_7	1	-2	0	[1]	0	0	0	1
σ_j			-6	1	3	0	-5	0	0

表 10.9: 第一阶段的初始单纯形表

□

$c_j \rightarrow$			0	0	0	0	0	-1	-1
$\mathbf{c_B}$	$\mathbf{x_B}$	**b**	x_1	x_2	x_3	x_4	x_5	x_6	x_7
0	x_4	10	3	0	0	1	0	2	-1
-1	x_6	1	0	[1]	0	0	-5	1	-2
0	x_3	1	-2	0	1	0	0	0	1
σ_j			0	1	0	0	-5	0	-3

表 10.10: 第一阶段第一次迭代的单纯形表

$c_j \rightarrow$			0	0	0	0	0	-1	-1
$\mathbf{c_B}$	$\mathbf{x_B}$	**b**	x_1	x_2	x_3	x_4	x_5	x_6	x_7
0	x_4	10	3	0	0	1	0	2	-1
0	x_2	1	0	1	0	0	-5	1	-2
0	x_3	1	-2	0	1	0	0	0	1
σ_j			0	0	0	0	0	0	0

表 10.11: 第一阶段第二次迭代的单纯形表

$c_j \rightarrow$			2	-1	-1	0	0
$\mathbf{c_B}$	$\mathbf{x_B}$	**b**	x_1	x_2	x_3	x_4	x_5
0	x_4	10	3	0	0	1	0
-1	x_2	1	0	1	0	0	-5
-1	x_3	1	-2	0	1	0	0
σ_j			0	0	0	0	-5

表 10.12: 第二阶段的初始单纯形表

在实际应用中, 尽管单纯形法非常有效, 但在理论上, 其算法时间复杂度仍是指数级别的. 一个自然的问题是: 能否找到算法时间复杂度更低的算法? 在此驱动下, 产生了**内点法**，其算法时间复杂度是多项式级别的, 是求解线性规划问题的另一有效方法, 可参见 [4].

10.4 非线性优化问题

在线性规划问题中, 目标函数和约束条件都是关于决策变量的线性函数, 而如果最优化问题中的目标函数或约束条件中有关于决策变量的非线性函数, 则称它是**非线性优化**问题或**非线性规划**问题.

例 10.7 在半径为 R 的圆内有一内接三角形 $\triangle ABC$, 其顶点 A, B, C 可以在圆上移动, 求 $\triangle ABC$ 面积的最大值.

解 设圆心为 O, 中心角 $\angle AOB$ 为 θ_1, $\angle BOC$ 为 θ_2, $\angle COA$ 为 θ_3, 则 $\triangle ABC$ 面积为

$$z = \frac{1}{2}R^2\big(\sin\theta_1 + \sin\theta_2 + \sin\theta_3\big),$$

其中 $\theta_1, \theta_2, \theta_3$ 满足约束条件

$$\begin{aligned} &\theta_1 + \theta_2 + \theta_3 = 2\pi, \\ &\theta_1, \theta_2, \theta_3 \geqslant 0. \end{aligned}$$

上述数学模型可写成非线性最优化问题的形式

$$\max\ z = \frac{1}{2}R^2\big(\sin\theta_1 + \sin\theta_2 + \sin\theta_3\big),$$
$$\text{s. t.} \begin{cases} \theta_1 + \theta_2 + \theta_3 = 2\pi \\ \theta_1, \theta_2, \theta_3 \geqslant 0. \end{cases}$$

利用数学分析的知识, 当 $\theta_1 = \theta_2 = \theta_3 = \dfrac{2\pi}{3}$ 时, 即等边三角形, $\triangle ABC$ 面积取到最大值 $\dfrac{3\sqrt{3}}{4}R^2$. □

例 10.8 现有一块椭球形的石材, 其半轴分别为 $a, b, c > 0$. 若要从中加工出一块长方体的石砖, 要求长方体各边平行于椭球的半轴且浪费最少, 问该如何设计切割方案?

解 设以椭球面的球心为坐标原点 O, 三条半轴为 x, y, z 轴, 建立空间直角坐标系, 椭球面的方程可写成

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=1.$$

显然, 当长方体的顶点都在椭球面上时, 石材浪费的较少. 记长方体在第一卦限内的顶点为 (x, y, z), 则长方体的体积 $z=8xyz$, 写成非线性最优化问题的形式

$$\begin{aligned}&\max\ z=8xyz,\\ &\text{s. t.}\begin{cases}\dfrac{x^2}{a^2}+\dfrac{y^2}{b^2}+\dfrac{z^2}{c^2}=1\\ x,y,z\geqslant 0.\end{cases}\end{aligned}$$

利用数学分析的知识, 当 $x=\dfrac{a}{\sqrt{3}}, y=\dfrac{b}{\sqrt{3}}, z=\dfrac{c}{\sqrt{3}}$ 时, 长方体的体积取到最大值 $\dfrac{8\sqrt{3}abc}{9}$. □

以上例子中, 目标函数或约束条件中的函数是非线性函数, 但是形式比较简单, 利用数学分析中多元函数条件极值的知识, 可以进行求解. 然而, 在实际应用中, 目标函数或约束条件中的函数可能非常复杂, 甚至写不出解析表达式; 决策变量个数很多, 譬如几百个甚至上万个, 这类复杂的最优化问题是不能显式得求解出来的. 因此, 有必要研究非线性优化问题的数值求解算法.

与线性规划类似, 因 $\max f(\mathbf{x})=-\min[-f(\mathbf{x})]$, 故可仅讨论最小化问题. 不失一般性, **非线性优化问题**可写成:

$$\begin{aligned}&\min\ f(\mathbf{x}),\quad \mathbf{x}\in\mathbb{R}^n\\ &\text{s. t.}\begin{cases}c_i(\mathbf{x})\geqslant 0, & i\in\mathcal{I}\\ c_i(\mathbf{x})=0, & i\in\mathcal{E}.\end{cases}\end{aligned}$$

其中 $f(\mathbf{x}), c_i(\mathbf{x})$ 是有 n 个自变量的连续实函数, $\mathcal{E},\mathcal{I}$ 分别是等式约束和不等式约束的有限指标集. 一般而言, 由于非线性函数的复杂性, 求解非线性规划问题要比线性规划问题困难的多. 而且, 也不像线性规划有单纯形法等通用方法, 非线性规划目前还没有普适的求解方法, 现有的算法都有特定的适用范围.

注解 10.3 如果非线性优化问题中的目标函数和约束条件中的函数均为凸函数, 则称它为**凸优化**问题. 由于凸函数具有许多好的性质, 凸优化问题存在更为有效的解法, 可参见专著 [3].

注解 10.4 如果目标函数的自变量为离散变量, 譬如整数或有限集合中的元素, 则称它为**离散优化**. 整数规划问题就是典型的离散优化问题. 另外, 旅行商问题、图着色问题、最小生成树等经典的组合优化问题, 也是离散优化问题. 与连续优化问题相比, 离散优化问题的求解更为困难, 算法的时间复杂度高, 常采用近似求解算法, 可参见专著 [2, 8].

在介绍具体求解算法之前, 我们先回顾数学分析中与函数极值有关的一些概念及定理.

定义 10.9 设函数 $f: D \to \mathbb{R}$，其中区域 $D \subset \mathbb{R}^n$. 如果对于 D 的内点 $\mathbf{x}^*$, 存在领域 $U(\mathbf{x}^*) \subset D$, 当 $\mathbf{x} \in U(\mathbf{x}^*)$ 时, 有 $f(\mathbf{x}) \geqslant f(\mathbf{x}^*)\,(f(\mathbf{x}) \leqslant f(\mathbf{x}^*))$, 则称函数 f 在 D 的内点 $\mathbf{x}^*$ 取**局部极小值 (局部极大值)**, 简称**极小值 (极大值)**. 如果对 $\mathbf{x} \in U(\mathbf{x}^*)\backslash\{\mathbf{x}^*\}$ 有严格不等式 $f(\mathbf{x}) > f(\mathbf{x}^*)\,(f(\mathbf{x}) < f(\mathbf{x}^*))$, 则称函数 f 在 $\mathbf{x}^*$ 取**严格局部极小值 (严格局部极大值)**.

定义 10.10 函数的局部极大值与局部极小值统称为函数的**局部极值**, 简称**极值**, 函数取极值的点称为**极值点**.

下面的定理给出了判别函数极值点的必要条件.

定理 10.5 设函数 $f: D \to \mathbb{R}$ 在内点 $\mathbf{x}^*$ 取极值且 $Jf(\mathbf{x}^*)$ 存在, 则 $Jf(\mathbf{x}^*) = \mathbf{0}$, 即

$$\frac{\partial f}{\partial x_i}(\mathbf{x}^*) = 0, \quad i = 1, 2, \cdots, n.$$

定义 10.11 设函数 $f: D \to \mathbb{R}$, 在 D 中使得 $Jf(\mathbf{x}) = \mathbf{0}$ 的一切内点称为函数 f 的**驻点**.

显然, 极值点一定是驻点, 但一般来说驻点未必是极值点. 若记

$$Jf(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \cdots, \frac{\partial f(\mathbf{x})}{\partial x_n}\right), \quad Hf(\mathbf{x}) = \begin{pmatrix} \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{pmatrix},$$

分别称 $Jf(\mathbf{x})$ 和 $Hf(\mathbf{x})$ 为函数 f 的 **Jacobian 矩阵**和 **Hessian 矩阵**. 利用 Taylor 展开公式与二次型的性质, 容易证明以下的定理.

定理 10.6 设函数 $f(\mathbf{x})$ 在 D 上具有连续的二阶偏导数, 即 $f(\mathbf{x}) \in C^2(D)$.

(1) (必要性) 若 $f(\mathbf{x})$ 在 $\mathbf{x}^*$ 取局部极小 (大) 值, 则 $Jf(\mathbf{x}^*)=\mathbf{0}$ 且 $Hf(\mathbf{x}^*)$ 半正 (负) 定;

(2) (充分性) 若 $Jf(\mathbf{x}^*)=\mathbf{0}$ 且 $Hf(\mathbf{x}^*)$ 正 (负) 定，则 $f(\mathbf{x})$ 在 $\mathbf{x}^*$ 取得严格局部极小 (大) 值.

上述定理给出了判别函数内部局部极值点的二阶必要条件和充分条件.

定义 10.12 设 $\mathbf{x}^*\in D$, 若对所有 $\mathbf{x}\in D$, 都有 $f(x)\geqslant f(\mathbf{x}^*)\,(f(x)\leqslant f(\mathbf{x}^*))$, 则称 $\mathbf{x}^*$ 为 $f(\mathbf{x})$ 在 D 上的**全局极小值点 (全局极大值点)**. 若对所有 $\mathbf{x}\in D\backslash\mathbf{x}^*$, 都有 $f(x)>f(\mathbf{x}^*)\,(f(x)<f(\mathbf{x}^*))$, 则称 $\mathbf{x}^*$ 为 $f(\mathbf{x})$ 在 D 上的**全局严格极小值点 (全局严格极大值点)**

需要指出的是, 全局极值点可能在定义域的边界点上取到. 然而, 对于这些边界上的极值点, 并没有实用的判别条件. 因此, 寻找全局极值点是一个相当困难的任务. 幸好, 在许多实际问题中, 局部极值点就已能满足应用的需求, 故非线性优化主要研究寻找局部极值点.

注解 10.5 对于定义在凸集上的凸函数 $f(\mathbf{x})$, 可以证明: $f(\mathbf{x})$ 的局部极值点就是全局极值点, 且所有极值点构成一个凸集. 进一步, 若 $f(\mathbf{x})$ 是严格凸函数, 则 $f(\mathbf{x})$ 的全局极值点是唯一的.

根据前面的叙述, 一个自然的想法是: 1) 令 $\nabla f(\mathbf{x})=\mathbf{0}$, 求出 $f(\mathbf{x})$ 的所有驻点; 2) 对每一驻点, 利用局部极值的充分条件进行判别, 找出所要的解. 对某些较简单的函数, 这样做是可行的, 但是, 对于一般的 n 元函数, 由条件 $\nabla f(\mathbf{x})=\mathbf{0}$ 得到的方程通常是一个非线性方程组, 求解非常困难. 当函数 $f(\mathbf{x})$ 不可微分时, 就更不能用上述方法了. 因此, 求解非线性优化问题常使用迭代法.

迭代法的基本思路是: 为求函数 $f(\mathbf{x})$ 的最优解, 首先给定一个初始估计 $\mathbf{x}_0$, 然后按某种规则找出比 $\mathbf{x}_0$ 更好的一个解 $\mathbf{x}_1$, 即 $f(\mathbf{x}_1)<f(\mathbf{x}_0)$, 重复上述过程, 得到一个解序列 $\{\mathbf{x}_k\}_{k=0}^{\infty}$. 若这个解序列有极限 $\mathbf{x}^*$, 即

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

则称它收敛于 $\mathbf{x}^*$. 迭代法的关键是设计有效的规则, 使得所产生的解序列 $\{\mathbf{x}_k\}_{k=0}^{\infty}$ 收敛于该问题的最优解之一. 因计算机只能进行有限次迭代, 一般不能得到问题的精确解, 故当满足所要求的精度时, 可终止迭代, 得到一个较好的近似解.

若由某种算法所产生的解序列 $\{\mathbf{x}_k\}_{k=0}^{\infty}$ 使目标函数 $f(\mathbf{x}_k)$ 逐步减少, 则称它为**下降算法**. 现假定已迭代到点 $\mathbf{x}_k$, 若从该点出发沿任何方向移动都不能使目标函数 $f(\mathbf{x})$

值下降, 则 $\mathbf{x}_k$ 是一个局部极小点, 迭代停止; 否则, 从该点出发至少存在一个方向使目标函数值有所下降, 则可选择使目标函数值下降的某一方向 $\mathbf{p}_k$ 作为搜索方向, 选择一个合适的步长 λ_k, 得到一个新的迭代点

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{p}_k,$$

其中称 $\mathbf{p}_k$ 为**搜索方向**, λ_k 为**步长**或**步长因子**.

下降迭代算法的计算步骤如下:

1. 选定某一初始点 $\mathbf{x}_0$, 并令 $k \leftarrow 0$;
2. 确定搜索方向 $\mathbf{p}_k$;
3. 从 $\mathbf{x}_k$ 出发, 沿搜索方向 $\mathbf{p}_k$ 求步长 λ_k, 以产生下一个迭代点 $\mathbf{x}_{k+1}$;
4. 检查新点 $\mathbf{x}_{k+1}$ 是否为极小点或近似极小点. 若是, 则算法终止; 否则, 令 $k \leftarrow k+1$, 重复步骤 2-4.

在上述步骤中, 选取搜索方向 $\mathbf{p}_k$ 是最关键的一步, 各种算法的主要差别之一在于确定搜索方向的方法不同. 确定步长 λ_k 亦可采用不同的方法. 最简单的一种是令它等于某一常数, 譬如 $\lambda_k = 1$, 这样做计算简便, 但不能保证使目标函数值下降; 第二种称为可接收点算法, 只要能使目标函数值下降, 可任意选择步长 λ_k; 第三种方法是沿着搜索方向使目标函数值下降最多, 即求解一个子优化问题

$$\min_{\lambda \in \mathbb{R}} f(\mathbf{x}_k + \lambda \mathbf{p}_k),$$

称这一过程为**最优一维搜索**或**精确一维搜索**, 所确定的步长为**最佳步长**. 最优一维搜索有一个重要的性质: 在搜索方向上所得到的最优点处, 目标函数的梯度和该搜索方向正交.

定理 10.7 设目标函数 $f(\mathbf{x})$ 具有一阶连续偏导数, $\mathbf{x}_{k+1}$ 按如下规则产生:

$$\begin{cases} \lambda_k = \arg\min\limits_{\lambda \in \mathbb{R}} f(\mathbf{x}_k + \lambda \mathbf{p}_k), \\ \mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{p}_k, \end{cases}$$

则有

$$\nabla f(\mathbf{x}_{k+1})^{\mathrm{T}} \mathbf{p}_k = 0.$$

证明 令 $\varphi(\lambda)=f(\mathbf{x}_k+\lambda\mathbf{p}_k)$, 因 $f(\mathbf{x})$ 具有一阶连续偏导数, 故 $\varphi(\lambda)$ 是关于变量 λ 的连续函数, 且有一阶连续的导数. 利用复合求导法则, 可得

$$\varphi'(\lambda)=\nabla f(\mathbf{x}_k+\lambda\mathbf{p}_k)^{\mathrm{T}}\mathbf{p}_k.$$

又因 λ_k 是 $\varphi(\lambda)$ 的局部极小值点, 故

$$\varphi'(\lambda)\Big|_{\lambda=\lambda_k}=\nabla f(\mathbf{x}_{k+1})^{\mathrm{T}}\mathbf{p}_k=0.$$

□

在设计算法时, 不仅要求它产生的解序列 $\{\mathbf{x}_k\}_{k=0}^{\infty}$ 能收敛到问题的最优解, 而且还希望它具有较快的收敛速度.

定义 10.13 设序列 $\{\mathbf{x}_k\}_{k=0}^{\infty}$ 收敛于 $\mathbf{x}^*$, 若存在常数 $c>0$, $\alpha\geqslant 1$ 及整数 $k_0>0$, 使得当 $k>k_0$ 时, 均有

$$\|\mathbf{x}_{k+1}-\mathbf{x}^*\|\leqslant c\|\mathbf{x}_k-\mathbf{x}^*\|^{\alpha}$$

成立, 则称 $\{\mathbf{x}_k\}_{k=0}^{\infty}$ 的**收敛阶**为 α, 或 $\{\mathbf{x}_k\}_{k=0}^{\infty}$ 是α **阶收敛**. 当 $\alpha=2$ 时, 称为**二阶收敛**; 当 $1<\alpha<2$ 时, 称为**超线性收敛**; 当 $\alpha=1$ 时, 称为**线性收敛**或**一阶收敛**.

通常, 线性收敛速度是比较慢的, 二阶收敛是很快的, 超线性收敛介于它们之间. 若一个算法具有超线性或更高的收敛速度, 一般就认为它是一个很好的算法了.

最后, 讨论一下迭代算法的收敛准则. 常用的有以下几种:

(1) 依据相邻两次迭代的绝对误差

$$\|\mathbf{x}_{k+1}-\mathbf{x}_k\|<\varepsilon_1,\quad |f(\mathbf{x}_{k+1})-f(\mathbf{x}_k)|<\varepsilon_2;$$

(2) 依据相邻两次迭代的相对误差

$$\frac{\|\mathbf{x}_{k+1}-\mathbf{x}_k\|}{\|\mathbf{x}_k\|}<\varepsilon_3,\quad \frac{|f(\mathbf{x}_{k+1})-f(\mathbf{x}_k)|}{|f(\mathbf{x}_k)|}<\varepsilon_4;$$

(3) 依据目标函数梯度的模

$$\|\nabla f(\mathbf{x}_k)\|<\varepsilon_5,$$

其中 $\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4,\varepsilon_5$ 为事先设定的精度参数. 此外, 为了防止程序陷入无限循环, 还可设置最大迭代次数 N.

10.5 一维搜索

所谓**一维搜索**, 又称**线性搜索**, 是指单变量函数的最优化, 它是多变量函数优化的基础. 在用迭代法求函数的极小值时, 经常用到一维搜索, 即沿着某一已知方向求目标函数的极小值点. 另外, 如果优化问题中的目标函数的导数或偏导数不存在或难以计算时, 也常采用搜索的方法求解. 一维搜索的方法很多, 常用的有

(1) 试探法, 譬如成功 - 失败法, 黄金分割法, 斐波那契 (Fibonacci) 法等;

(2) 插值法, 譬如抛物线法, 三次插值法等;

(3) 数学分析中的求根法, 譬如切线法, 二分法等.

一维搜索的主要步骤: 首先确定包含问题最优解的搜索区间; 然后采用某种分割技术或插值方法缩小这个区间, 进行搜索求解.

定义 10.14 设 $\varphi : \mathbb{R} \mapsto \mathbb{R}$ 是一个连续函数, $\lambda^* \in [0, +\infty)$, 且有

$$\varphi(\lambda^*) = \min_{\lambda \geqslant 0} \varphi(\lambda).$$

若存在闭区间 $[a, b] \subset [0, +\infty)$, 使 $\lambda^* \in [a, b]$, 则称 $[a, b]$ 是一维优化问题 $\min\limits_{\lambda \geqslant 0} \varphi(\lambda)$ 的**搜索区间**.

确定搜索区间的一种简单方法叫进退法, 它的基本思想为: 从定义域内某一点出发, 按一定步长, 试图确定某一区间, 使得函数值呈现 “高 - 低 - 高” 的形状. 一个方向不成功, 就退回来, 再沿相反方向进行搜索. 进退法的计算步骤如下:

1. 设置初始值. $\lambda_0 \in [0, +\infty)$, $h_0 > 0$, 加倍系数 $t > 1$, 常取 $t = 2$, 计算 $\varphi(\lambda_0)$, $k \leftarrow 0$;
2. 比较目标函数值. 令 $\lambda_{k+1} \leftarrow \lambda_k + h_k$, 计算 $\varphi(\lambda_{k+1})$. 若 $\varphi(\lambda_{k+1}) < \varphi(\lambda_k)$, 进入下一步; 否则, 转步骤 4;
3. 更新搜索步长. 令 $h_{k+1} \leftarrow t \times h_k$, $\lambda \leftarrow \lambda_k$, $\lambda_k \leftarrow \lambda_{k+1}$, $k \leftarrow k + 1$, 转步骤 2;
4. 反向搜索. 若 $k = 0$, 转换搜索方向, 令 $h_k \leftarrow (-h_k)$, $\lambda_k \leftarrow \lambda_{k+1}$, 转步骤 2; 否则, 停止迭代, 并令

$$a = \min\{\lambda, \lambda_{k+1}\}, \quad b = \max\{\lambda, \lambda_{k+1}\},$$

输出 $[a, b]$.

一维搜索方法主要针对单峰函数在单峰区间上的优化问题.

定义 10.15 设 $\varphi:\mathbb{R}\mapsto\mathbb{R}$ 是一个连续函数, 若存在 $[a,b]$ 及 $\lambda^*\in[a,b]$, 使 $\varphi(\lambda)$ 在 $[a,\lambda^*]$ 上严格递减, 在 $[\lambda^*,b]$ 上严格递增, 则称 $[a,b]$ 是函数 $\varphi(\lambda)$ 的**单峰区间**, $\varphi(\lambda)$ 是区间 $[a,b]$ 上的**单峰函数**.

引理 10.8 设 $\varphi:\mathbb{R}\mapsto\mathbb{R}$ 是一个连续函数, $[a,b]$ 是 $\varphi(\lambda)$ 的单峰区间, $\alpha_1,\alpha_2\in[a,b]$, 且 $\alpha_1<\alpha_2$.

(1) 若 $\varphi(\alpha_1)\leqslant\varphi(\alpha_2)$, 则 $[a,\alpha_2]$ 是 $\varphi(\lambda)$ 的单峰区间;

(2) 若 $\varphi(\alpha_1)\geqslant\varphi(\alpha_2)$, 则 $[\alpha_1,b]$ 是 $\varphi(\lambda)$ 的单峰区间.

证明 根据单峰函数和单峰区间的定义, 存在 $\lambda^*\in[a,b]$, 使 $\varphi(\lambda)$ 在 $[a,\lambda^*]$ 上严格递减, 在 $[\lambda^*,b]$ 上严格递增. 因 $\varphi(\alpha_1)\leqslant\varphi(\alpha_2)$, 故 $\lambda^*\in[a,\alpha_2]$. 又因 $\varphi(\lambda)$ 是 $[a,b]$ 上的单峰函数, 故 $\varphi(\lambda)$ 也是 $[a,\alpha_2]$ 上的单峰函数, 即 $[a,\alpha_2]$ 是 $\varphi(\lambda)$ 的单峰区间. 类似地可证明 $\varphi(\alpha_1)\geqslant\varphi(\alpha_2)$ 的情形. □

利用上述引理知, 如果 $\varphi(\lambda)$ 是 $[a,b]$ 上的单峰函数, 则可通过比较 $\varphi(\lambda)$ 的函数值, 来缩小搜索区间. 设 $\varphi(\lambda)=f(\mathbf{x}_k+\lambda\mathbf{p}_k)$, $\varphi(\lambda)$ 是搜索区间 $[a,b]$ 上的单峰函数. 记第 k 次迭代所产生的搜索区间为 $[a_k,b_k]$, 取两个试探点 $\alpha_k,\beta_k\in[a_k,b_k]$, 且 $\alpha_k<\beta_k$, 计算 $\varphi(\alpha_k)$ 和 $\varphi(\beta_k)$, 根据引理10.8, 知

(1) 若 $\varphi(\alpha_k)\leqslant\varphi(\beta_k)$, 则令 $a_{k+1}\leftarrow a_k$, $b_{k+1}\leftarrow\beta_k$;

(2) 若 $\varphi(\alpha_k)>\varphi(\beta_k)$, 则令 $a_{k+1}\leftarrow\alpha_k$, $b_{k+1}\leftarrow b_k$.

反复执行上述步骤, 搜索区间逐渐变小, 当达到所需的精度时, 算法终止. 显然, 剩下的问题是如何确定试探点 α_k,β_k 的位置.

下面, 我们介绍几种选取的方法. 一种常用的取法是**黄金分割法**, 它要求试探点 α_k,β_k 满足以下条件:

(1) α_k 和 β_k 到搜索区间 $[a_k,b_k]$ 的端点距离相等, 即

$$b_k-\alpha_k=\beta_k-a_k; \tag{10.4}$$

(2) 每次迭代, 搜索区间长度的缩短率相等, 即

$$b_{k+1}-a_{k+1}=\tau(b_k-a_k). \tag{10.5}$$

将式 (10.4) 与式 (10.5) 联立, 可得

$$\alpha_k = a_k + (1-\tau)(b_k - a_k), \tag{10.6}$$

$$\beta_k = a_k + \tau(b_k - a_k). \tag{10.7}$$

现考虑 $\varphi(\alpha_k) \leqslant \varphi(\beta_k)$ 的情形, 则新的搜索区间为 $[a_{k+1}, b_{k+1}] = [a_k, \beta_k]$. 为进一步缩短搜索区间, 需取新的试探点 $\alpha_{k+1}, \beta_{k+1}$. 由式 (10.7) 知,

$$\begin{aligned}\beta_{k+1} &= a_{k+1} + \tau(b_{k+1} - a_{k+1}) \\ &= a_k + \tau(\beta_k - a_k) \\ &= a_k + \tau[a_k + \tau(b_k - a_k) - a_k] \\ &= a_k + \tau^2(b_k - a_k).\end{aligned}$$

若令 $\tau^2 = 1 - \tau$, 则

$$\beta_{k+1} = a_k + (1-\tau)(b_k - a_k) = \alpha_k.$$

此时, 新的试探点 β_{k+1} 不需要重新计算, 只要取 α_k 即可. 从而在每次迭代中, 仅需取一个新的试探点即可. 类似地, 考虑 $\varphi(\alpha_k) > \varphi(\beta_k)$ 的情形, 由式 (10.6) 知, 新的试探点 $\alpha_{k+1} = \beta_k$, 它也不需要重新计算.

由于 $\tau^2 = 1 - \tau$, 且 $\tau > 0$, 解得

$$\tau = \frac{-1 + \sqrt{5}}{2} \approx 0.618.$$

因 0.618 等于黄金分割率, 故上述方法常称为黄金分割法或 0.618 法. 试探点 α_k 与 β_k 的取法为

$$\alpha_k = a_k + 0.382(b_k - a_k),$$

$$\beta_k = a_k + 0.618(b_k - a_k).$$

黄金分割法的计算步骤如下:

1. 选取初始数据. 确定初始搜索区间 $[a_0, b_0] = [a, b]$ 和精度要求 $\delta > 0$. 计算初始试探点 α_0, β_0, 并令 $k \leftarrow 0$;
2. 比较函数值. 若 $\varphi(\alpha_k) \leqslant \varphi(\beta_k)$, 进入下一步; 否则, 转步骤 4;
3. 若 $\beta_k - a_k \leqslant \delta$, 则停止计算, 输出 α_k; 否则, 令 $a_{k+1} \leftarrow a_k, b_{k+1} \leftarrow \beta_k, \beta_{k+1} \leftarrow \alpha_k$, $\varphi(\beta_{k+1}) \leftarrow \varphi(\alpha_k)$, $\alpha_{k+1} \leftarrow a_{k+1} + 0.382(b_{k+1} - a_{k+1})$, 计算 $\varphi(\alpha_{k+1})$, 转步骤 5;

4. 若 $b_k-\alpha_k\leqslant\delta$, 则停止计算, 输出 β_k; 否则, 令 $a_{k+1}\leftarrow\alpha_k, b_{k+1}\leftarrow b_k, \alpha_{k+1}\leftarrow\beta_k$, $\varphi(\alpha_{k+1})\leftarrow\varphi(\beta_k), \beta_{k+1}\leftarrow a_{k+1}+0.618(b_{k+1}-a_{k+1})$, 计算 $\varphi(\beta_{k+1})$, 进入下一步;
5. 令 $k\leftarrow k+1$, 转步骤 2.

容易看出, 每次迭代后, 搜索区间的缩短率为 τ. 若搜索初始区间为 $[a,b]$, 则经过 n 迭代后, 搜索区间的长度为 $\tau^n(b-a)$, 算法是收敛的, 且收敛速度是线性的.

另一种常用的取法是**斐波那契法**, 它与黄金分割法的主要区别: 搜索区间长度的缩短率不是采用黄金分割数, 而是采用斐波那契数列 $\{F_k\}_{k=0}^{\infty}$, 即

$$\begin{aligned}F_0&=F_1=1,\\F_{k+1}&=F_k+F_{k-1},\quad k=1,2,\cdots.\end{aligned}$$

在该方法中, 试探点 α_k, β_k 的取法为

$$\begin{aligned}\alpha_k&=a_k+\Big(1-\frac{F_{n-k}}{F_{n-k+1}}\Big)(b_k-a_k),\\\beta_k&=a_k+\frac{F_{n-k}}{F_{n-k+1}}(b_k-a_k).\end{aligned}$$

显然, 上述公式中的 $\dfrac{F_{n-k}}{F_{n-k+1}}$ 相当于黄金分割法的 τ, 每次缩短率满足

$$b_{k+1}-a_{k+1}=\frac{F_{n-k}}{F_{n-k+1}}(b_k-a_k),$$

其中 n 是计算函数值的次数. 若要求经过 n 次迭代后, 所得的区间长度不超过 δ, 即

$$b_n-a_n\leqslant\delta,$$

那么

$$\begin{aligned}b_n-a_n&=\frac{F_1}{F_2}(b_{n-1}-a_{n-1})\\&=\frac{F_1}{F_2}\cdot\frac{F_2}{F_3}\cdots\frac{F_{n-1}}{F_n}(b_1-a_1)\\&=\frac{1}{F_n}(b_1-a_1).\end{aligned}$$

从而有

$$F_n\geqslant\frac{b_1-a_1}{\delta}.$$

此外, 根据斐波那契数列的定义, 可求出其通项公式

$$F_k = \frac{1}{\sqrt{5}}\left\{\left(\frac{1+\sqrt{5}}{2}\right)^{k+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{k+1}\right\}, \quad k = 0, 1, 2, \cdots.$$

因此, 对于给定的初始搜索区间 $[a, b]$ 及精度参数 δ, 可先确定 F_n 的下界, 然后利用通项公式求出满足要求的最小 n, 再通过采用与黄金分割法类似的迭代过程, 求出最终的搜索区间. 容易看出

$$\lim_{k\to\infty} \frac{F_{k-1}}{F_k} = \frac{\sqrt{5}-1}{2} = \tau,$$

故斐波那契法的区间缩短率与黄金分割法相同, 也是线性收敛的. 可以证明: 斐波那契法是用分割方法求一维极小化问题的最优策略, 而黄金分割法是近似最优的. 但由于后者简单易行, 因而得到了广泛的应用.

还有一种简单的分割方法是**二分法**, 基本思想是: 通过计算函数导数值来缩短搜索区间. 设初始搜索区间为 $[a_0, b_0] = [a, b]$, 第 k 步时的搜索区间为 $[a_k, b_k]$, 满足 $\varphi'(a_k) \leqslant 0, \varphi'(b_k) \geqslant 0$, 取中点

$$c_k = \frac{a_k + b_k}{2}.$$

若 $\varphi'(c_k) \geqslant 0$, 则令 $a_{k+1} = a_k, b_{k+1} = c_k$; 否则, 令 $a_{k+1} = c_k, b_{k+1} = b_k$, 从而得到新的搜索区间 $[a_{k+1}, b_{k+1}]$. 重复上述步骤, 直到搜索区间的长度小于事先设定的精度为止. 容易看出, 二分法每次迭代都将区间缩短一半, 故二分法的收敛速度也是线性的.

最后, 介绍一类重要的一维搜索方法, 称为**插值法**. 它的基本思想是: 在搜索区间中不断用低次 (通常不超过三次) 多项式来近似目标函数, 并逐步用插值多项式的极小点来逼近一维搜索问题

$$\varphi(\lambda^*) = \min_{\lambda\geqslant 0} \varphi(\lambda)$$

的极小点. 当函数具有较好的解析性质时, 插值法比直接方法, 譬如黄金分割法和斐波那契法等, 效果更好.

考虑利用一点处的函数值、一阶和二阶导数值来构造二次插值函数. 设二次插值多项式为

$$q(x) = ax^2 + bx + c,$$

则 $q(x)$ 在

$$x = -\frac{b}{2a}$$

处取到极小值, 故可作为计算近似极小点的公式. 若已知函数在 λ 处的函数值、一阶和二阶导数值, 即 $\varphi(\lambda), \varphi'(\lambda), \varphi''(\lambda)$, 则 $q(x)$ 需满足插值条件

$$\begin{aligned} q(\lambda) &= a\lambda^2 + b\lambda + c = \varphi(\lambda), \\ q'(\lambda) &= 2a\lambda + b = \varphi'(\lambda), \\ q''(\lambda) &= 2a = \varphi''(\lambda). \end{aligned}$$

解得

$$a = \frac{\varphi''(\lambda)}{2}, \quad b = \varphi'(\lambda) - \varphi''(\lambda)\lambda.$$

从而有迭代计算公式

$$\lambda_{k+1} = \lambda_k - \frac{\varphi'(\lambda_k)}{\varphi''(\lambda_k)}, \quad k = 0, 1, \cdots,$$

称为**一点二次插值法**, 或称为**牛顿法**. 可以证明: 当初始值 λ_0 充分靠近 λ^* 时, 由牛顿法产生的迭代序列 $\{\lambda_k\}_{k=1}^{\infty}$ 是收敛的, 且收敛速度是二阶的.

牛顿法的优点是速度速度快, 但是需要用到二阶导数值, 计算代价高. 因此, 一个自然的想法是: 用两个点 λ_1, λ_2 处的函数值及其中一个点的导数值 $\varphi'(\lambda_1)$(或 $\varphi'(\lambda_2)$)来构造二次插值函数. 通过插值条件并求解, 可得计算公式:

$$\lambda_{k+1} = \lambda_k - \frac{(\lambda_k - \lambda_{k-1})\varphi'(\lambda_k)}{2\left[\varphi'(\lambda_k) - \dfrac{\varphi(\lambda_k) - \varphi(\lambda_{k-1})}{\lambda_k - \lambda_{k-1}}\right]}, \quad k = 1, \cdots,$$

称为**两点二次插值法**, 其收敛速度的阶为 $(1+\sqrt{5})/2 \approx 1.618$. 类似地, 还有三点二次插值法, 二点三次插值法等.

一维搜索过程是非线性优化方法的基本组成部分, 前述通过求解一维搜索最优化问题的方法, 譬如黄金分割法, 牛顿法等, 统称为**精确一维搜索**, 需要花费很大的计算量. 特别地, 当迭代点远离问题的解时, 精确得求解一个一维子问题通常不是十分有效. 因此, 提出了**非精确一维搜索**, 即选取步长 λ_k 使得目标函数 $f(\mathbf{x})$ 有可接收的下降量 $f(\mathbf{x}_k) - f(\mathbf{x}_k + \lambda_k \mathbf{p}_k) > \delta$. 非精确一维搜索包括 Armijo-Goldstein 方法, Wolfe-Powell 方法等. 在非线性优化中, 一个可用来代替一维线搜索的方法是**信赖域方法** (Trust-Region Methods), 其基本思想是: 先设定一个可信的步长, 然后利用局部 n 维二次模型寻找最优下降方向和步长. 因步长受到使 Taylor 展开式有效的信赖域限制, 故又称**限步长法**. 信赖域方法具有快速的局部收敛性, 又有理想的总体收敛性, 可参见 [4, 6].

10.6 无约束非线性优化

本小节研究一类常见的非线性优化问题: **无约束非线性优化**, 即

$$\min\ f(\mathbf{x}),\quad \mathbf{x}\in\mathbb{R}^n \tag{10.8}$$

其中 $f(\mathbf{x}):\mathbb{R}^n\mapsto\mathbb{R}$ 是有 n 个自变量的连续函数.

求解无约束最优化问题最简单的方法是**最速下降法**, 它以负梯度方向作为极小化算法的下降方向, 故又称为**梯度法**. 设函数 $f(\mathbf{x})$ 在 $\mathbf{x}_k$ 附件连续可微, 且 $\nabla f(\mathbf{x}_k)\neq 0$. 利用 Taylor 展开式

$$f(\mathbf{x})=f(\mathbf{x}_k)+(\mathbf{x}-\mathbf{x}_k)^{\mathrm{T}}\nabla f(\mathbf{x}_k)+o(\|\mathbf{x}-\mathbf{x}_k\|)$$

可知, 若记 $\mathbf{x}-\mathbf{x}_k=\lambda\mathbf{p}_k$, $\mathbf{g}_k=\nabla f(\mathbf{x}_k)$, 则满足 $\mathbf{p}_k^{\mathrm{T}}\mathbf{g}_k<0$ 的方向 $\mathbf{p}_k$ 都是下降方向. 当 λ 的值固定后, 若 $\mathbf{p}_k^{\mathrm{T}}\mathbf{g}_k$ 的值越小, 即 $-\mathbf{p}_k^{\mathrm{T}}\mathbf{g}_k$ 的值越大, 则函数 $f(\mathbf{x})$ 的值下降的越快. 另外, 由 Cauchy-Schwartz 不等式知最优化问题

$$\min_{\mathbf{p}_k\in\mathbb{R}^n}\ \mathbf{p}_k^{\mathrm{T}}\mathbf{g}_k,\quad \text{subject to}\ \|\mathbf{p}_k\|=\|\mathbf{g}_k\|,$$

当且仅当 $\mathbf{p}_k=-\mathbf{g}_k$ 时, $\mathbf{p}_k^{\mathrm{T}}\mathbf{g}_k$ 的值最小, 称 $-\mathbf{g}_k$ 为最速下降方向. 最速下降法的计算步骤如下:

1. 设定初始值 $\mathbf{x}_0\in\mathbb{R}^n$ 及精度参数 $\varepsilon>0$, 并令 $k\leftarrow 0$;
2. 计算梯度 $\mathbf{g}_k=\nabla f(\mathbf{x}_k)$, 令 $\mathbf{p}_k=-\mathbf{g}_k$;
3. 若 $\|\mathbf{g}_k\|<\varepsilon$, 则算法终止; 否则, 利用一维搜索求步长因子 λ_k, 使得

$$f(\mathbf{x}_k+\lambda_k\mathbf{p}_k)=\min_{\lambda\geqslant 0}f(\mathbf{x}_k+\lambda\mathbf{p}_k);$$

4. 计算 $\mathbf{x}_{k+1}\leftarrow\mathbf{x}_k+\lambda_k\mathbf{p}_k$, 令 $k\leftarrow k+1$, 转步骤 2.

可以证明: 设 $f(\mathbf{x})\in C^1$, 在最速下降法中采用精确一维搜索, 则产生的迭代点序列 $\{\mathbf{x}_k\}_{k=1}^{\infty}$ 的每一个聚点都是驻点. 进一步, 若 $\lim\limits_{k\to\infty}\mathbf{x}_k=\mathbf{x}^*$ 且 $f(\mathbf{x})$ 在 $\mathbf{x}^*$ 的某一邻域内二次连续可微, 存在 $\varepsilon>0$ 和 $M>m>0$, 使得当 $\|\mathbf{x}-\mathbf{x}^*\|<\varepsilon$ 时, 有

$$m\|\mathbf{y}\|^2\leqslant\mathbf{y}^{\mathrm{T}}G(\mathbf{x})\mathbf{y}\leqslant M\|\mathbf{y}\|^2,\quad\forall\mathbf{y}\in\mathbb{R}^n,$$

其中 $G(\mathbf{x})=\nabla^2 f(\mathbf{x})$, 则最速下降法至少是线性收敛的.

由于最速下降方向是函数的局部性质, 对许多实际问题, 最速下降法并非 “最速下降”, 而是下降非常缓慢. 数值实验表明, 当目标函数的等值线或面接近于一个圆或球时, 最速下降法下降较快; 而当目标函数的等值线或面接近于一个扁长的椭圆或椭球时, 最速下降法开始几步下降较快, 后来就出现锯齿现象, 下降十分缓慢. 事实上, 因采用一维精确搜索, 由定理10.7知

$$\mathbf{g}_{k+1}^{\mathrm{T}}\mathbf{p}_k = \mathbf{p}_{k+1}^{\mathrm{T}}\mathbf{p}_k = 0,$$

即在相邻的两个迭代点上, 两个搜索方向是相互正交的, 这便是产生锯齿状现象的原因. 当接近极小点时, 步长越小, 前进越慢.

容易看出, 最速下降法仅用到了函数 $f(\mathbf{x})$ 的一阶局部信息. 若目标函数 $f(\mathbf{x})$ 是二次连续可微的, 则在 $\mathbf{x}_k$ 附件可用 $f(\mathbf{x})$ 的二次 Taylor 展开式

$$q_k(\mathbf{h}) = f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^{\mathrm{T}}\mathbf{h} + \frac{1}{2}\mathbf{h}^{\mathrm{T}}\nabla^2 f(\mathbf{x}_k)\mathbf{h}$$

来近似 $f(\mathbf{x})$, 其中 $\mathbf{h} = \mathbf{x} - \mathbf{x}_k$. 若矩阵 $\nabla^2 f(\mathbf{x}_k)$ 正定, 则 n 维二次函数 $q_k(\mathbf{x})$ 在

$$\frac{\partial q_k(\mathbf{h})}{\partial \mathbf{h}} = \mathbf{0} \Leftrightarrow \nabla^2 f(\mathbf{x}_k)\mathbf{h} = -\nabla f(\mathbf{x}_k) \Leftrightarrow \mathbf{h} = -\left[\nabla^2 f(\mathbf{x}_k)\right]^{-1}\nabla f(\mathbf{x}_k)$$

处取到极小值. 因此, 可设计迭代格式

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \left[\nabla^2 f(\mathbf{x}_k)\right]^{-1}\nabla f(\mathbf{x}_k), \quad k = 0, 1, \cdots,$$

称为**牛顿迭代法**. 若记 $G_k = \nabla^2 f(\mathbf{x}_k)$, $\mathbf{g}_k = \nabla f(\mathbf{x}_k)$, 则牛顿迭代可简写成

$$\mathbf{x}_{k+1} = \mathbf{x}_k - G_k^{-1}\mathbf{g}_k, \quad k = 0, 1, \cdots.$$

不难看出, 牛顿迭代法是在椭球范数 $\|\cdot\|_{G_k}$ 下的最速下降法. 事实上, 对于局部近似函数 $f(\mathbf{x}_k + \mathbf{h}) \approx f(\mathbf{x}_k) + \mathbf{g}_k^{\mathrm{T}}\mathbf{h}$, 考虑极小化问题

$$\min_{\mathbf{h}\in\mathbb{R}^n} \frac{\mathbf{g}_k^{\mathrm{T}}\mathbf{h}}{\|\mathbf{h}\|}$$

的最优解, 其中 $\|\cdot\|$ 是某种向量范数. 当采用 l_2 范数时, 即 $\|\mathbf{x}\|^2 = \mathbf{x}^{\mathrm{T}}\mathbf{x}$, 可得

$$\mathbf{h} = -\mathbf{g}_k,$$

所得的方法是最速下降法. 当采用椭球范数 $\|\cdot\|_{G_k}$ 时, 即 $\|\mathbf{x}\|^2 = \mathbf{x}^{\mathrm{T}}G_k\mathbf{x}$, 可得

$$\mathbf{h} = -G_k^{-1}\mathbf{g}_k,$$

所得的方法是牛顿迭代法.

显然, 当 $f(\mathbf{x})$ 是正定二次函数时, $q_k(\mathbf{x}) = f(\mathbf{x})$, 故牛顿迭代法一步就可达最优解. 而对于非二次函数, 牛顿迭代法不能保证经过有限次迭代求得最优解, 但由于目标函数在极小点附件可近似于二次函数, 故当初始点靠近极小点时, 牛顿迭代法的收敛速度一般是快的. 可以证明: 设 $f(\mathbf{x}) \in C^2$, $\mathbf{x}_k$ 充分靠近 $\mathbf{x}^*$, 如果 $\nabla^2 f(\mathbf{x}^*)$ 正定, 且 $G(x) = \nabla^2 f(\mathbf{x})$ 的每一个元素的函数满足 Lipschitz 条件, 则牛顿迭代法所得的序列 $\{\mathbf{x}_k\}_{k=1}^{\infty}$ 收敛于 $\mathbf{x}^*$, 且具有二阶的收敛速度.

不难发现, 当初始点远离最优解时, G_k 不一定正定, 牛顿方向不一定是下降方向, 收敛性不能保证, 这说明恒取步长为 1 的牛顿迭代法是不合适的. 因此, 应该引入某种一维搜索来确定步长因子, 这种方法称为**带步长因子的牛顿法**, 计算步骤如下:

1. 设定初始值 $\mathbf{x}_0 \in \mathbb{R}^n$ 及精度参数 $\varepsilon > 0$, 并令 $k \leftarrow 0$;
2. 计算梯度 $\mathbf{g}_k = \nabla f(\mathbf{x}_k)$. 若 $\|\mathbf{g}_k\| < \varepsilon$, 则算法终止; 否则, 进入下一步;
3. 计算矩阵 G_k, 并求解线性方程组 $G_k\mathbf{p} = -\mathbf{g}_k$ 得牛顿方向 $\mathbf{p}_k$;
4. 利用一维搜索求步长因子 λ_k, 使得

$$f(\mathbf{x}_k + \lambda_k \mathbf{p}_k) = \min_{\lambda \geqslant 0} f(\mathbf{x}_k + \lambda \mathbf{p}_k);$$

5. 计算 $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k + \lambda_k \mathbf{p}_k$, 令 $k \leftarrow k+1$, 转步骤 2.

接下来, 介绍一类重要的方法: **共轭方向法**, 一种介于最速下降法与牛顿法之间的方法, 它仅需利用一阶导数信息, 但克服了最速下降法收敛慢的缺点, 又避免了存储和计算牛顿法所需要的二阶导数信息. 共轭方向法是从研究二次函数的极小化问题产生的, 是一种迭代求解大型稀疏正定线性方程组的有效方法, 同时它还可以推广到处理非二次函数的极小化问题.

定义 10.16 设 G 是 n 阶正定方阵, $\mathbf{p}_1, \mathbf{p}_2$ 是 n 维非零向量, 如果 $\mathbf{p}_1^{\mathrm{T}} G \mathbf{p}_2 = 0$, 则称 $\mathbf{p}_1$ 与 $\mathbf{p}_2$ 是**G-共轭的**. 类似地, 设 $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n$ 是一组 n 维非零向量, 如果

$$\mathbf{p}_i^{\mathrm{T}} G \mathbf{p}_j = 0, \quad \forall 1 \leqslant i \neq j \leqslant n,$$

则称 $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n$ 是 G-共轭的.

显然, 如果 $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n$ 是 G-共轭的, 则它们是线性无关的. 当 $G = I$ 时, G-共轭性就是通常的正交性.

一般共轭方向法的计算步骤如下:

1. 设定初始值 $\mathbf{x}_0 \in \mathbb{R}^n$ 及初始方向 $\mathbf{g}_0$, 计算 $\mathbf{p}_0$, 使得 $\mathbf{p}_0^{\mathrm{T}}\mathbf{g}_0 < 0$, 并令 $k \leftarrow 0$;
2. 计算 λ_k 和 $\mathbf{x}_{k+1}$, 使得

$$f(\mathbf{x}_k + \lambda_k \mathbf{p}_k) = \min_{\lambda \geqslant 0} f(\mathbf{x}_k + \lambda \mathbf{p}_k),$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{p}_k;$$

3. 计算 $\mathbf{p}_{k+1}$, 使得 $\mathbf{p}_{k+1}^{\mathrm{T}} G \mathbf{p}_j = 0$, 其中 $j = 0, 1, 2, \cdots, k$;
4. 令 $k \leftarrow k+1$, 转步骤 2.

共轭方向法有一个基本性质: 只要执行精确一维搜索, 就具有二次终止性, 即下面的定理.

定理 10.9 (共轭方向基本定理) 设 $f(\mathbf{x}) = \dfrac{1}{2}\mathbf{x}^{\mathrm{T}} G \mathbf{x} + \mathbf{g}^{\mathrm{T}}\mathbf{x} + c$ 是正定二次函数, 其中 G 是 n 阶正定方阵, $\mathbf{g}$ 是 n 维向量, c 是常数, 则共轭方向法至多经过 n 步精确线性搜索终止; 且每一个 $\mathbf{x}_{k+1}$ 都是 $f(\mathbf{x})$ 在点 $\mathbf{x}_0$ 和方向 $\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_k$ 所张成的线性流形 $\left\{\mathbf{x} \,\middle|\, \mathbf{x} = \mathbf{x}_0 + \sum\limits_{j=0}^{k} \lambda_j \mathbf{p}_j, \quad \forall \lambda_j \in \mathbb{R}\right\}$ 中的极小点.

证明 因 $f(\mathbf{x})$ 是正定二次函数, 故存在唯一的 $\mathbf{x}^*$, 使得 $f(\mathbf{x})$ 在 $\mathbf{x}^*$ 取极小值, $\mathbf{x}^*$ 满足

$$G\mathbf{x}^* + \mathbf{g} = 0.$$

设共轭方向法执行 n 步所产生的搜索方向依次为 $\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_{n-1}$, 满足 G- 共轭条件, 是线性无关的, 故构成线性空间 $\mathbb{R}^n$ 的一组基. 因此, 存在 $\mu_0, \mu_1, \cdots, \mu_{n-1} \in \mathbb{R}$, 使得

$$\mathbf{x}^* - \mathbf{x}_0 = \mu_0 \mathbf{p}_0 + \mu_1 \mathbf{p}_1 + \cdots + \mu_{n-1}\mathbf{p}_{n-1},$$

上式两端同乘以 $\mathbf{p}_k^{\mathrm{T}} G$, 并利用 G-共轭条件, 可得

$$\mu_k = \frac{\mathbf{p}_k^{\mathrm{T}} G(\mathbf{x}^* - \mathbf{x}_0)}{\mathbf{p}_k^{\mathrm{T}} G \mathbf{p}_k}.$$

另一方面, 记共轭方向法所产生的序列为

$$\mathbf{x}_k = x_0 + \lambda_0 \mathbf{p}_0 + \lambda_1 \mathbf{p}_1 + \cdots + \lambda_{k-1}\mathbf{p}_{k-1},$$

上式两端同乘以 $\mathbf{p}_k^{\mathrm{T}} G$, 并利用 G-共轭条件, 可得

$$\mathbf{p}_k^{\mathrm{T}} G(\mathbf{x}_k - \mathbf{x}_0) = 0.$$

记 $\mathbf{g}_k = G\mathbf{x}_k + \mathbf{g} = \nabla f(\mathbf{x}_k)$, 于是有

$$\mathbf{p}_k^{\mathrm{T}} G(\mathbf{x}^* - \mathbf{x}_0) = \mathbf{p}_k^{\mathrm{T}} G(\mathbf{x}^* - \mathbf{x}_k) = -\mathbf{p}_k^{\mathrm{T}}(G\mathbf{x}_k + \mathbf{g}) = -\mathbf{p}_k^{\mathrm{T}}\mathbf{g}_k.$$

在共轭方向法中, λ_k 是通过求解一维优化问题

$$\begin{aligned} f(\mathbf{x}_k + \lambda_k \mathbf{p}_k) &= \min_{\lambda \geqslant 0} f(\mathbf{x}_k + \lambda \mathbf{p}_k) \\ &= \min_{\lambda \geqslant 0} \left\{ \frac{1}{2}\lambda^2 \mathbf{p}_k^{\mathrm{T}} G \mathbf{p}_k + \lambda(G\mathbf{x}_k + \mathbf{g})^{\mathrm{T}} \mathbf{p}_k + \frac{1}{2}\mathbf{x}_k^{\mathrm{T}} G \mathbf{x}_k + \mathbf{g}^{\mathrm{T}} \mathbf{x}_k + c \right\} \end{aligned}$$

确定, 容易求得

$$\lambda_k = -\frac{(G\mathbf{x}_k + \mathbf{g})^{\mathrm{T}} \mathbf{p}_k}{\mathbf{p}_k^{\mathrm{T}} G \mathbf{p}_k} = -\frac{\mathbf{p}_k^{\mathrm{T}} \mathbf{g}_k}{\mathbf{p}_k^{\mathrm{T}} G \mathbf{p}_k}.$$

因此, 有 $\lambda_k = \mu_k$, 命题得证. □

在精确线性搜索的条件下, 利用 $\mathbf{g}_{i+1} - \mathbf{g}_i = G(\mathbf{x}_{i+1} - \mathbf{x}_i) = \lambda_i G \mathbf{p}_i$ 可知, 共轭方向法的梯度 $\mathbf{g}_{k+1}$ 满足

$$\mathbf{g}_{k+1}^{\mathrm{T}} \mathbf{p}_j = \mathbf{g}_{j+1}^{\mathrm{T}} \mathbf{p}_j + \sum_{i=j+1}^{k} (\mathbf{g}_{i+1}^{\mathrm{T}} - \mathbf{g}_i^{\mathrm{T}}) \mathbf{p}_j = 0, \quad j = 0, 1, \cdots, k. \tag{10.9}$$

若记 $\mathbf{x}^* = \mathbf{x}_0 + \sum_{i=0}^{n-1} y_i^* \mathbf{p}_i$, $\mathbf{x} = \mathbf{x}_0 + \sum_{i=0}^{n-1} y_i \mathbf{p}_i$, 则二次函数 $f(\mathbf{x})$ 可改写成

$$q(\mathbf{y}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}} G (\mathbf{x} - \mathbf{x}^*) + \tilde{c} = \frac{1}{2}(\mathbf{y} - \mathbf{y}^*)^{\mathrm{T}} S^{\mathrm{T}} G S (\mathbf{y} - \mathbf{y}^*) + \tilde{c},$$

其中 $S = (\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_{n-1})$. 利用 G-共轭性条件, 得

$$q(\mathbf{y}) = \frac{1}{2} \sum_{i=0}^{n-1} (y_i - y_i^*)^2 d_{ii},$$

其中 $d_{ii} = \mathbf{p}_i^{\mathrm{T}} G \mathbf{p}_i$. 因此, 通过选择 $y_i = y_i^*$, $i = 0, 1, \cdots, n-1$, 即可使 $q(\mathbf{y})$ 极小化, 这等价于分别沿着共轭方向 $\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_{n-1}$ 作精确一维搜索. 从坐标变换的角度看, 共轭性意味着存在一个恰当的坐标变换 S, 使得 G 在新的坐标系下是一个对角矩阵 $S^T G S$, 新的变量在二次函数中是相互分离的. 于是, 一个共轭方向法是在新坐标系中的一个交替变量法.

共轭梯度法是最著名的共轭方向法, 它使得最速下降方向具有共轭性, 从而提高算法的有效性和可靠性. 由于解正定线性方程组 $A\mathbf{x}=\mathbf{b}$ 等价于极小化一个正定二次函数

$$\min_{\mathbf{x}\in\mathbb{R}^n}\frac{1}{2}\mathbf{x}^{\mathrm{T}}A\mathbf{x}-\mathbf{b}^{\mathrm{T}}\mathbf{x},$$

共轭梯度法首先是作为解线性方程组的方法被提出来的. 不失一般性, 考虑正定二次函数

$$f(\mathbf{x})=\frac{1}{2}\mathbf{x}^{\mathrm{T}}G\mathbf{x}+\mathbf{g}^{\mathrm{T}}\mathbf{x}+c,$$

$f(\mathbf{x})$ 的梯度为 $\nabla f(\mathbf{x})=G\mathbf{x}+\mathbf{g}$. 共轭梯度法有一个非常特殊的性质: 在构造共轭方向时, $\mathbf{p}_k$ 仅依赖于 $\mathbf{p}_{k-1}$. 这意味着在实现算法时, 可以减少存储量和计算时间, 是一个很好的性质. 另外, 在搜索方向 $\mathbf{p}_k$ 上, 希望函数值是下降的, 故一个自然的选择是

$$\mathbf{p}_k=-\mathbf{g}_k+\beta_{k-1}\mathbf{p}_{k-1}, \tag{10.10}$$

其中 $\mathbf{g}_k=\nabla f(\mathbf{x_k})=G\mathbf{x_k}+\mathbf{g}$, β_{k-1} 是待定系数. 因 $\mathbf{p}_k$ 与 $\mathbf{p}_{k-1}$ 需满足 G-共轭条件, 上式两端左乘以 $\mathbf{p}_{k-1}^{\mathrm{T}}G$, 故有

$$\beta_{k-1}=\frac{\mathbf{g}_k^{\mathrm{T}}G\mathbf{p}_{k-1}}{\mathbf{p}_{k-1}^{\mathrm{T}}G\mathbf{p}_{k-1}},\quad k=1,2,\cdots,n-1.$$

对于 $\mathbf{p}_0$, 直接取 $\mathbf{p}_0=-\mathbf{g}_0$. 可以证明 $\mathbf{p}_0,\mathbf{p}_1,\cdots,\mathbf{p}_{n-1}$ 是 G-共轭的, 且每一次迭代所产生的搜索方向 $\mathbf{p}_k$ 和梯度方向 $\mathbf{g}_k$ 都包含于

$$\mathcal{K}(\mathbf{g}_0;k)\triangleq\operatorname{span}\left\{\mathbf{g}_0,G\mathbf{g}_0,\cdots,G^k\mathbf{g}_0\right\},$$

称 $\mathcal{K}(\mathbf{g}_0;k)$ 为 $\mathbf{g}_0$ 的 k 次 Krylov 子空间.

定理 10.10 设 $f(\mathbf{x})=\frac{1}{2}\mathbf{x}^{\mathrm{T}}G\mathbf{x}+\mathbf{g}^{\mathrm{T}}\mathbf{x}+c$ 是正定二次函数, 其中 G 是 n 阶正定方阵, $\mathbf{g}$ 是 n 维向量, c 是常数, 则采用精确线性搜索的共轭梯度法经过 $m\leqslant n$ 步终止, 且对任意的 $k\leqslant m$ 有

$$\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_i=0,\quad i=0,1,\cdots,k-1, \tag{10.11}$$

$$\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_i=0,\quad i=0,1,\cdots,k-1, \tag{10.12}$$

$$\mathbf{p}_i^{\mathrm{T}}\mathbf{g}_i=-\mathbf{g}_i^{\mathrm{T}}\mathbf{g}_i,\quad i=0,1,\cdots,k-1, \tag{10.13}$$

$$\operatorname{span}\left\{\mathbf{g}_0,\mathbf{g}_1,\cdots,\mathbf{g}_k\right\}=\mathcal{K}(\mathbf{g}_0;k), \tag{10.14}$$

$$\operatorname{span}\left\{\mathbf{p}_0,\mathbf{p}_1,\cdots,\mathbf{p}_k\right\}=\mathcal{K}(\mathbf{g}_0;k). \tag{10.15}$$

证明 由于采用精确搜索确定步长因子, 显然有

$$\begin{aligned}\mathbf{p}_i^{\mathrm{T}}\mathbf{g}_i &= -\mathbf{g}_i^{\mathrm{T}}\mathbf{g}_i + \beta_{i-1}\mathbf{p}_{i-1}^{\mathrm{T}}\mathbf{g}_i \\ &= -\mathbf{g}_i^{\mathrm{T}}\mathbf{g}_i, \quad i = 1, 2, \cdots, k-1,\end{aligned}$$

且 $\mathbf{p}_0^{\mathrm{T}}\mathbf{g}_0 = -\mathbf{g}_0^{\mathrm{T}}\mathbf{g}_0$, 故式 (10.13) 得证.

下面用数学归纳法. 先证式 (10.11) 和 (10.12). 当 $k = 1$ 时, 由式 (10.9) 和 (10.10) 知,

$$\mathbf{g}_1^{\mathrm{T}}\mathbf{g}_0 = -\mathbf{g}_1^{\mathrm{T}}\mathbf{p}_0 = 0, \quad \mathbf{p}_1^{\mathrm{T}}G\mathbf{p}_0 = (-\mathbf{g}_1^{\mathrm{T}} + \beta_0\mathbf{p}_0^{\mathrm{T}})G\mathbf{p}_0 = 0.$$

设式 (10.11) 和 (10.12) 对 k 成立, 下面证明对 $k+1$ 亦成立.

因 $f(\mathbf{x})$ 是正定二次函数, 故有

$$\mathbf{g}_{k+1} = \mathbf{g}_k + G(\mathbf{x}_{k+1} - \mathbf{x}_k) = \mathbf{g}_k + \lambda_k G\mathbf{p}_k, \tag{10.16}$$

上式左乘 $\mathbf{p}_k^{\mathrm{T}}$, 并注意到 λ_k 是精确搜索确定的步长因子, 有

$$\lambda_k = -\frac{\mathbf{p}_k^{\mathrm{T}}\mathbf{g}_k}{\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_k} = \frac{\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k}{\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_k} \neq 0.$$

由式 (10.16) 和 (10.10), 得

$$\begin{aligned}\mathbf{g}_{k+1}^{\mathrm{T}}\mathbf{g}_j &= \mathbf{g}_k^{\mathrm{T}}\mathbf{g}_j + \lambda_k\mathbf{p}_k^{\mathrm{T}}G\mathbf{g}_j \\ &= \mathbf{g}_k^{\mathrm{T}}\mathbf{g}_j - \lambda_k\mathbf{p}_k^{\mathrm{T}}G(\mathbf{p}_j - \beta_{j-1}\mathbf{p}_{j-1}).\end{aligned}$$

当 $j < k$ 时, 利用归纳假设可知上式为零; 当 $j = k$ 时, 上式成为

$$\mathbf{g}_{k+1}^{\mathrm{T}}\mathbf{g}_k = \mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k - \frac{\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k}{\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_k}\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_k = 0,$$

故式 (10.11) 得证.

类似地, 由式 (10.16) 和 (10.10) 得

$$\begin{aligned}\mathbf{p}_{k+1}^{\mathrm{T}}G\mathbf{p}_j &= -\mathbf{g}_{k+1}^{\mathrm{T}}G\mathbf{p}_j + \beta_k\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_j \\ &= \mathbf{g}_{k+1}^{\mathrm{T}}(\mathbf{g}_j - \mathbf{g}_{j+1})/\lambda_j + \beta_k\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_j.\end{aligned}$$

当 $j<k$ 时, 利用归纳假设可知上式为零; 当 $j=k$ 时, 上式成为

$$\begin{aligned}\mathbf{p}_{k+1}^{\mathrm{T}}G\mathbf{p}_k &= -\frac{\mathbf{g}_{k+1}^{\mathrm{T}}(\mathbf{g}_{k+1}-\mathbf{g}_k)}{\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k}\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_k+\frac{\mathbf{g}_{k+1}^{\mathrm{T}}G\mathbf{p}_k}{\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_k}\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_k\\ &= -\frac{\mathbf{g}_{k+1}^{\mathrm{T}}(\mathbf{g}_{k+1}-\mathbf{g}_k)}{\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k}\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_k+\frac{\mathbf{g}_{k+1}^{\mathrm{T}}(\mathbf{g}_{k+1}-\mathbf{g}_k)}{\mathbf{p}_k^{\mathrm{T}}(\mathbf{g}_{k+1}-\mathbf{g}_k)}\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_k\\ &= -\frac{\mathbf{g}_{k+1}^{\mathrm{T}}\mathbf{g}_{k+1}}{\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k}\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_k+\frac{\mathbf{g}_{k+1}^{\mathrm{T}}\mathbf{g}_{k+1}}{-\mathbf{p}_k^{\mathrm{T}}\mathbf{g}_k}\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_k\\ &= -\frac{\mathbf{g}_{k+1}^{\mathrm{T}}\mathbf{g}_{k+1}}{\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k}\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_k+\frac{\mathbf{g}_{k+1}^{\mathrm{T}}\mathbf{g}_{k+1}}{\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k}\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_k=0,\end{aligned}$$

故式 (10.12) 得证.

最后, 证明式 (10.14) 和 (10.15). 当 $k=0$ 时, 结论显然成立. 设式 (10.14) 和 (10.15) 对 k 成立, 下面证明对 $k+1$ 亦成立.

利用归纳假设知

$$\mathbf{g}_k, G\mathbf{p}_k \in \operatorname{span}\left\{\mathbf{g}_0, G\mathbf{g}_0, \cdots, G^k\mathbf{g}_0, G^{k+1}\mathbf{g}_0\right\},$$

又因式 (10.16) 有

$$\operatorname{span}\left\{\mathbf{g}_0, \mathbf{g}_1, \cdots, \mathbf{g}_k, \mathbf{g}_{k+1}\right\} \subset \operatorname{span}\left\{\mathbf{g}_0, G\mathbf{g}_0, \cdots, G^k\mathbf{g}_0, G^{k+1}\mathbf{g}_0\right\}.$$

另一方面, 利用归纳假设知

$$G^{k+1}\mathbf{g}_0 = G(G^k\mathbf{g}_0) \in \operatorname{span}\left\{G\mathbf{p}_0, G\mathbf{p}_1, \cdots, G\mathbf{p}_k\right\},$$

又因式 (10.16) 有

$$G\mathbf{p}_k = (\mathbf{g}_{k+1}-\mathbf{g}_k)/\lambda_k,$$

可得

$$G^{k+1}\mathbf{g}_0 \in \operatorname{span}\left\{\mathbf{g}_0, \mathbf{g}_1, \cdots, \mathbf{g}_k, \mathbf{g}_{k+1}\right\},$$

从而有

$$\operatorname{span}\left\{\mathbf{g}_0, G\mathbf{g}_0, \cdots, G^k\mathbf{g}_0, G^{k+1}\mathbf{g}_0\right\} \subset \operatorname{span}\left\{\mathbf{g}_0, \mathbf{g}_1, \cdots, \mathbf{g}_k, \mathbf{g}_{k+1}\right\},$$

故式 (10.14) 得证.

利用式 (10.10) 和归纳假设, 可知

$$\begin{aligned}\text{span}\left\{\mathbf{p}_0,\mathbf{p}_1,\cdots,\mathbf{p}_k,\mathbf{p}_{k+1}\right\} &= \text{span}\left\{\mathbf{p}_0,\mathbf{p}_1,\cdots,\mathbf{p}_k,\mathbf{g}_{k+1}\right\}\\ &= \text{span}\left\{\mathbf{g}_0,G\mathbf{g}_0,\cdots,G^k\mathbf{g}_0,\mathbf{g}_{k+1}\right\}\\ &= \text{span}\left\{\mathbf{g}_0,\mathbf{g}_1,\cdots,\mathbf{g}_k,\mathbf{g}_{k+1}\right\}\\ &= \text{span}\left\{\mathbf{g}_0,G\mathbf{g}_0,\cdots,G^k\mathbf{g}_0,G^{k+1}\mathbf{g}_0\right\},\end{aligned}$$

故式 (10.15) 得证. □

值得指出的是, $\mathbf{p}_0$ 必须取 $-\mathbf{g}_0$, 否则上述定理是不成立的. 式 (10.11) 表明梯度向量序列 $\{\mathbf{g}_i\}$ 是相互正交的, 而式 (10.12) 表明搜索方向序列 $\{\mathbf{p}_i\}$ 才是 G-共轭的. 由式 (10.13) 知, 共轭梯度法的每一步都是值下降的, 精确一维搜索的步长 λ_k 可写成:

$$\lambda_k = -\frac{\mathbf{p}_k^{\mathrm{T}}\mathbf{g}_k}{\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_k} = \frac{\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k}{\mathbf{p}_k^{\mathrm{T}}G\mathbf{p}_k}.$$

利用 $G\mathbf{p}_k = (\mathbf{g}_{k+1}-\mathbf{g}_k)/\lambda_k$ 及式 (10.11) 和 (10.13), 可得计算搜索方向的步长 β_k 几个常用公式

$$\begin{aligned}\beta_k &= \frac{\mathbf{g}_{k+1}^{\mathrm{T}}(\mathbf{g}_{k+1}-\mathbf{g}_k)}{\mathbf{p}_k^{\mathrm{T}}(\mathbf{g}_{k+1}-\mathbf{g}_k)} \text{ (Crowder-Wolfe 公式)}\\ &= \frac{\mathbf{g}_{k+1}^{\mathrm{T}}\mathbf{g}_{k+1}}{\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k} \text{ (Fletcher-Reeves 公式)}\\ &= \frac{\mathbf{g}_{k+1}^{\mathrm{T}}(\mathbf{g}_{k+1}-\mathbf{g}_k)}{\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k} \text{ (Polak-Ribiere-Polyak 公式)}\\ &= -\frac{\mathbf{g}_{k+1}^{\mathrm{T}}\mathbf{g}_{k+1}}{\mathbf{p}_k^{\mathrm{T}}\mathbf{g}_k} \text{ (Dixon 公式).}\end{aligned}$$

由上面的公式知, 共轭梯度法仅比最速下降法稍微复杂一点, 但却具有二次终止性, 所以共轭梯度法是一个很有效的方法.

容易看出, 共轭梯度法可推广至一般的目标函数 $f(\mathbf{x})$, 此时, 步长 λ_k 可通过精确一维搜索或非精确一维搜索得到, 而 β_k 的计算公式保持不变. 但是, 需要注意的是, n 步以后共轭梯度法所产生的搜索方向 $\mathbf{p}_n$ 不再与 $\mathbf{p}_0,\mathbf{p}_1,\cdots,\mathbf{p}_{n-1}$ 满足 G-共轭条件, 故需重新取最速下降方向作为搜索方向, 即

$$\mathbf{p}_{cn} = -\mathbf{g}_{cn},\quad c = 1,2,\cdots,$$

这种方法称为**再开始共轭梯度法**, 其计算步骤如下:

1. 设定初始值 $\mathbf{x}_0 \in \mathbb{R}^n$ 及精度参数 $\varepsilon > 0$, 计算 $\mathbf{g}_0 = \nabla f(\mathbf{x}_0)$, 令 $k \leftarrow 0$;
2. 若 $\|\mathbf{g}_0\| < \varepsilon$, 则算法终止; 否则, 令 $\mathbf{p}_0 \leftarrow -\mathbf{g}_0$;
3. 利用一维搜索求步长因子 λ_k, 使得

$$f(\mathbf{x}_k + \lambda_k \mathbf{p}_k) = \min_{\lambda \geqslant 0} f(\mathbf{x}_k + \lambda \mathbf{p}_k),$$

   令 $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k + \lambda_k \mathbf{p}_k$, 计算 $\mathbf{g}_{k+1} = \nabla f(\mathbf{x}_{k+1})$;
4. 若 $\|\mathbf{g}_{k+1}\| < \varepsilon$, 则算法终止; 否则, 进入下一步;
5. 若 $k = n-1$, 令 $\mathbf{x}_0 \leftarrow \mathbf{x}_{k+1}$, 转步骤 1; 否则, 进入下一步;
6. 计算 $\beta_k = \mathbf{g}_{k+1}^{\mathrm{T}} \mathbf{g}_{k+1} / \mathbf{g}_k^{\mathrm{T}} \mathbf{g}_k$, 令 $\mathbf{p}_{k+1} \leftarrow -\mathbf{g}_{k+1} + \beta_k \mathbf{p}_k$, $k \leftarrow k+1$;
7. 若 $\mathbf{p}_k^{\mathrm{T}} \mathbf{g}_k > 0$, 令 $\mathbf{x}_0 \leftarrow \mathbf{x}_k$, 转步骤 1; 否则, 转步骤 3.

在上述算法中, 步长因子 λ_k 也可以采用非精确一维搜索. 此时, β_k 的几种计算公式的效果是不同的, 参见 [4, 6].

共轭梯度法在一定条件下是收敛的. 可以证明: 若 $f(\mathbf{x})$ 在有界水平集 $L = \{\mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) \leqslant f(\mathbf{x}_0)\}$ 上连续可微, 则采用 Fletcher-Reeves 公式和精确一维搜索的共轭梯度法产生的序列 $\{\mathbf{x}_k\}$ 至少有一个聚点是驻点, 即

(1) 当 $\{\mathbf{x}_k\}$ 是有穷点列时, 其最后一个点 $\mathbf{x}^*$ 是 $f(\mathbf{x})$ 的驻点;

(2) 当 $\{\mathbf{x}_k\}$ 是无穷点列时, 它必有极限点, 且其任一极限点都是 $f(\mathbf{x})$ 的驻点.

采用其他公式和搜索策略也有类似的结论.

当 $f(\mathbf{x})$ 是正定二次函数时, 采用精确一维搜索的共轭梯度法至多在 n 次迭代后终止, 这是一个很好的性质. 而对于一般的非线性函数 $f(\mathbf{x})$, 因目标函数在极小点附件可近似于一个正定二次函数, 故若将 n 次迭代视为一次大的迭代, 则共轭梯度法应该与牛顿法有类似的收敛速度. 事实上, 可以证明: 设 $f(\mathbf{x}) \in C^3$, 且存在常数 $m, M > 0$, 使得

$$m\|\mathbf{y}\|^2 \leqslant \mathbf{y}^{\mathrm{T}} \nabla^2 f(\mathbf{x}) \mathbf{y} \leqslant M\|\mathbf{y}\|^2, \quad \forall \mathbf{y} \in \mathbb{R}^n, \mathbf{x} \in L,$$

其中 $L = \{\mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) \leqslant f(\mathbf{x}_0)\}$ 是有界水平集, 则采用 Fletcher-Reeves 公式和精确一维搜索的再开始共轭梯度法产生的序列 $\{\mathbf{x}_k\}$ 是 n 步二阶收敛的, 即存在常数 $c > 0$, 使得

$$\lim_{k \to \infty} \sup \frac{\|\mathbf{x}_{kn+n} - \mathbf{x}^*\|}{\|\mathbf{x}_{kn} - \mathbf{x}^*\|^2} \leqslant c < \infty.$$

最后, 介绍一类很有效的方法: **拟牛顿法**. 从前面的分析中不难发现, 牛顿法收敛速度快的一个关键是利用了目标函数的二阶信息, 即 Hessian 矩阵. 然而, 计算和存储 Hessian 矩阵的代价比较高, 有时甚至无法求出. 因此, 一个自然的想法是: 利用目标函数 $f(\mathbf{x})$ 及一阶导数 $\nabla f(\mathbf{x})$ 的信息来构造近似的 Hessian 矩阵, 从而实现加快收敛速度的目标.

设 $f(\mathbf{x}):\mathbb{R}^n\to\mathbb{R}$ 是二次连续可微的函数, 则 $f(\mathbf{x})$ 在 $\mathbf{x}_{k+1}$ 附件的二次近似函数为

$$f(\mathbf{x})\approx f(\mathbf{x}_{k+1})+\mathbf{g}_{k+1}^{\mathrm{T}}(\mathbf{x}-\mathbf{x}_{k+1})+\frac{1}{2}(\mathbf{x}-\mathbf{x}_{k+1})^{\mathrm{T}}G_{k+1}(\mathbf{x}-\mathbf{x}_{k+1}),$$

其中 $\mathbf{g}_{k+1}=\nabla f(\mathbf{x}_{k+1})$, $G_{k+1}=\nabla^2 f(\mathbf{x}_{k+1})$, 从而有

$$\nabla f(\mathbf{x})\approx \mathbf{g}_{k+1}+G_{k+1}(\mathbf{x}-\mathbf{x}_{k+1}).$$

令 $\mathbf{x}=\mathbf{x}_k$, 记 $\mathbf{s}_k=\mathbf{x}_{k+1}-\mathbf{x}_k$, $\mathbf{y}_k=\mathbf{g}_{k+1}-\mathbf{g}_k$, 则

$$G_{k+1}^{-1}\mathbf{y}_k\approx\mathbf{s}_k.$$

显然, 当 $f(\mathbf{x})$ 是二次函数时, 上述近似关系取等号. 在拟牛顿法中, 要求构造出来的 Hessian 矩阵的逆近似 H_{k+1} 满足这种关系, 即

$$H_{k+1}\mathbf{y}_k=\mathbf{s}_k, \tag{10.17}$$

称为**拟牛顿条件**或**拟牛顿方程**. 若记 $B_{k+1}=H_{k+1}^{-1}$, 则上述条件也可写为

$$B_{k+1}\mathbf{s}_k=\mathbf{y}_k,$$

是关于 Hessian 矩阵的拟牛顿条件. 容易验证, 如果 H_{k+1} 满足拟牛顿条件, 那么局部二次近似函数

$$q(\mathbf{x})=f(\mathbf{x}_{k+1})+\mathbf{g}_{k+1}^{\mathrm{T}}(\mathbf{x}-\mathbf{x}_{k+1})+\frac{1}{2}(\mathbf{x}-\mathbf{x}_{k+1})^{\mathrm{T}}H_{k+1}^{-1}(\mathbf{x}-\mathbf{x}_{k+1})$$

满足插值条件

$$\begin{cases}\nabla q(\mathbf{x}_k)=\mathbf{g}_k,\\ q(\mathbf{x}_{k+1})=f(\mathbf{x}_{k+1}),\\ \nabla q(\mathbf{x}_{k+1})=\mathbf{g}_{k+1}.\end{cases}$$

一般拟牛顿法的计算步骤如下:

1. 设定初始值 $\mathbf{x}_0\in\mathbb{R}^n$ 及精度参数 $\varepsilon>0$, $H_0\in\mathbb{R}^{n\times n}$, 并令 $k\leftarrow 0$;

2. 计算梯度 $\mathbf{g}_k = \nabla f(\mathbf{x}_k)$, 若 $\|\mathbf{g}_k\| < \varepsilon$, 则算法终止; 否则, 计算 $\mathbf{p}_k = -H_k\mathbf{g}_k$;
3. 利用一维搜索求步长因子 λ_k, 使得

$$f(\mathbf{x}_k + \lambda_k\mathbf{p}_k) = \min_{\lambda\geqslant 0} f(\mathbf{x}_k + \lambda\mathbf{p}_k);$$

4. 校正 H_k 产生 H_{k+1}, 使得拟牛顿条件 (10.17) 成立;
5. 计算 $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k + \lambda_k\mathbf{p}_k$, 令 $k \leftarrow k+1$, 转步骤 2.

类似地, 拟牛顿法也可采用近似 Hessian 矩阵 B_k 进行. 在实际应用中, 初始矩阵 H_0 可取单位阵, 此时, 拟牛顿法的第一次迭代等价于一个最速下降迭代.

不难看出, 拟牛顿法是在椭球范数 $\|\cdot\|_{H_k^{-1}}$ 下的最速下降法, 搜索方向

$$\mathbf{p}_k = -H_k\mathbf{g}_k$$

是 $f(\mathbf{x})$ 从 $\mathbf{x}_k$ 点出发的最速下降方向. 因在每一次迭代中, 度量矩阵 H_k^{-1} 都是变化的, 故方法也称为**变尺度法**.

与牛顿法相比, 拟牛顿法有以下优点:

(1) 仅需一阶导数信息;

(2) H_k 保持正定, 使得方法具有下降性质;

(3) 每次迭代需要 $O(n^2)$ 次乘法运算.

剩下的问题是: 如何校正 H_k 产生 H_{k+1}, 使得拟牛顿条件 (10.17) 成立, 并保持正定性. 设 H_{k+1} 是 H_k 通过对称秩二校正得到的, 即

$$H_{k+1} = H_k + a\mathbf{u}\mathbf{u}^{\mathrm{T}} + b\mathbf{v}\mathbf{v}^{\mathrm{T}},$$

其中 $a, b \in \mathbb{R}$, $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{n\times 1}$. 代入拟牛顿条件 (10.17) 得

$$H_k\mathbf{y}_k + a\mathbf{u}\mathbf{u}^{\mathrm{T}}\mathbf{y}_k + b\mathbf{v}\mathbf{v}^{\mathrm{T}}\mathbf{y}_k = \mathbf{s}_k.$$

显然, 向量 $\mathbf{u}$ 和 $\mathbf{v}$ 不能由上式唯一确定, 但是 $\mathbf{u}$ 和 $\mathbf{v}$ 可取为

$$\mathbf{u} = \mathbf{s}_k, \quad \mathbf{v} = H_k\mathbf{y}_k,$$

这时, 只要 a 和 b 满足

$$a\mathbf{u}^{\mathrm{T}}\mathbf{y}_k = 1, \quad b\mathbf{v}^{\mathrm{T}}\mathbf{y}_k = -1,$$

拟牛顿条件 (10.17) 就成立了. 解得

$$H_{k+1} = H_k + \frac{\mathbf{s}_k\mathbf{s}_k^{\mathrm{T}}}{\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k} - \frac{H_k\mathbf{y}_k\mathbf{y}_k^{\mathrm{T}}H_k}{\mathbf{y}_k^{\mathrm{T}}H_k\mathbf{y}_k}, \tag{10.18}$$

称为 **DFP 校正公式**, 是由 Davidon, Fletcher 和 Powell 发展出来的, 具有很多重要的性质:

(1) 校正保持正定性, 故下降性质成立;

(2) 每次迭代需要 $3n^2 + O(n)$ 次乘法运算;

(3) 具有超线性的收敛速度;

(4) 对于凸函数, 当采用精确一维搜索时, 方法具有总体收敛性.

利用 $H_{k+1} \longleftrightarrow B_{k+1}, \mathbf{s}_k \longleftrightarrow \mathbf{y}_k$ 之间的对偶关系及矩阵逆的秩一校正公式, 可得

$$H_{k+1} = H_k + \left(1 + \frac{\mathbf{y}_k^{\mathrm{T}}H_k\mathbf{y}_k}{\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k}\right)\frac{\mathbf{s}_k\mathbf{s}_k^{\mathrm{T}}}{\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k} - \frac{\mathbf{s}_k\mathbf{y}_k^{\mathrm{T}}H_k + H_k\mathbf{y}_k\mathbf{s}_k^{\mathrm{T}}}{\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k}, \tag{10.19}$$

称为 **BFGS 校正公式**, 是由 Broyden, Fletcher, Goldfarb 和 Shanno 发展出来的. BFGS 校正公式具备与 DFP 校正公式相同的性质, 并且对于凸函数, 当采用非精确一维搜索时, 仍具有总体收敛性. 在实际应用中, BFGS 校正公式的结果也优于 DFP 校正公式, 是最常用的拟牛顿方法之一.

一般地, 当初始点靠近极小点时, 拟牛顿迭代法在一定条件下是线性收敛的, 很多时候甚至是超线性收敛的. 对于凸函数, 拟牛顿法还有整体收敛性. 可以证明: 当 $f(\mathbf{x})$ 是一致凸的二阶连续可微函数时, 采用精确一维搜索和 DFP 校正公式的拟牛顿法整体收敛; 当 $f(\mathbf{x})$ 是凸的二阶连续可微函数时, 采用非精确一维搜索的 Wolfe-Powell 准则和 BFGS 校正公式的拟牛顿法整体收敛.

前面介绍了一些求解无约束非线性优化问题的常用方法, 都是确定性的方法. 在许多实际应用中, 譬如深度学习, 压缩感知, 低秩矩阵填充等, 优化变量的个数可能会很大, 如几十万甚至更多, 这时即使采用梯度下降法, 每一步迭代所需的计算量也是相当可观的. 因此, 人们提出了**随机梯度下降法**, 基本思想是: 梯度向量可以视为期望, 而期望可以使用小规模的样本来近似估计. 设优化的目标函数为 $f(\mathbf{x})$, 优化变量 $\mathbf{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$, 先任取它的一个有 m 个元素的子集 $\{x_{i1}, x_{i2}, \cdots, x_{im}\}$, m 远小于 n, 然后计算 $f(\mathbf{x})$ 关于变量 $x_{i1}, x_{i2}, \cdots, x_{im}$ 的梯度并乘以一个称为学习率的常数 r, 其他变量视为常数, 执行一次梯度下降. 重复以上步骤, 直至算法满足收敛准则, 可参见 [5]. 此外, 如果非线性优化问题中含有等式或不等式约束条件, 则称为**带约束的非线性优化问题**, 这类问题求解起来更加困难, 感兴趣的读者可参见 [1, 4, 6].

参考文献

[1] Roger Fletcher. Practical Methods of Optimization. 2nd, Ed. John Wiley & Sons, 1987.

[2] William J. Cook, William H. Cunningham, William R. Pulleyblank, Alexander Schrijver. Combinatorial Optimization. John Wiley & Sons, 1998.

[3] Stephen Boyd, Lieven Vandenberghe. Convex Optimization. Cambridge University Press, 2004.

[4] Jorge Nocedal, Stephen J. Wright. Numerical Optimization. 2nd, Ed. Springer, 2006.

[5] Ian Goodfellow, Yoshua Bengio, Aaron Courville. Deep Learning. MIT Press, 2016.

[6] 袁亚湘, 孙文瑜. 最优化理论与方法. 科学出版社, 1997.

[7] 黄云清, 舒适, 陈艳萍, 金继承, 文立平. 数值计算方法. 科学出版社, 2009.

[8] 孙小玲, 李端. 整数规划. 科学出版社, 2010.

[9] 《运筹学》教材编写组. 运筹学. 第 4 版. 清华大学出版社, 2012.

[10] 胡运权主编. 运筹学教程. 清华大学出版社, 2018.

习　题

1. 某养殖场靠饲养生猪盈利, 若每头猪每天至少需要 700g 蛋白质、30g 矿物质、10mg 维生素, 现有五种饲料可供选择, 各种饲料每 kg 的营养成分含量及单价如下表所示

饲料	蛋白质 (g)	矿物质 (g)	维生素 (mg)	价格 (元/kg)
A	50	2	1.0	2.0
B	30	5	0.1	2.4
C	20	8	0.8	1.5
D	40	4	0.2	3.0
E	80	1	0.4	1.8

试建立确定既满足生猪营养需求, 又使费用最低的饲料配方所使用的最优化模型.

2. 用图解法求解下列线性规划问题, 并指出问题是否有唯一最优解、无穷多最优

解、无界解还是无可行解?

(1) $\max z = 2x_1 + 3x_2$,

$$\text{s. t.} \begin{cases} x_1 + 2x_2 \leqslant 8, \\ 2x_1 + x_2 \geqslant 1, \\ x_2 \leqslant 3, \\ x_1, x_2 \geqslant 0. \end{cases}$$

(2) $\min z = x_1 + x_2$,

$$\text{s. t.} \begin{cases} x_1 + 3x_2 \geqslant 3, \\ x_1 - x_2 \geqslant 2, \\ x_1, x_2 \geqslant 0. \end{cases}$$

(3) $\max z = x_1 - x_2$,

$$\text{s. t.} \begin{cases} 2x_1 - x_2 \geqslant -1, \\ -0.5x_1 + x_2 \leqslant 2, \\ x_1, x_2 \geqslant 0. \end{cases}$$

(4) $\max z = 2x_1 + x_2$,

$$\text{s. t.} \begin{cases} x_1 - 2x_2 \geqslant 1, \\ 3x_1 - x_2 \leqslant -1, \\ x_1, x_2 \geqslant 0. \end{cases}$$

3. 将下列线性规划问题化为标准形式, 并列出初始单纯形表.

(1) $\min z = -x_1 + 2x_2 - 3x_3 + 2x_4$,

$$\text{s. t.} \begin{cases} 4x_1 - x_2 + 2x_3 - x_4 = -2, \\ x_1 + x_2 - x_3 + 2x_4 \leqslant 14, \\ -2x_1 + 3x_2 + x_3 - x_4 \geqslant 2, \\ x_1, x_2, x_3 \geqslant 0, x_4\text{无约束}. \end{cases}$$

(2) $\max z = x_1 - 3x_2 + 2x_3$,

$$\text{s. t.} \begin{cases} -x_1 + x_2 + x_3 = 4, \\ -2x_1 + x_2 - x_3 \leqslant 6, \\ x_1 \leqslant 0, x_2 \geqslant 0, x_3\text{无约束}. \end{cases}$$

4. 求下列线性规划问题中满足约束条件的所有基解, 并指出哪些是基可行解, 并代入目标函数, 确定哪一个是最优解.

(1) $\min z = 2x_1 - x_2 + 3x_3 + 2x_4$,

$$\text{s. t.} \begin{cases} 2x_1 + 3x_2 - x_3 - 4x_4 = 8, \\ x_1 - 2x_2 + 6x_3 - 7x_4 = -3, \\ x_1, x_2, x_3, x_4 \geqslant 0. \end{cases}$$

(2) $\min z = 3x_1 + 2x_2 - 3x_3 + 6x_4$,

$$\text{s. t.} \begin{cases} x_1 + 2x_2 + 3x_3 + 2x_4 = 7, \\ 2x_1 + 4x_2 + x_3 - 2x_4 = 3, \\ x_1, x_2, x_3, x_4 \geqslant 0. \end{cases}$$

5. 用单纯形方法求解以下线性规划问题:

(1) $\max z = 2x_1 + 3x_2$,

$$\text{s. t.} \begin{cases} 2x_1 + x_2 \leqslant 8, \\ x_1 + 2x_2 \leqslant 7, \\ x_1, x_2 \geqslant 0. \end{cases}$$

(2) $\max z = 3x_1 - 2x_2 + 5x_3$,

$$\text{s. t.} \begin{cases} 3x_1 + 2x_3 \leqslant 13, \\ x_2 + 3x_3 \leqslant 17, \\ 2x_1 + x_2 + x_3 \leqslant 13, \\ x_1, x_2, x_3 \geqslant 0. \end{cases}$$

6. 分别用大 M 法和两阶段法求解下列线性规划问题, 并指出属于哪一类解.

(1) $\min z = 3x_1 - x_2$,

$$\text{s. t.} \begin{cases} 3x_1 + x_2 \geqslant 3, \\ 2x_1 - 3x_2 \geqslant 1, \\ x_1, x_2 \geqslant 0. \end{cases}$$

(2) $\min z = 3x_1 + 2x_2 + x_3$,

$$\text{s. t.} \begin{cases} 2x_1 + 4x_2 + 5x_3 \geqslant 0, \\ 3x_1 - x_2 + 7x_3 \geqslant 2, \\ 2x_1 + 2x_2 + x_3 \geqslant 16, \\ x_1, x_2, x_3 \geqslant 0. \end{cases}$$

7. 考虑线性规划问题

$$\max z = c_1 x_1 + c_2 x_2,$$

$$\text{s. t.} \begin{cases} a_{11} x_1 + a_{12} x_2 \leqslant b_1, \\ a_{21} x_1 + a_{22} x_2 \leqslant b_2, \\ x_1, x_2 \geqslant 0, \end{cases}$$

其中 $1 \leqslant c_1 \leqslant 3, 4 \leqslant c_2 \leqslant 6, -1 \leqslant a_{11} \leqslant 3, 2 \leqslant a_{12} \leqslant 5, 8 \leqslant b_1 \leqslant 12, 2 \leqslant a_{21} \leqslant 4, 4 \leqslant a_{22} \leqslant 6, 10 \leqslant b_2 \leqslant 14$, 试求目标函数 z 最优值的上界和下界.

8. 分别用黄金分割法与斐波那契法求函数

$$f(x) = 2x^2 - 8x + 3$$

在区间 $[0, 4]$ 上的极小点, 要求缩短后的区间长度不大于原区间长度的 5%.

9. 分别用最速下降法与牛顿法求函数

$$f(\mathbf{x}) = x_1^2 - x_1 x_2 + x_2^2 + x_1 x_3 + x_3^2 - 2x_1 + 4x_2 + 2x_3 - 2, \quad \mathbf{x} = (x_1, x_2, x_3)^{\mathrm{T}} \in \mathbb{R}^3,$$

的极小点, 初始点 $\mathbf{x}_0 = (0, 0, 0)^{\mathrm{T}}$, 要求: (1) 最速下降法进行三次迭代, 并验证相邻两步的搜索方向正交. (2) 牛顿法进行一次迭代.

10. 试用共轭梯度法求解线性方程组

$$\begin{cases} 3x_1 - 2x_2 + x_3 = -2, \\ -2x_1 + 2x_2 + x_3 = 1, \\ x_1 + x_2 + 6x_3 = -3, \end{cases}$$

初始点 $\mathbf{x}_0 = (0,0,0)^{\mathrm{T}}$, 要求计算过程中无舍入误差, 并验证算法经过三次迭代后终止.

11. 试用 BFGS 校正公式的拟牛顿法求函数

$$f(\mathbf{x}) = x_1^2 + 4x_2^2 - 4x_1 - 8x_2, \quad \mathbf{x} = (x_1, x_2)^{\mathrm{T}} \in \mathbb{R}^2,$$

的极小点 $\mathbf{x}^*$, 初始点 $\mathbf{x}_0 = (0,0)^{\mathrm{T}}$, 初始矩阵 $H_0 = I$. 验证直线 $\mathbf{x} = \mathbf{x}_1 + \lambda_1 \mathbf{p}_1$ 经过 $\mathbf{x}^*$, 算法执行两步即终止.

12. 设 $f(\mathbf{x}) = \dfrac{1}{2}\mathbf{x}^{\mathrm{T}} A\mathbf{x} - \mathbf{b}^{\mathrm{T}}\mathbf{x}$, 其中 A 是 n 阶正定对称方阵, $\mathbf{b}$ 是 n 维列向量, $\mathbf{x}^*$ 是最优化问题

$$\min\ f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n$$

的解, 即 $A\mathbf{x}^* = \mathbf{b}$. 若采用精确一维搜索的最速下降法来求解上述问题产生的序列记为 $\{\mathbf{x}_k\}$, 证明:

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_A^2 \leqslant \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 \|\mathbf{x}_k - \mathbf{x}^*\|_A^2 \Longleftrightarrow \frac{f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)}{f(\mathbf{x}_k) - f(\mathbf{x}^*)} \leqslant \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2,$$

其中 $\|\mathbf{x}\|_A^2 = \mathbf{x}^{\mathrm{T}} A\mathbf{x}$, λ_1 和 λ_n 分别是 A 的最小和最大特征值. (提示: 先证明

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_A^2 = \left[1 - \frac{(\nabla f_k^{\mathrm{T}} \nabla f_k)^2}{(\nabla f_k^{\mathrm{T}} A \nabla f_k)(\nabla f_k^{\mathrm{T}} A^{-1} \nabla f_k)}\right] \|\mathbf{x}_k - \mathbf{x}^*\|_A^2,$$

再证明 Kantorovich 不等式

$$\frac{(\mathbf{x}^{\mathrm{T}}\mathbf{x})^2}{(\mathbf{x}^{\mathrm{T}} A\mathbf{x})(\mathbf{x}^{\mathrm{T}} A^{-1}\mathbf{x})} \geqslant \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2}, \quad \forall \mathbf{x} \in \mathbb{R}^{n\times 1},$$

最后结合上述两式即可完成证明)

13. 设 A 是一个 n 阶正定对称方阵, $\{\mathbf{x}_i\}_{i=1}^n$ 是一组 A-共轭的向量, 证明:

$$A^{-1} = \sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}}{\mathbf{x}_i^{\mathrm{T}} A \mathbf{x}_i}.$$

14. 设 $A \in \mathbb{R}^{n\times n}$ 是非奇异矩阵, $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ 是任意向量, 若

$$1 + \mathbf{v}^{\mathrm{T}} A^{-1}\mathbf{u} \neq 0,$$

证明: A 的秩一校正

$$\left(A + \mathbf{u}\mathbf{v}^{\mathrm{T}}\right)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^{\mathrm{T}}A^{-1}}{1 + \mathbf{v}^{\mathrm{T}}A^{-1}\mathbf{u}},$$

即著名的 Sherman-Morrison 定理. 利用 $H_{k+1} \longleftrightarrow B_{k+1}$, $\mathbf{s}_k \longleftrightarrow \mathbf{y}_k$ 之间的对偶关系及该定理, 试从 DFP 校正公式 (10.18) 推导出 BFGS 校正公式 (10.19).