普通高等学校“十一五”精品规划教材

数值计算方法

主　编　　韩旭里

内 容 简 介

本书旨在讲述现代科学计算中常用的数值计算方法及其理论，包括插值法、函数的最佳逼近、数值积分和数值微分、线性方程组的直接解法和迭代解法、非线性方程和方程组的数值解法、矩阵特征值问题的数值解法和常微分方程的数值解法. 每章都配有较丰富的习题和数值试验题，书末附有部分习题答案.

本书注重内容的实用性、基本思想的阐述、数值计算方法的应用. 取材精炼、叙述清晰、系统性强、数值计算的例子较多是本书的特色.

本书可作为高等院校理工科专业数值计算方法课程的教材，也可供从事科学与工程计算的科技人员学习参考.

前　言

随着计算科学的迅速发展以及在其他科学技术问题中的广泛应用，继理论方法和实验方法之后，数值计算已成为科学研究的第三种基本手段. 数值计算方法日益受到数学、计算机科学以及各种工程技术科学领域的专家和科技工作者的重视. 数值计算方法课程已经成为高等学校理工科专业的一门重要基础课程. 本书是作者根据多年数值计算方法课程教学实践的感受，根据理工科专业本科生、研究生教学的要求，在教学内容不断充实与更新的基础上编写的. 教材的内容为数值计算的基本理论与基本方法，包括插值法、函数的最佳逼近、数值积分与数值微分、线性方程组的直接解法和迭代解法、非线性方程和方程组的数值解法、矩阵特征值问题的计算与常微分方程的数值解法.

编写本教材的指导思想是：着重内容的实用性，着重基本原理和方法的基本思想的阐述，着重数值计算方法的应用能力的培养和提高. 教材在体系结构和内容取材上，致力于取材精炼、由浅入深、衔接顺畅，尽量综合国内外同类教材的优点. 在理论方法的分析和内容表述上，力求重点突出、思路清晰、脉络分明、便于理解. 在实际应用上，尽量联系问题的应用背景，各章节配备了丰富的数值计算例题、习题与数值试验题.

数值计算方法课程既像通常的数学课程那样有自身严密的科学体系，它又是一门应用性和实践性很强的课程. 希望通过本教材的学习使读者掌握数值计算的基本理论和基本方法，提高数学素养，提高应用计算机进行科学与工程计算的能力，提高应用数学与计算机解决实际问题的能力.

本教材可作为高等院校理工科专业数值计算方法课程的教材或教学参考书，也可供从事科学与工程计算的科技人员学习参考. 教材内容涉及的范围和深度具有一定的弹性，教学时可根据学生的实际情况选用. 60 学时左右可以讲授本书的主要内容，结合数值试验讲授完全部内容需 80 学时左右. 当然，使用本书

的讲授次序不是固定不变的，大致地说，可以分为两条主要线索，即按照书上各章的自然顺序，或者是按第 1,5,6,7,8 章及第 2,3,4,9 章的顺序讲授.

本书的选材和内容的叙述可能会有不当甚至错误之处，诚请读者和同行们批评指正.

编者

2008 年 **7** 月

目　　录

第1章　绪　　论

1.1　数值计算方法的研究对象和特点

数值计算方法也称数值分析，它研究用计算机求解各种数学问题的数值方法及其理论。数学学科内容十分广泛，数值计算方法属于计算数学的范畴，这里只涉及科学和工程计算中常见的数学问题，如函数的插值、逼近、离散数据的拟合、数值积分与数值微分、线性和非线性方程数值解法和矩阵特征值问题数值解法和微分方程数值解法等.

由于计算机科学与技术的迅速发展，数值计算方法的应用已经普遍深入到各个科学领域，很多复杂和大规模的计算问题都可以在计算机上进行计算，新的、有效的数值方法不断出现. 现在，科学与工程中的数值计算已经成为各门自然科学和工程技术科学研究的一种重要手段，成为与实验和理论并列的一个不可缺少的环节. 所以，数值计算方法既是一个基础性的，同时也是一个应用性的数学学科分支，与其他学科的联系十分紧密.

用数值方法求解数学问题首先要构造算法，即由运算规则（包括算术运算、逻辑运算和运算顺序）构成的完整的解题过程. 同一个数学问题可能有多种数值计算方法，但不一定都有效. 评价一个算法的好坏主要有两条标准：计算结果的精度和得到结果所付出的代价. 我们自然应该选择代价小又能满足精度要求的算法. 计算代价也称为计算复杂性，包括时间复杂性和空间复杂性. 时间复杂性好是指节省时间，主要由运算次数决定. 空间复杂性好是指节省存储量，主要由使用的数据量决定.

用计算机求数学问题的数值解不是简单地构造算法，它涉及多方面的理论问题，例如，算法的收敛性和稳定性等. 除理论分析外，一个数值方法是否有效，最终要通过大量的数值实验来检验. 数值计算方法具有理论性、实用性和实践性都很强的特点.

作为数值计算方法的基础知识，本课程不可能面面俱到. 除构造算法外，各章根据内容自身的特点，讨论的问题有所侧重. 学习时我们首先要注意掌握方法的基本原理和思想，要注意方法处理的技巧及其与计算机的结合，要重视误差分析、收敛性和稳定性的基本理论. 其次，要通过例子，学习使用各种数值方法解决

实际计算问题，熟悉数值方法的计算过程. 最后，为了掌握本课程的内容，还应做一定数量的理论分析与计算练习.

1.2 数值计算的误差

1.2.1 误差的来源

应用数学工具解决实际问题，首先，要对被描述的实际问题进行抽象、简化，得到实际问题的数学模型. 数学模型与实际问题之间会出现的误差，我们称之为模型误差. 在数学模型中，通常要包含一些由观测数据确定的参数. 数学模型中一些参数观测结果一般不是绝对准确的. 我们把观测模型参数值产生的误差称为观测误差. 例如，设一根铝棒在温度 t 时的实际长度为 L_t，在 $t=0$ 时的实际长度为 L_0，用 l_t 来表示铝棒在温度为 t 时的长度计算值，并建立一个数学模型

$$l_t=L_0(1+at),\ a\approx 0.000\ 023\ 8/℃,$$

其中 a 是由实验观测得到的常数，$a\in[0.000\ 023\ 7,0.000\ 023\ 9]$，则称 L_t-l_t 为模型误差，$a-0.000\ 023\ 8$ 是 a 的观测误差.

在解实际问题时，数学模型往往很复杂，因而不易获得分析解，这就需要建立一套行之有效的近似方法和数值方法. 我们可能用容易计算的问题代替不易计算的问题而产生误差，也可能用有限的过程代替无限的过程而产生误差. 我们将模型的准确解与用数值方法求得的准确解之间的误差称为截断误差或方法误差. 例如，对函数

$$\sin x=x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\cdots+(-1)^n\frac{x^{2n+1}}{(2n+1)!}+\cdots,$$

该式右边有无限多项，计算机上无法计算. 然而，根据微积分学中的泰勒(Taylor)定理，当 $|x|$ 较小时，我们若用前 3 项作为 $\sin x$ 的近似值，则截断误差的绝对值不超过 $\frac{|x|^7}{7!}$.

用计算机做数值计算时，一般也不能获得数值计算公式的准确解，需要对原始数据、中间结果和最终结果取有限位数字. 我们将计算过程中取有限位数字进行运算而引起的误差称为舍入误差. 例如，$\frac{1}{3}=0.333\ 33\cdots$，如果我们取小数点后 4 位数字，则 $\frac{1}{3}-0.333\ 3=0.000\ 033\cdots$ 就是舍入误差.

在数值分析中，除了研究数学问题的算法外，还要研究计算结果的误差是否满足精度要求，这就是误差估计问题. 在数值计算方法中，主要讨论的是截断误

差和舍入误差.

1.2.2　误差与有效数字

定义 1.1　设 x 是某实数的精确值，x_A 是它的一个近似值，则称 $x-x_A$ 为近似值 x_A 的绝对误差，或简称误差. $\dfrac{x-x_A}{x}$ 称为 x_A 的相对误差.

当 $x=0$ 时，相对误差没有意义. 在实际计算中，精确值 x 往往是不知道的，所以通常把 $\dfrac{x-x_A}{x_A}$ 作为 x_A 的相对误差.

定义 1.2　设 x 是某实值的精确值，x_A 是它的一个近似值，并可对 x_A 的绝对误差作估计 $|x-x_A|\leqslant\varepsilon_A$，则称 ε_A 是 x_A 的绝对误差界，或简称误差界. 称 $\dfrac{\varepsilon_A}{|x_A|}$ 是 x_A 的相对误差界.

例 1.1　我们知道 $\pi=3.141\,592\,6\cdots$，若取近似值 $\pi_A=3.14$，则 $\pi-\pi_A=0.001\,592\,6\cdots$，可以估计绝对误差界为 0.002，相对误差界为 $0.000\,6$.

例 1.2　测量一木板长是 954 cm，问测量的相对误差界是多大？

解　因为实际问题中所截取的近似数，其绝对误差界一般不超过最小刻度的半个单位，所以当 $x=954$ cm 时，有 $\varepsilon_A=0.5$ cm，其相对误差界为

$$\frac{\varepsilon_A}{|x|}=\frac{0.5}{954}=0.000\,524\,1\cdots<0.053\%.$$

定义 1.3　设 x_A 是 x 的一个近似值，将 x_A 写成

$$x_A=\pm10^k\times0.a_1a_2\cdots a_i\cdots, \tag{1.1}$$

它可以是有限或无限小数的形式，其中 $a_i(i=1,2,\cdots)$ 是 $0,1,\cdots,9$ 中的一个数字，$a_1\neq0$，k 为整数. 如果

$$|x-x_A|\leqslant0.5\times10^{k-n},$$

则称 x_A 为 x 的具有 n 位有效数字的近似值.

可见，若近似值 x_A 的误差界是某一位的半个单位，该位到 x_A 的第一位非零数字共有 n 位，则 x_A 有 n 位有效数字.

通常在 x 的准确值已知的情况下，若要取有限位数的数字作为近似值，就采用四舍五入的原则，不难验证，采用四舍五入得到的近似值，其绝对误差界可以取为被保留的最后数位上的半个单位. 例如

$$|\pi-3.14|\leqslant0.5\times10^{-2},\ |\pi-3.142|\leqslant0.5\times10^{-3}.$$

按定义，3.14 和 3.142 分别是具有 3 位和 4 位有效数字的近似值.

显然，近似值的有效数字位数越多，相对误差界就越小，反之也对. 下面，我们给出相对误差界与有效数字的关系.

定理 1.1 设 x 的近似值 x_A 有(1.1)式的表达式.

(1) 如果 x_A 有 n 位有效数字,则

$$\frac{|x-x_A|}{|x_A|}\leqslant\frac{1}{2a_1}\times10^{1-n};\tag{1.2}$$

(2) 如果

$$\frac{|x-x_A|}{|x_A|}\leqslant\frac{1}{2(a_1+1)}\times10^{1-n},\tag{1.3}$$

则 x_A 至少具有 n 位有效数字.

证 由(1.1)式可得到

$$a_1\times10^{k-1}\leqslant|x_A|\leqslant(a_1+1)\times10^{k-1}.\tag{1.4}$$

所以,当 x_A 有 n 位有效数字时

$$\frac{|x-x_A|}{|x_A|}\leqslant\frac{0.5\times10^{k-n}}{a_1\times10^{k-1}}=\frac{1}{2a_1}\times10^{1-n},$$

即(1.2)式得证.

由(1.3)式和(1.4)式有

$$|x-x_A|\leqslant(a_1+1)\times10^{k-1}\times\frac{1}{2(a_1+1)}\times10^{1-n}=0.5\times10^{k-n},$$

即说明 x_A 有 n 位有效数字,(2)得证.

例 1.3 要使 $\sqrt{20}$ 的近似值的相对误差界小于 0.1%,应取几位有效数字?

解 由于 $4<\sqrt{20}<5$,因此 $a_1=4$,设有 n 位有效数字,则由(1.2)式,可令

$$\frac{1}{2a_1}\times10^{1-n}\leqslant0.1\%,$$

即 $10^{n-4}\geqslant\frac{1}{8}$,得 $n\geqslant4$. 故只要对 $\sqrt{20}$ 的近似数取 4 位有效数字,其相对误差就可小于 0.1%,因此,可取 $\sqrt{20}\approx4.472$.

例 1.4 已知近似数 x_A 的相对误差界为 0.3%,问 x_A 至少有几位有效数字?

解 设 x_A 有 n 位有效数字,由于 x_A 的第一个有效数 a_1 没有具体给定,而我们知道 a_1 一定是 1,2,…,9 中的一个,由于

$$\frac{|x-x_A|}{x_A}\leqslant\frac{3}{1\,000}<\frac{1}{2\times10^2}=\frac{1}{2(9+1)}\times10^{-1},$$

故由(1.3)式知 $n=2$,即 x_A 至少有 2 位有效数字.

1.2.3 函数求值的误差估计

对一元函数 $f(x)$,自变量 x 的一个近似值为 x_A,以 $f(x_A)$ 近似 $f(x)$,其误

差界记作 $\varepsilon(f(x_A))$. 若 $f(x)$ 具有二阶连续导数，$f'(x_A)$ 与 $f''(x_A)$ 的比值不太大，则可忽略 $|x-x_A|$ 的二次项，由 Taylor 展开式得到 $f(x_A)$ 的一个近似误差界

$$\varepsilon(f(x_A))\approx|f'(x_A)|\varepsilon(x_A).$$

对 n 元函数 $f(x_1,x_2,\cdots,x_n)$，自变量 $x_1,x_2,\cdots,x_n$ 的近似值分别为 x_{1A}，$x_{2A},\cdots,x_{nA}$，则有

$$f(x_1,x_2,\cdots,x_n)-f(x_{1A},x_{2A},\cdots,x_{nA})\approx\sum_{k=1}^{n}\left(\frac{\partial f}{\partial x_k}\right)_A(x_k-x_{kA}),$$

其中 $\left(\dfrac{\partial f}{\partial x_k}\right)_A=\dfrac{\partial}{\partial x_k}f(x_{1A},x_{2A},\cdots,x_{nA})$. 因此，可以得到函数值的一个近似误差界

$$\varepsilon(f(x_{1A},x_{2A},\cdots,x_{nA}))\approx\sum_{k=1}^{n}\left|\left(\frac{\partial f}{\partial x_k}\right)_A\right|\varepsilon(x_{kA}).$$

特别地，对 $f(x_1,x_2)=x_1\pm x_2$ 有

$$\varepsilon(x_{1A}\pm x_{2A})=\varepsilon(x_{1A})+\varepsilon(x_{2A}).$$

同样，可以得到

$$\varepsilon(x_{1A}x_{2A})\approx|x_{1A}|\varepsilon(x_{2A})+|x_{2A}|\varepsilon(x_{1A}),$$

$$\varepsilon\left(\frac{x_{1A}}{x_{2A}}\right)\approx\frac{|x_{1A}|\varepsilon(x_{2A})+|x_{2A}|\varepsilon(x_{1A})}{|x_{2A}|^2},\quad x_{2A}\neq0.$$

例 1.5　设有长为 l，宽为 d 的某场地. 现测得 l 的近似值 $l_A=120$ m，d 的近似值 $d_A=90$ m，并已知它们的误差界为 $|l-l_A|\leqslant0.2$ m，$|d-d_A|\leqslant0.2$ m. 试估计该场地面积 $S=ld$ 的误差界和相对误差界.

解　这里 $\varepsilon(l_A)=0.2$，$\varepsilon(d_A)=0.2$，并且有

$$\frac{\partial S}{\partial l}=d,\frac{\partial S}{\partial d}=l,\ S_A=l_Ad_A=10\,800\ \text{m}^2.$$

于是有误差界

$$\varepsilon(S_A)\approx120\times0.2+90\times0.2=42\ \text{m}^2,$$

相对误差界

$$\varepsilon_r(S_A)=\frac{\varepsilon(S_A)}{l_Ad_A}\approx\frac{42}{10\,800}=0.39\%.$$

例 1.6　设有 3 个近似数

$$a=2.31,\ b=1.93,\ c=2.24,$$

它们都有 3 位有效数字. 试计算 $p=a+bc$ 的误差界和相对误差界，并问 p 的计算结果能有几位有效数字？

解　$p=2.31+1.93\times2.24=6.6332$. 于是有误差界

$$\varepsilon(p)=\varepsilon(a)+\varepsilon(bc)$$

$$\approx \varepsilon(a)+|b|\varepsilon(c)+|c|\varepsilon(b)$$
$$=0.005+0.005(1.93+2.24)=0.025\,85,$$

相对误差界

$$\varepsilon_r(p)=\frac{\varepsilon(p)}{|p|}\approx\frac{0.025\,85}{6.633\,2}\approx 0.39\%.$$

因为 $\varepsilon(p)\approx 0.025\,85<0.05$，所以 $p=6.633\,2$ 能有 2 位有效数字.

1.2.4 计算机中数的表示

任意一个非零实数用(1.1)式表示，是规格化的十进制科学记数方法. 在计算机中通常采用二进制的数系(或其变形的十六进制等)，并且表示成与十进制类似的规格化形式，即浮点形式

$$\pm 2^m\times 0.\beta_1\beta_2\cdots\beta_t,$$

这里整数 m 称为阶码，用二进制表示为 $m=\pm\alpha_1\alpha_2\cdots\alpha_s$，$\alpha_j=0$ 或 $1(j=1,2,\cdots,s)$，s 是阶的位数. 小数 $0.\beta_1\beta_2\cdots\beta_t$ 称为尾数，其中 $\beta_1=1,\beta_j=0$ 或 $1(j=2,3,\cdots,t)$，t 是尾数部位的位数. s 和 t 与具体的机器有关. 由于计算机的字长总是有限位的，所以计算机所能表示的数系是一个特殊的离散集合，此集合的数称为机器数. 用浮点方式表示的数有比较大的取值范围.

十进制输入计算机时转换成二进制，并对 t 位后面的数作舍入处理，使得尾数为 t 位，因此一般都有舍入误差. 两个二进制数作算术运算时，对计算结果也要作类似的舍入处理，使得尾数为 t 位，从而也有舍入误差.

在实现算法时，计算的最后结果与算法的精确解之间的误差，从根本上说是由机器的舍入误差造成的，包括输入数据和算术运算的舍入误差. 因此有必要对计算机中数的浮点表示方法和舍入误差有一个初步的了解. 有时为了分析某一个计算方法可能出现的误差现象，为了适应人们的习惯，我们会采用十进制实数系统进行误差分析.

1.3 数值稳定性和要注意的若干原则

1.3.1 数值方法的稳定性

实际计算时，给定的数据会有误差，数值计算中也会产生误差，并且，这些误差在进一步的计算中会有误差传播. 因此，尽管数值计算中的误差估计比较困难，我们还是应该重视计算过程中的误差分析.

定义 **1.4** 对于某个数值计算方法，如果输入数据的误差在计算过程中迅

速增长而得不到控制，则称该算法是数值不稳定的，否则是数值稳定的.

下面举例说明误差传播的现象.

例 **1.7**　计算积分值 $I_n=\int_0^1 \frac{x^n}{x+5}\mathrm{d}x$，$n=0,1,\cdots,6$.

解　由于要计算系列的积分值，我们先推导 I_n 的一个递推公式. 由

$$\begin{aligned} I_n+5I_{n-1} &= \int_0^1 \frac{x^n+5x^{n-1}}{x+5}\mathrm{d}x \\ &= \int_0^1 x^{n-1}\mathrm{d}x=\frac{1}{n}, \end{aligned}$$

可得下面两个递推算法.

算法 **1**：　$I_n=\frac{1}{n}-5I_{n-1}$，　$n=1,2,\cdots,6$.

算法 **2**：　$I_{n-1}=\frac{1}{5}\left(\frac{1}{n}-I_n\right)$，　$n=6,5,\cdots,1$.

直接计算可得 $I_0=\ln 6-\ln 5$. 如果我们用 4 位数字计算，得 I_0 的近似值为 $I_0^*=0.1823$. 记 $E_n=I_n-I_n^*$，I_n^* 为 I_n 的近似值.

对算法 1，有

$$E_n=-5E_{n-1}=\cdots=(-5)^nE_0.$$

按以上初始值 I_0 的取法有 $|E_0|\leqslant 0.5\times 10^{-4}$，事实上 $|E_0|\approx 0.22\times 10^{-4}$. 这样，我们得到 $|E_6|=5^6|E_0|\approx 0.34$. 这个数已经大大超过了 I_6 的大小，所以 I_6^* 连一位有效数字也没有了，误差掩盖了真值.

对算法 2，有

$$E_{k-n}=\left(-\frac{1}{5}\right)^nE_k,$$

$$|E_0|=\left(\frac{1}{5}\right)^6|E_6|.$$

如果我们能够给出 I_6 的一个近似值，则可由算法 2 计算 $I_n(n=5,4,\cdots,0)$ 的近似值. 并且，即使 E_6 较大，得到的近似值的误差将较小. 由于

$$\frac{1}{6(k+1)}=\int_0^1\frac{x^k}{6}\mathrm{d}x<I_k<\int_0^1\frac{x^k}{5}\mathrm{d}x=\frac{1}{5(k+1)},$$

因此，可取 I_k 的一个近似值为

$$I_k^*=\frac{1}{2}\left[\frac{1}{6(k+1)}+\frac{1}{5(k+1)}\right].$$

对 $k=6$ 有 $I_6^*=0.0262$.

按 $I_0^*=0.1823$ 和 $I_6^*=0.0262$，分别按算法 1 和算法 2 计算，计算结果如表 1-1，其中 $I_n^{(1)}$ 为算法 1 的计算值，$I_n^{(2)}$ 为算法 2 的计算值. 易知，对于任何自然

数 n，都有 $0<I_n<1$，并且 I_n 单调减. 可见，算法 1 是不稳定的，算法 2 是稳定的.

表 1－1

n	$I_n^{(1)}$	$I_n^{(2)}$	I_n（4 位）
0	0.182 3	0.182 3	0.182 3
1	0.088 5	0.088 4	0.088 4
2	0.057 5	0.058 0	0.058 0
3	0.045 8	0.043 1	0.043 1
4	0.021 0	0.034 4	0.034 3
5	0.095 0	0.028 1	0.028 5
6	－0.308 3	0.026 2	0.024 3

当然，数值不稳定的方法一般在实际计算中不能采用. 数值不稳定的现象属于误差危害现象. 下面讨论误差危害现象的其他表现及如何避免问题.

1.3.2 避免有效数字的损失

在数值计算中，参加运算的数有时数量级相差很大，而计算机位数有限，如不注意，“小数”的作用可能消失，即出现“大数”吃“小数”的现象.

例 1.8 用 3 位十进制数字计算

$$x = 101 + \delta_1 + \delta_2 + \cdots + \delta_{100},$$

其中 $0.1 \leqslant \delta_i \leqslant 0.4, i = 1,2,\cdots,100$.

解 在计算机内计算时，要写成浮点数形式，且要对阶. 如果是 101 与 δ_1 相加，对阶时，$101 = 0.101 \times 10^3$，$\delta_1 = 0.000 \times 10^3$. 因此，如果我们自左至右逐个相加，则所有的 δ_i 都会被舍掉，得 $x \approx 101$. 但若把所有的 δ_i 先加起来，再与 101 相加，就有

$$111 = 101 + 100 \times 0.1 \leqslant x \leqslant 101 + 100 \times 0.4 = 141.$$

可见，计算的次序会产生很大的影响. 这是因为用计算机计算时，在运算中要“对阶”，对阶引起了大数吃小数的现象. 大数吃小数在有些情况下是允许的，但有些情况下则会造成谬误.

在数值计算中，两个相近数相减会使有效数字严重损失.

例 1.9 求实系数二次方程 $ax^2 + bx + c = 0$ 的根，其中 $b^2 - 4ac > 0, ab \neq 0$.

解 考虑两种算法.

算法 1： $x_{1,2} = \dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

算法 2：$x_1=\dfrac{-b-\mathrm{sign}(b)\sqrt{b^2-4ac}}{2a}$，$x_2=\dfrac{c}{ax_1}$，

其中 sign 表示取数的符号，即

$$\mathrm{sign}(b)=\begin{cases}1, & b>0,\\ 0, & b=0,\\ -1, & b<0.\end{cases}$$

对算法 1，若 $b^2\gg 4ac$，则是不稳定的，否则是稳定的. 这是因为在算法 1 中分子会有相近数相减的情形，会造成有效数字的严重损失，从而结果的误差很大. 算法 2 不存在这个问题，在任何情况下都是稳定的. 因此称算法 1 是条件稳定的，算法 2 是无条件稳定的. 例如，对于方程

$$x^2+62.10x+1.000=0,$$

用 4 位有效数字计算，结果如下：

算法 1：$x_1=-62.08$，$x_2=-0.020\,00$.

算法 2：$x_1=-62.08$，$x_2=-0.016\,11$.

准确解是 $x_1=-62.083\,892\cdots$，$x_2=-0.016\,107\,237\cdots$. 这里，$b^2\gg 4ac$，所以算法 1 不稳定，舍入误差对 x_2 的影响大.

在进行数值计算时，如果遇到两相近数相减的情形，可通过变换计算公式来避免或减少有效数字的损失. 例如，如果 $|x|\approx 0$，有变换公式

$$\frac{1-\cos x}{\sin x}=\frac{\sin x}{1+\cos x}.$$

如果 $x_1\approx x_2$，有变换公式

$$\lg x_1-\lg x_2=\lg\frac{x_1}{x_2}.$$

如果 $x\gg 1$，有变换公式

$$\sqrt{x+1}-\sqrt{x}=\frac{1}{\sqrt{x+1}+\sqrt{x}}.$$

此外，用绝对值很小的数作除数时，舍入误差会很大，可能对计算结果带来严重影响. 因此，要避免除数绝对值远远小于被除数绝对值的除法运算.

如果无法改变算法，则采用增加有效位数进行计算，或在计算上采用双精度运算，但这要增加机器计算时间和多占内存单元.

1.3.3　减少运算次数

在数值计算中，要注意简化计算步骤，减少运算次数，这也是数值分析中所要研究的重要内容. 同样一个计算问题，如果能减少运算次数，不但可节省计算

机的计算时间，还能减少误差的积累. 下面举例说明简化计算公式的重要性.

例 **1.10** 给定 x，计算多项式

$$P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$$

的值. 如果我们先求 $a_k x^k$，需要进行 k 次乘法，再相加，则总共需要 $\frac{n(n+1)}{2}$ 次乘法和 n 次加法才能得到一个多项式的值. 如果我们将多项式写成下面的形式

$$P_n(x) = x\{x\cdots[x(a_n x + a_{n-1}) + a_{n-2}] + \cdots + a_1\} + a_0,$$

则只需 n 次乘法和 n 次加法即可得到一个多项式的值，这就是著名的秦九韶算法，可描述为

$$\begin{cases} u_n = a_n, \\ u_k = u_{k+1}x + a_k, \quad k = n-1, n-2, \cdots, 0, \end{cases}$$

最后有 $u_0 = P_n(x)$.

例 **1.11** 计算 ln2 的值.

解 如果利用级数

$$\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n}$$

计算 ln2，若要精确到误差的绝对值小于 10^{-5}，要计算 10 万项求和，计算量很大，并且舍入误差的积累也十分严重. 如果改用级数

$$\ln \frac{1+x}{1-x} = 2\left(x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + \frac{x^{2n+1}}{(2n+1)!} + \cdots\right)$$

来计算 ln2，取 $x = \frac{1}{3}$，则只要计算前 9 项，截断误差便小于 10^{-10}.

1.4 向量和矩阵的范数

为了对矩阵计算进行数值分析，我们需要对向量和矩阵的“大小”引进某种度量. 在解析几何中，向量的大小和两个向量之差的大小是用“长度”和“距离”的概念来度量的. 在实数域中，数的大小和两个数之间的距离是通过绝对值来度量的. 范数是绝对值概念的自然推广.

1.4.1 向量的范数

定义 **1.5** 如果向量 $\boldsymbol{x} \in \mathbf{R}^n$ 的某个实值函数 $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ 满足

(1) 正定性：$\|\boldsymbol{x}\| \geqslant 0$，且 $\|\boldsymbol{x}\| = 0$ 当且仅当 $\boldsymbol{x} = \boldsymbol{0}$；

(2) 齐次性：对任意实数 α，都有 $\|\alpha\boldsymbol{x}\| = |\alpha| \, \|\boldsymbol{x}\|$；

(3) 三角不等式：对任意 $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^n$，都有

$$\| \boldsymbol{x}+\boldsymbol{y} \| \leqslant \| \boldsymbol{x} \| + \| \boldsymbol{y} \|,$$

则称 $\| \boldsymbol{x} \|$ 为 $\mathbf{R}^n$ 上的一个向量范数.

在 $\mathbf{R}^n$ 中,记 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$,实际计算中最常用的向量范数有:

(1) 向量的 ∞ 范数

$$\| \boldsymbol{x} \|_{\infty} = \max_{1\leqslant i\leqslant n} | x_i |;$$

(2) 向量的 1 范数

$$\| \boldsymbol{x} \|_1 = \sum_{i=1}^{n} | x_i |;$$

(3) 向量的 2 范数

$$\| \boldsymbol{x} \|_2 = \Big(\sum_{i=1}^{n} x_i^2 \Big)^{\frac{1}{2}}.$$

容易验证,向量的 ∞ 范数和 1 范数满足定义 1.5 中的条件. 对于 2 范数,满足定义 1.5 中的条件(1) 和(2) 是显然的,对于条件(3),利用向量内积的 Cauchy-Schwarz 不等式可以验证. 更一般地,有如下向量的 p 范数

$$\| \boldsymbol{x} \|_p = \Big(\sum_{i=1}^{n} | x_i |^p \Big)^{\frac{1}{p}},$$

其中 $p\in[1,+\infty)$.

容易验证

$$\| \boldsymbol{x} \|_{\infty} \leqslant \| \boldsymbol{x} \|_p \leqslant n^{\frac{1}{p}} \| \boldsymbol{x} \|_{\infty},$$

由此可得如下定理.

定理 1.2　$\lim\limits_{p\to\infty} \| \boldsymbol{x} \|_p = \| \boldsymbol{x} \|_{\infty}$.

下面,我们利用向量范数的连续性来说明向量范数的重要特征.

定理 1.3　设给定 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$,$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}\in\mathbf{R}^n$,则对 $\mathbf{R}^n$ 上每一种向量范数,$\| \boldsymbol{Ax} \|$ 都是 $x_1,x_2,\cdots,x_n$ 的 n 元连续函数.

证　设 $\boldsymbol{a}_j$ 为 $\boldsymbol{A}$ 的列向量,将 $\boldsymbol{A}$ 写成

$$\boldsymbol{A}=(\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n).$$

则由三角不等式,对 $\boldsymbol{h}=(h_1,h_2,\cdots,h_n)^{\mathrm{T}}\in\mathbf{R}^n$,有

$$\begin{aligned} | \, \| \boldsymbol{A}(\boldsymbol{x}+\boldsymbol{h}) \| - \| \boldsymbol{Ax} \| \, | &\leqslant \| \boldsymbol{Ah} \| = \Big\| \sum_{i=1}^{n} h_i \boldsymbol{a}_i \Big\| \\ &\leqslant \sum_{i=1}^{n} | h_i | \, \| \boldsymbol{a}_i \| \\ &\leqslant M \max_i | h_i |, \end{aligned}$$

其中 $M=\sum\limits_{i=1}^{n} \| \boldsymbol{a}_i \|$. 所以,对任意的 $\varepsilon>0$,当 $\max\limits_i | h_i | < \dfrac{\varepsilon}{M}$ 时,有

$$| \| \boldsymbol{A}(\boldsymbol{x}+\boldsymbol{h}) \| - \| \boldsymbol{A}\boldsymbol{x} \| | < \varepsilon,$$

这就证明了 $\| \boldsymbol{A}\boldsymbol{x} \|$ 的连续性.

推论 **1.1** $\| \boldsymbol{x} \|$ 是 $\boldsymbol{x}$ 的各分量的连续函数.

向量范数的一个重要特征是具有等价性.

定理 **1.4** $\mathbf{R}^n$ 上的所有向量范数是彼此等价的,即对 $\mathbf{R}^n$ 上的任意两种向量范数 $\| \boldsymbol{x} \|_s$ 和 $\| \boldsymbol{x} \|_t$,存在常数 $c_1, c_2 > 0$,使得对任意 $\boldsymbol{x}$,有

$$c_1 \| \boldsymbol{x} \|_s \leqslant \| \boldsymbol{x} \|_t \leqslant c_2 \| \boldsymbol{x} \|_s.$$

证 只要就 $\| \boldsymbol{x} \|_s = \| \boldsymbol{x} \|_\infty$ 证明上式成立即可,即证明存在常数 c_1, $c_2 > 0$,对一切 $x \in \mathbf{R}^n$ 且 $\boldsymbol{x} \neq \boldsymbol{0}$,有

$$c_1 \| \boldsymbol{x} \|_\infty \leqslant \| \boldsymbol{x} \|_t \leqslant c_2 \| \boldsymbol{x} \|_\infty.$$

记 $\mathbf{R}^n$ 上的有界闭集

$$D = \{\boldsymbol{x} : \boldsymbol{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}, \| \boldsymbol{x} \|_\infty = 1\}.$$

由定理 1.3 的推论知,$\| \boldsymbol{x} \|_t$ 是 D 上的 n 元连续函数,所以在 D 上有最大值 c_2 和最小值 c_1,且 $\boldsymbol{x} \in D$ 时有 $\boldsymbol{x} \neq 0$,故有 $c_2 \geqslant c_1 > 0$. 现考虑 $\boldsymbol{x} \in \mathbf{R}^n$,且 $\boldsymbol{x} \neq \boldsymbol{0}$,则有 $\dfrac{\boldsymbol{x}}{\| \boldsymbol{x} \|_\infty} \in D$,所以有

$$c_1 \leqslant \left\| \frac{\boldsymbol{x}}{\| \boldsymbol{x} \|_\infty} \right\|_t \leqslant c_2, \quad \forall \boldsymbol{x} \in \mathbf{R}^n, \boldsymbol{x} \neq \boldsymbol{0}.$$

从而对 $\boldsymbol{x} \neq \boldsymbol{0}$ 有

$$c_1 \| \boldsymbol{x} \|_\infty \leqslant \| \boldsymbol{x} \|_t \leqslant c_2 \| \boldsymbol{x} \|_\infty.$$

而 $\boldsymbol{x} = \boldsymbol{0}$ 时上式自然成立,定理得证.

由于向量范数之间具有等价性,对于范数的极限性质,我们只需对一种范数进行讨论,其余范数也都具有相似的结论. 比如,我们可以方便地讨论向量序列的收敛性.

定义 **1.6** 设向量序列 $\boldsymbol{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, \cdots, x_n^{(k)})^{\mathrm{T}} \in \mathbf{R}^n, k = 1, 2, \cdots$,若存在 $\boldsymbol{x}^* = (x_1^*, x_2^*, \cdots, x_n^*)^{\mathrm{T}} \in \mathbf{R}^n$,使得

$$\lim_{k \to \infty} x_i^{(k)} = x_i^*, \ i = 1, 2, \cdots, n,$$

则称序列 $\{\boldsymbol{x}^{(k)}\}$ 收敛于 $\boldsymbol{x}^*$,记为

$$\lim_{k \to \infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}^*.$$

按定义有

$$\lim_{k \to \infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}^* \Leftrightarrow \lim_{k \to \infty} \| \boldsymbol{x}^{(k)} - \boldsymbol{x}^* \|_\infty = 0.$$

又因为

$$c_1 \| \boldsymbol{x}^{(k)} - \boldsymbol{x}^* \|_\infty \leqslant \| \boldsymbol{x}^{(k)} - \boldsymbol{x}^* \| \leqslant c_2 \| \boldsymbol{x}^{(k)} - \boldsymbol{x}^* \|_\infty,$$

所以有

$$\lim_{k\to\infty}\boldsymbol{x}^{(k)}=\boldsymbol{x}^{*}\Leftrightarrow\lim_{k\to\infty}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{*}\|=0.$$

因此，若向量序列在一种范数下收敛，则在其他范数下也收敛. 不必强调是在哪种范数意义下收敛.

1.4.2　矩阵的范数

定义 1.7　如果矩阵 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 的某个实值函数 $f(\boldsymbol{A})=\|\boldsymbol{A}\|$ 满足

(1) 正定性：$\|\boldsymbol{A}\|\geqslant 0$，且 $\|\boldsymbol{A}\|=0$ 当且仅当 $\boldsymbol{A}=0$；

(2) 齐次性：对任意实数 α，都有 $\|\alpha\boldsymbol{A}\|=|\alpha|\,\|\boldsymbol{A}\|$；

(3) 三角不等式：对任意 $\boldsymbol{A},\boldsymbol{B}\in\mathbf{R}^{n\times n}$，都有 $\|\boldsymbol{A}+\boldsymbol{B}\|\leqslant\|\boldsymbol{A}\|+\|\boldsymbol{B}\|$；

(4) 相容性：对任意 $\boldsymbol{A},\boldsymbol{B}\in\mathbf{R}^{n\times n}$，都有 $\|\boldsymbol{AB}\|\leqslant\|\boldsymbol{A}\|\,\|\boldsymbol{B}\|$；

则称 $\|\boldsymbol{A}\|$ 为 $\mathbf{R}^{n\times n}$ 上的一个矩阵范数.

可以验证，对 $\boldsymbol{A}=(a_{ij})_{n\times n}$，

$$\|\boldsymbol{A}\|_{\mathrm{F}}=\Big(\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}^{2}\Big)^{\frac{1}{2}}$$

是一种矩阵范数，称之为 **Frobenius** 范数，简称 **F** 范数.

由于矩阵与向量常常同时参与讨论与计算，矩阵范数与向量范数之间需要有一种联系.

定义 1.8　对于给定的 $\mathbf{R}^{n}$ 上的一种向量范数 $\|\boldsymbol{x}\|$ 和 $\mathbf{R}^{n\times n}$ 上的一种矩阵范数 $\|\boldsymbol{A}\|$，如果满足

$$\|\boldsymbol{Ax}\|\leqslant\|\boldsymbol{A}\|\,\|\boldsymbol{x}\|,$$

则称矩阵范数 $\|\boldsymbol{A}\|$ 与向量范数 $\|\boldsymbol{x}\|$ 相容.

上面的定义 1.7 是矩阵范数的一般定义，下面我们通过已给的向量范数来定义与之相容的矩阵范数.

定义 1.9　设 $\boldsymbol{x}\in\mathbf{R}^{n}$，$\boldsymbol{A}\in\mathbf{R}^{n\times n}$，对给出的一种向量范数 $\|\boldsymbol{x}\|_{v}$，相应地定义一个矩阵的非负函数

$$\|\boldsymbol{A}\|_{v}=\max_{x\neq 0}\frac{\|\boldsymbol{Ax}\|_{v}}{\|\boldsymbol{x}\|_{v}}.$$

称之为由向量范数导出的矩阵范数，也称为算子范数或从属范数.

由定义可得

$$\|\boldsymbol{Ax}\|_{v}\leqslant\|\boldsymbol{A}\|_{v}\,\|\boldsymbol{x}\|_{v},\quad\|\boldsymbol{A}\|_{v}=\max_{\|\boldsymbol{x}\|_{v}=1}\|\boldsymbol{Ax}\|_{v}.$$

算子范数满足矩阵范数一般定义中的条件(1) 和(2) 是显然的，现验证满足条件(3) 和(4).

对任意的 $\boldsymbol{A},\boldsymbol{B}\in\mathbf{R}^{n\times n}$，有

$$\|\boldsymbol{A}+\boldsymbol{B}\|_v=\max_{\|\boldsymbol{x}\|_v=1}\|(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{x}\|_v$$
$$\leqslant\max_{\|\boldsymbol{x}\|_v=1}\|\boldsymbol{A}\boldsymbol{x}\|_v+\max_{\|\boldsymbol{x}\|_v=1}\|\boldsymbol{B}\boldsymbol{x}\|_v=\|\boldsymbol{A}\|_v+\|\boldsymbol{B}\|_v,$$
$$\|\boldsymbol{A}\boldsymbol{B}\|_v=\max_{\|\boldsymbol{x}\|_v=1}\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{x}\|_v$$
$$\leqslant\max_{\|\boldsymbol{x}\|_v=1}\|\boldsymbol{A}\|_v\|\boldsymbol{B}\|_v\|\boldsymbol{x}\|_v=\|\boldsymbol{A}\|_v\|\boldsymbol{B}\|_v.$$

因此,算子范数满足矩阵范数一般定义中的条件(3)和(4).

由常用的向量范数,可以导出与其相容的矩阵算子范数.

定理 1.5 设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$,记 $\boldsymbol{A}=(a_{ij})_{n\times n}$,则

(1) $\|\boldsymbol{A}\|_\infty=\max\limits_{1\leqslant i\leqslant n}\sum\limits_{j=1}^{n}|a_{ij}|$,称之为矩阵 $\boldsymbol{A}$ 的行范数;

(2) $\|\boldsymbol{A}\|_1=\max\limits_{1\leqslant j\leqslant n}\sum\limits_{i=1}^{n}|a_{ij}|$,称之为矩阵 $\boldsymbol{A}$ 的列范数;

(3) $\|\boldsymbol{A}\|_2=\sqrt{\lambda_{\max}(A^{\mathrm{T}}A)}$,称之为矩阵 $\boldsymbol{A}$ 的 2 范数或谱范数,其中,$\lambda_{\max}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})$ 表示 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ 的最大特征值.

证 这里只对(1)和(3)给出证明,(2)的证明同理可得.

先证明(1):设 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}\neq0$,不妨设 $\boldsymbol{A}\neq0$,则有

$$\|\boldsymbol{A}\boldsymbol{x}\|_\infty=\max_{1\leqslant i\leqslant n}\left|\sum_{j=1}^{n}a_{ij}\boldsymbol{x}_j\right|\leqslant\|\boldsymbol{x}\|_\infty\max_{1\leqslant i\leqslant n}\sum_{j=1}^{n}|a_{ij}|.$$

于是有

$$\|\boldsymbol{A}\|_\infty=\max_{\|\boldsymbol{x}\|_\infty=1}\|\boldsymbol{A}\boldsymbol{x}\|_\infty\leqslant\max_{1\leqslant i\leqslant n}\sum_{j=1}^{n}|a_{ij}|.$$

设矩阵 $\boldsymbol{A}$ 的第 p 行元素的绝对值之和达到最大,即

$$\sum_{j=1}^{n}|a_{pj}|=\max_{1\leqslant i\leqslant n}\sum_{j=1}^{n}|a_{ij}|.$$

取向量 $\boldsymbol{\xi}=(\xi_1,\xi_2,\cdots,\xi_n)^{\mathrm{T}}$,其中

$$\xi_j=\begin{cases}1, & a_{pj}\geqslant0,\\ -1, & a_{pj}<0.\end{cases}$$

显然,$\|\boldsymbol{\xi}\|_\infty=1$,而且

$$\|\boldsymbol{A}\|_\infty=\max_{\|\boldsymbol{x}\|_\infty=1}\|\boldsymbol{A}\boldsymbol{x}\|_\infty\geqslant\|\boldsymbol{A}\boldsymbol{\xi}\|_\infty=\max_{1\leqslant i\leqslant n}\left|\sum_{j=1}^{n}a_{ij}\xi_j\right|=\sum_{j=1}^{n}|a_{pj}|.$$

于是(1)得证.

再证明(3):显然,$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ 是对称半正定矩阵,它的全部特征值均非负,设为

$$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n\geqslant0.$$

由实对称矩阵的性质,各特征值对应的特征向量必正交.设对应的标准正交特征

向量为 $\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_n$，即 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{u}_i=\lambda_i\boldsymbol{u}_i(i=1,2,\cdots,n),(\boldsymbol{u}_i,\boldsymbol{u}_j)=\delta_{ij}(i,j=1,2,\cdots,n)$.

对向量 $\boldsymbol{x}\in\mathbf{R}^n$，$\|\boldsymbol{x}\|_2=1$，可由 $\mathbf{R}^n$ 的一组基 $\boldsymbol{u}_i(i=1,2,\cdots,n)$ 线性表示，即有

$$\boldsymbol{x}=\sum_{i=1}^{n}c_i\boldsymbol{u}_i,\quad \|\boldsymbol{x}\|_2^2=\sum_{i=1}^{n}c_i^2=1,$$

于是有

$$\|\boldsymbol{A}\boldsymbol{x}\|_2^2=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}=\sum_{i=1}^{n}\lambda_i c_i^2\leqslant\lambda_1\sum_{i=1}^{n}c_i^2=\lambda_1.$$

另一方面，取 $\boldsymbol{\xi}=\boldsymbol{u}_1$，显然有 $\|\boldsymbol{\xi}\|_2=1$，

$$\|\boldsymbol{A}\boldsymbol{\xi}\|_2^2=\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\xi}=\lambda_1\boldsymbol{u}_1^{\mathrm{T}}\boldsymbol{u}_1=\lambda_1.$$

因此，$\|\boldsymbol{A}\|_2=\max\limits_{\|\boldsymbol{x}\|_2=1}\|\boldsymbol{A}\boldsymbol{x}\|_2=\sqrt{\lambda_1}$，(3) 得证.

由定理 1.5 可见，计算一个矩阵的行范数和列范数是比较容易的，而矩阵的 2 范数计算却不方便，但由于它有许多好的性质，所以在理论上还是有用的.

例 1.12　设矩阵

$$\boldsymbol{A}=\begin{bmatrix}1&-2\\-3&4\end{bmatrix},$$

计算 $\boldsymbol{A}$ 的各种算子范数.

解　$\|\boldsymbol{A}\|_\infty=\max\{3,7\}=7$，$\|\boldsymbol{A}\|_1=\max\{4,6\}=6$，

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}=\begin{bmatrix}10&-14\\-14&20\end{bmatrix},$$

$$\det(\lambda\boldsymbol{I}-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})=\begin{vmatrix}\lambda-10&14\\14&\lambda-20\end{vmatrix}=\lambda^2-30\lambda+4,$$

求得 $\lambda_1=15+\sqrt{221}$，$\lambda_2=15-\sqrt{221}$. 因此

$$\|\boldsymbol{A}\|_2=\sqrt{\lambda_1}=\sqrt{15+\sqrt{221}}\approx 5.46.$$

定义 1.10　设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 的特征值为 $\lambda_i(i=1,2,\cdots,n)$，称

$$\rho(\boldsymbol{A})=\max_{1\leqslant i\leqslant n}|\lambda_i|$$

为 $\boldsymbol{A}$ 的谱半径.

谱半径在几何上可解释为以原点为圆心，能包含 $\boldsymbol{A}$ 的全部特征值的圆的半径中最小者.

例 1.13　计算例 1.12 中矩阵的谱半径.

解　由 $\boldsymbol{A}$ 的特征方程

$$\det(\lambda \boldsymbol{I}-\boldsymbol{A})=\begin{vmatrix}\lambda-1 & 2\\ 3 & \lambda-4\end{vmatrix}=\lambda^2-5\lambda-2=0$$

得 $\lambda_1=\dfrac{5+\sqrt{33}}{2}$，$\lambda_2=\dfrac{5-\sqrt{33}}{2}$. 所以 $\rho(\boldsymbol{A})=\dfrac{5+\sqrt{33}}{2}\approx 5.37$.

定理 1.6 设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$，则有

$$\rho(\boldsymbol{A})\leqslant \|\boldsymbol{A}\|.$$

证 设 $\boldsymbol{A}\boldsymbol{x}=\lambda\boldsymbol{x}$，$\boldsymbol{x}\neq 0$，且 $|\lambda|=\rho(\boldsymbol{A})$，必存在向量 $\boldsymbol{y}$，使 $\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}$ 不是零矩阵. 于是

$$\rho(\boldsymbol{A})\|\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}\| = \|\lambda\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}\| = \|\boldsymbol{A}\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}\|\leqslant\|\boldsymbol{A}\|\,\|\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}\|,$$

即得 $\rho(\boldsymbol{A})\leqslant\|\boldsymbol{A}\|$.

例 1.14 设矩阵 $\boldsymbol{A}$ 与矩阵 $\boldsymbol{B}$ 是对称的，求证

$$\rho(\boldsymbol{A}+\boldsymbol{B})\leqslant\rho(\boldsymbol{A})+\rho(\boldsymbol{B}).$$

证 因 $\boldsymbol{A}=\boldsymbol{A}^{\mathrm{T}}$，于是有

$$\|\boldsymbol{A}\|_2^2=\lambda_{\max}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})=\lambda_{\max}(\boldsymbol{A}^2)=[\rho(\boldsymbol{A})]^2,$$

即 $\|\boldsymbol{A}\|_2=\rho(\boldsymbol{A})$. 同理 $\|\boldsymbol{B}\|_2=\rho(\boldsymbol{B})$.

由于 $\boldsymbol{A}+\boldsymbol{B}=(\boldsymbol{A}+\boldsymbol{B})^{\mathrm{T}}$，因此

$$\rho(\boldsymbol{A}+\boldsymbol{B})=\|\boldsymbol{A}+\boldsymbol{B}\|_2\leqslant\|\boldsymbol{A}\|_2+\|\boldsymbol{B}\|_2=\rho(\boldsymbol{A})+\rho(\boldsymbol{B}).$$

定理 1.7 如果 $\|\boldsymbol{B}\|<1$，则 $\boldsymbol{I}\pm\boldsymbol{B}$ 为非奇异矩阵，且

$$\|(\boldsymbol{I}\pm\boldsymbol{B})^{-1}\|\leqslant\frac{1}{1-\|\boldsymbol{B}\|},$$

这里的矩阵范数是指矩阵的算子范数.

证 若 $\boldsymbol{I}\pm\boldsymbol{B}$ 奇异，则存在向量 $\boldsymbol{x}\neq 0$，使 $(\boldsymbol{I}\pm\boldsymbol{B})\boldsymbol{x}=0$，故有 $\rho(\boldsymbol{B})\geqslant 1$，这与 $\|\boldsymbol{B}\|<1$ 矛盾，所以 $\boldsymbol{I}\pm\boldsymbol{B}$ 非奇异. 由于

$$(\boldsymbol{I}\pm\boldsymbol{B})^{-1}=\boldsymbol{I}\mp\boldsymbol{B}(\boldsymbol{I}\pm\boldsymbol{B})^{-1},$$

于是得

$$\|(\boldsymbol{I}\pm\boldsymbol{B})^{-1}\|\leqslant\|\boldsymbol{I}\|+\|\boldsymbol{B}\|\,\|(\boldsymbol{I}\pm\boldsymbol{B})^{-1}\|.$$

当矩阵范数取算子范数时，$\|\boldsymbol{I}\|=1$. 因此，定理得证.

类似于向量范数，矩阵范数也具有下面的等价性.

定理 1.8 $\mathbf{R}^{n\times n}$ 上的任意两种矩阵范数都是等价的，即对 $\mathbf{R}^{n\times n}$ 上的任意两种矩阵范数 $\|\boldsymbol{A}\|_s$ 和 $\|\boldsymbol{A}\|_t$，存在常数 $c_1,c_2>0$，使得

$$c_1\|\boldsymbol{A}\|_s\leqslant\|\boldsymbol{A}\|_t\leqslant c_2\|\boldsymbol{A}\|_s.$$

由矩阵范数的等价性，我们可以用矩阵的范数描述矩阵序列的极限性质.

定义 1.11 设矩阵序列 $\boldsymbol{A}^{(k)}=(a_{ij}^{(k)})_{n\times n}\in\mathbf{R}^{n\times n}$，$k=1,2,\cdots$，若存在 $\boldsymbol{A}^*=$

$(a_{ij}^{*})_{n\times n}\in\mathbf{R}^{n\times n}$，使得

$$\lim_{k\to\infty}a_{ij}^{(k)}=a_{ij}^{*},\quad i,j=1,2,\cdots,n,$$

则称序列$\{\boldsymbol{A}^{(k)}\}$收敛于$\boldsymbol{A}^{*}$，记为

$$\lim_{k\to\infty}\boldsymbol{A}^{(k)}=\boldsymbol{A}^{*}.$$

可以验证

$$\lim_{k\to\infty}\boldsymbol{A}^{(k)}=\boldsymbol{A}^{*}\Leftrightarrow\lim_{k\to\infty}\|\boldsymbol{A}^{(k)}-\boldsymbol{A}^{*}\|=0.$$

评　　注

本章介绍了数值计算的研究对象、误差及相关概念、数值计算的稳定性及构造算法的基本原则. 考虑到矩阵计算的数值分析，本章还介绍了向量范数和矩阵范数的基本概念和常用定理.

误差分析问题是数值分析中重要而困难的问题. 误差的基本概念和误差分析的若干原则，对学习本课程是很有必要的. 但是，作为工程或科学计算的实际问题则要复杂得多，往往要根据不同问题分门别类地进行分析. 例如，由于舍入误差有随机性，有人应用概率的观点研究误差规律. 在工程计算中，常用几种不同办法(包括实验方法)进行比较，以确定计算结果的可靠性. 20 世纪 60 年代以来，发展了两种估计误差的理论：一种是 J. H. Wilkinson 等人针对计算机浮点算法提出了一套预先估计的研究误差的方法，使矩阵运算的舍入误差研究获得了新发展；另一种是 R. E. Moore 等人应用区间分析理论估计误差，开创了研究误差的新方法.

关于范数方面，所述内容是为以下各章服务的一些初步概念和常用的定理，对本书够用就可以了. 例如只讨论了 $\mathbf{R}^{n\times n}$ 的范数，而没有顾及 $\mathbf{R}^{n\times m}$. 又例如介绍了 $\mathbf{R}^{n}$ 和 $\mathbf{R}^{n\times n}$ 上范数的等价性，此性质对有限维空间都是成立的，而对于 $C[a,b]$ 则没有这个性质，这些都是赋范线性空间有关的问题，详细讨论这些问题是泛函分析的内容.

习　题　1

1.1　已知 e=2.718 28…，问下列近似值 x_A 有几位有效数字，相对误差界是多少？

(1) $x=\mathrm{e}$，$x_A=2.7$；　　(2) $x=\mathrm{e}$，$x_A=2.718$；

(3) $x=\dfrac{\mathrm{e}}{100}$，$x_A=0.027$；　　(4) $x=\dfrac{\mathrm{e}}{100}$，$x_A=0.027\,18$.

1.2 设原始数据的下列近似值每位都是有效数字：

$$x_1^*=1.1021,\ x_2^*=0.031,\ x_3^*=56.430.$$

试计算(1) $x_1^*+x_2^*+x_3^*$；(2)$\dfrac{x_2^*}{x_3^*}$，并估计它们的相对误差界.

1.3 设 x 的相对误差界为 δ，求 x^n 的相对误差界.

1.4 设 $x>0$，x 的相对误差界为 δ，求 $\ln 2$ 的绝对误差界.

1.5 为了使计算球体体积时的相对误差不超过 1%，问测量半径 R 时的允许相对误差界是多少？

1.6 三角函数值取 4 位有效数字，怎样计算 $1-\cos 2^\circ$ 才能保证精度？

1.7 设 $Y_0=28$，按递推公式

$$Y_n=Y_{n-1}-\frac{1}{100}\sqrt{783},\quad n=1,2,\cdots,$$

计算. 若取 $\sqrt{783}\approx 27.982$(5 位有效数字)，试问计算 Y_{100} 将有多大误差？

1.8 求解方程 $x^2+56x+1=0$，使其根至少具有 4 位有效数字(用 $\sqrt{783}\approx 27.982$).

1.9 正方形的边长大约为 100 cm，应怎样测量才能使其面积的误差不超过 1 cm^2？

1.10 序列$\{y_n\}$满足递推关系

$$y_n=10y_{n-1}-1,\quad n=1,2,\cdots.$$

若 $y_0=\sqrt{2}\approx 1.41$(3 位有效数字)，计算到 y_{10} 时的误差有多大？这个计算过程稳定吗？

1.11 对积分 $I_n=\int_0^1 x^n e^{x-1}dx$，$n=0,1,\cdots$，验证 $I_0=1-e^{-1}$，$I_n=1-nI_{n-1}$. 若取 $e^{-1}\approx 0.3679$，按递推公式 $I=1-nI_{n-1}$，用 4 位有效数字计算 $I_0,I_1,\cdots,I_9$，并证明这种算法是不稳定的.

1.12 反双曲正弦函数为 $f(x)=\ln(x+\sqrt{x^2+1})$. 如何计算 $f(x)$ 才能避免有效数字的损失？试计算 $f(30)$ 和 $f(-30)$(开方和对数用 6 位函数表).

1.13 下列公式是否要作变换才能避免有效数字的损失？如何变换？

(1) $\sin x-\sin y$；　　(2) $\arctan x-\arctan y$；

(3) $\sqrt{x+4}-2$；　　(4) $\dfrac{e^{2x}-1}{2}$.

1.14 已知三角形面积 $s=\dfrac{1}{2}ab\sin C$，其中 C 为弧度，$0<C<\dfrac{\pi}{2}$，且测量 a，b，C 的误差分别为 Δa，Δb，ΔC，证明面积的误差 Δs 满足

$$\left|\frac{\Delta s}{s}\right| \leqslant \left|\frac{\Delta a}{a}\right| + \left|\frac{\Delta b}{b}\right| + \left|\frac{\Delta C}{C}\right|.$$

1.15　设 $\boldsymbol{P}\in\mathbf{R}^{n\times n}$ 且非奇异，又设 $\|\boldsymbol{x}\|$ 为 $\mathbf{R}^n$ 上的一种向量范数，定义

$$\|\boldsymbol{x}\|_P = \|\boldsymbol{P}\boldsymbol{x}\|.$$

试证明 $\|\boldsymbol{x}\|_P$ 是 $\mathbf{R}^n$ 上的一种向量范数.

1.16　设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 为对称正定矩阵，定义

$$\|\boldsymbol{x}\|_A = (\boldsymbol{A}\boldsymbol{x},\boldsymbol{x})^{\frac{1}{2}}.$$

试证明 $\|\boldsymbol{x}\|_A$ 为 $\mathbf{R}^n$ 上的一种向量范数.

1.17　设矩阵

$$\boldsymbol{A} = \begin{bmatrix} 0.6 & 0.5 \\ 0.1 & 0.3 \end{bmatrix}.$$

计算 $\boldsymbol{A}$ 的行范数、列范数、2 范数及 F 范数.

1.18　证明 $\|\boldsymbol{A}\|_2 \leqslant \|\boldsymbol{A}\|_{\mathrm{F}} \leqslant \sqrt{n}\,\|\boldsymbol{A}\|_2$，并说明 $\|\boldsymbol{A}\|_{\mathrm{F}}$ 与 $\|\boldsymbol{x}\|_2$ 相容.

1.19　设 $\boldsymbol{P}\in\mathbf{R}^{n\times n}$ 且非奇异，又设 $\|\boldsymbol{x}\|$ 为 $\mathbf{R}^n$ 上的一种向量范数，定义范数 $\|\boldsymbol{x}\|_P = \|\boldsymbol{P}\boldsymbol{x}\|$. 证明对应于 $\|\boldsymbol{x}\|_P$ 的算子范数 $\|\boldsymbol{A}\|_P = \|\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}\|$.

1.20　设 $\boldsymbol{A}$ 为非奇异矩阵，求证：

$$\frac{1}{\|\boldsymbol{A}^{-1}\|_\infty} = \min_{y\neq 0}\frac{\|\boldsymbol{A}\boldsymbol{y}\|_\infty}{\|\boldsymbol{y}\|_\infty}.$$

1.21　设 $\boldsymbol{A}$ 为 n 阶方阵，U 为 n 阶正交矩阵，试证：

$$\|\boldsymbol{A}U\|_2 = \|U\boldsymbol{A}\|_2 = \|\boldsymbol{A}\|_2,$$

$$\|\boldsymbol{A}U\|_{\mathrm{F}} = \|U\boldsymbol{A}\|_{\mathrm{F}} = \|\boldsymbol{A}\|_{\mathrm{F}}.$$

1.22　对算子范数，设 $\|\boldsymbol{B}\| < 1$，求证：

$$\frac{1}{1+\|\boldsymbol{B}\|} \leqslant \|(\boldsymbol{I}\pm\boldsymbol{B})^{-1}\|.$$

数值试验题 1

1.1　设 $f(x) = x(\sqrt{x+1}-\sqrt{x})$，$g(x) = \dfrac{1}{\sqrt{x+1}-\sqrt{x}}$. 用软件工具或自编程序计算 $x=1$，$x=10^5$，$x=10^{10}$ 时 $f(x)$ 和 $g(x)$ 的值，并对计算结果和计算方法进行分析.

1.2　有下列两种方式计算 e^{-5} 的近似值：

(1) $\mathrm{e}^{-5} \approx \sum\limits_{n=0}^{9}(-1)^n\dfrac{5^n}{n!}$；　　(2) $\mathrm{e}^{-5} \approx \left(\sum\limits_{n=0}^{9}\dfrac{5^n}{n!}\right)$.

用软件工具或自编程序计算这两个表达式的值，并对计算结果和计算方法进行分析.

1.3 序列$\{3^{-n}\}$可由下列两种递推公式生成：

(1) $x_0=1$，$x_n=\frac{1}{3}x_{n-1}$，$n=1,2,\cdots$；

(2) $y_0=1$，$y_1=\frac{1}{3}$，$y_n=\frac{5}{3}y_{n-1}-\frac{4}{9}y_{n-2}$，$n=2,3,\cdots$.

用软件工具或自编程序递推地计算$\{x_n\}$和$\{y_n\}$，并对计算结果和计算方法进行分析.

1.4 设 $p(x)=(x-1)(x-2)\cdots(x-20)$，显然，该多项式的全部根为 1，2，…，20 共 20 个. 取多个非常小的数 ε，用软件工具解方程 $p(x)+\varepsilon x^{19}=0$，并对计算结果进行分析.

第2章　插　值　法

在许多实际问题中，我们需要用函数 $y=f(x)$ 来表示某种内在规律的数量关系，其中相当一部分函数是基于实际或观测数据而得到的. 虽然 $f(x)$ 在某个区间 $[a,b]$ 上是存在的，有的还是连续的，但都只能给出 $[a,b]$ 上一系列点 x_i 的函数值 $y_i=f(x_i)$，$i=0,1,\cdots,n$，这只是一个函数表. 有的函数虽有解析表达式，但由于计算复杂，使用不方便，通常也造一个函数表. 为了研究函数的变化规律，往往需要求出不在给定数据点上的其他函数值. 因此，我们希望根据给定的函数值做一个既能反映函数 $f(x)$ 的特性，又便于计算的简单函数 $\varphi(x)$，用 $\varphi(x)$ 近似 $f(x)$，粗略地说，就是要对函数的离散数据建立简单的数学模型.

例如，从人口普查统计，已知某国新生儿累计分布为 $y_i=f(x_i)$，x_i 为母亲年龄，y_i 为新生儿母亲的年龄低于或等于 x_i 的新生儿数目. 我们需要建立 $y=f(x)$ 的简单数学模型. 又如，由化学实验得到某种物质浓度与时间的关系 $y_i=f(x_i)$，x_i 为时间，y_i 为对应的浓度，我们需要建立 $y=f(x)$ 的简单数学模型.

插值法就是用一个便于计算的简单的函数 $\varphi(x)$ 去代替 $f(x)$，使得

$$\varphi(x_i)=y_i,\ i=0,1,2,\cdots,n.$$

通常称 $f(x)$ 为被插值函数，$x_0,x_1,\cdots,x_n$ 为插值节点，$\varphi(x)$ 为插值函数. 将求 $\varphi(x)$ 的方法称为插值法.

2.1　Lagrange 插值多项式

2.1.1　多项式插值问题

用代数多项式作为插值函数的插值法称为多项式插值，相应的多项式称为插值多项式. 设函数 $f(x)$ 在 $n+1$ 个相异点 $x_0,x_1,\cdots,x_n$ 上的值为

$$y_i=f(x_i),\ i=0,1,\cdots,n, \tag{2.1}$$

则存在唯一的次数不超过 n 的多项式 $\varphi(x)$ 满足条件(2.1)式. 事实上，令

$$\varphi(x)=a_0+a_1x+\cdots+a_nx^n,$$

由插值条件(2.1)式有

$$a_0+a_1x_i+\cdots+a_nx_i^n=y_i,\ i=0,1,\cdots,n. \tag{2.2}$$

这是关于未知数 $a_0,a_1,\cdots,a_n$ 的线性方程组，它的系数矩阵的行列式是

Vandermonde 行列式

$$\boldsymbol{\Delta}=\begin{vmatrix}1 & x_0 & x_0^2 & \cdots & x_0^n\\ 1 & x_1 & x_1^2 & \cdots & x_1^n\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_n & x_n^2 & \cdots & x_n^n\end{vmatrix}=\prod_{i>j}(x_i-x_j).$$

因为 $x_i\neq x_j(i\neq j)$，所以 $\boldsymbol{\Delta}\neq 0$. 这表明(2.2)式存在唯一解 $a_0,a_1,\cdots,a_n$.

上述的存在唯一性说明，满足插值条件的多项式存在，并且插值多项式与构造方法无关. 然而，直接求解方程组(2.2)的方法，不但计算复杂，而且难于得到 $\varphi(x)$的简单表达式. 下面，我们将给出不同形式的便于使用的插值多项式.

2.1.2 Lagrange 插值多项式

先考察低次插值多项式. 当 $n=1$ 时，要构造通过两点(x_0,y_0)和(x_1,y_1)的不超过 1 次的多项式 $L_1(x)$，使得 $L_1(x_0)=y_0$，$L_1(x_1)=y_1$. 显然，$L_1(x)$可以写成

$$L_1(x)=y_0\,\frac{x-x_1}{x_0-x_1}+y_1\,\frac{x-x_0}{x_1-x_0}.$$

它是两个线性函数

$$l_0(x)=\frac{x-x_1}{x_0-x_1},\ l_1(x)=\frac{x-x_0}{x_1-x_0} \tag{2.3}$$

的线性组合. 显然，$l_0(x)$和 $l_1(x)$也是线性插值多项式，满足

$$l_0(x_0)=1,\ l_0(x_1)=0,\ l_1(x_0)=0,\ l_1(x_1)=1.$$

我们称(2.3)式为线性插值基函数. 这种用插值基函数表示的方法容易推广到一般情形.

设 $x_0<x_1<\cdots<x_n$ 为插值节点，若 n 次多项式 $l_k(x)(k=0,1,\cdots,n)$满足条件

$$l_k(x_i)=\delta_{ik}=\begin{cases}1, & i=k,\\ 0, & i\neq k,\end{cases} \tag{2.4}$$

则由此可得

$$l_k(x)=\prod_{\substack{i=0\\ i\neq k}}^{n}\frac{x-x_i}{x_k-x_i},\quad k=0,1,\cdots,n, \tag{2.5}$$

称其为 Lagrange 插值基函数.

引入记号

$$\omega_{n+1}(x)=\prod_{i=0}^{n}(x-x_i). \tag{2.6}$$

容易求得

$$\omega_{n+1}'(x_k) = \prod_{\substack{i=0 \\ i\neq k}}^{n}(x_k - x_i).$$

于是,(2.5)式可以写成

$$l_k(x)=\frac{\omega_{n+1}(x)}{(x-x_k)\omega_{n+1}'(x_k)}. \tag{2.7}$$

值得注意的是,插值基函数仅由节点确定,与被插函数无关.

由(2.4)式易见,满足插值条件

$$L_n(x_i)=y_i, \quad i=0,1,\cdots,n, \tag{2.8}$$

的插值多项式 $L_n(x)$为

$$L_n(x) = \sum_{k=0}^{n} y_k l_k(x), \tag{2.9}$$

称之为 Lagrange 插值多项式.

显然,如此构造的 $L_n(x)$是不超过 n 次的多项式. 当 $n=1$ 时,称为线性插值多项式. 当 $n=2$ 时,称为抛物线插值多项式.

例 **2.1**　已知$\sqrt{4}=2$,$\sqrt{9}=3$,用线性插值求$\sqrt{7}$的近似值.

解　考虑函数 $f(x)=\sqrt{x}$. 若 $x_0=4$,$x_1=9$,则 $y_0=2$,$y_1=3$. 基函数分别为

$$l_0(x)=\frac{x-9}{4-9}=-\frac{1}{5}(x-9), \quad l_1(x)=\frac{x-4}{9-4}=\frac{1}{5}(x-4).$$

插值多项式为

$$L_1(x)=y_0 l_0(x)+y_1 l_1(x)=\frac{1}{5}(x+6).$$

所以

$$\sqrt{7}\approx L_1(7)=2.6.$$

2.1.3　插值余项

插值公式(2.9)是在节点 $x_0,x_1,\cdots,x_n$ 上关于 $f(x)$的插值多项式. 我们希望知道,当 $x\neq x_i(i=0,1,\cdots,n)$时,$f(x)$与 $L_n(x)$的偏差(不包括计算 $L_n(x)$时出现的舍入误差)有多大. 称

$$R_n(x)=f(x)-L_n(x)$$

为插值多项式的余项,也就是插值的截断误差. 下面给出关于插值余项的基本结果.

定理 **2.1**　设 $x_0,x_1,\cdots,x_n$ 为区间$[a,b]$上相异节点,$f(x)\in C^n[a,b]$,并且 $f^{(n+1)}(x)$在(a,b)内存在,$L_n(x)$为满足插值条件(2.8)式的插值多项式,则对任

何 $x\in[a,b]$，存在 $\xi\in(a,b)$，使得

$$R_n(x)=f(x)-L_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x). \tag{2.10}$$

证　当 x 为插值节点时，(2.10)式显然成立. 下面假设 $x\in[a,b]$，但 x 不是节点. 引入辅助函数

$$G(t)=R_n(t)-\frac{\omega_{n+1}(t)}{\omega_{n+1}(x)}R_n(x),$$

其中 x 是固定的，t 是自变量.

显然，$G(t)\in C^n[a,b]$，$G^{(n+1)}(t)$ 在 (a,b) 内存在. 当 $t=x_0,x_1,\cdots,x_n$ 时，$G(t)=0$. 根据 Rolle 定理，$G'(t)$ 在 (a,b) 内至少存在 $n+1$ 个相异的零点. 一般地，$G^{(i)}(t)$ 在 (a,b) 内至少存在 $n+2-i$ 个相异的零点，$i=1,2,\cdots,n+1$. 设 $\xi=\xi(x)\in(a,b)$ 是 $G^{(n+1)}(t)$ 的零点，即 $G^{(n+1)}(\xi)=0$. 由于

$$G^{(n+1)}(t)=f^{(n+1)}(t)-\frac{(n+1)!}{\omega_{n+1}(x)}R_n(x),$$

所以，(2.10)式得证.

应当指出，余项表达式仅在 $f(x)$ 的高阶导数存在时才能应用. ξ 在 (a,b) 内的具体位置通常不可能给出. 如果我们可以求出 $\max\limits_{a\leqslant x\leqslant b}|f^{(n+1)}(x)|=M_{n+1}$，那么插值多项式 $L_n(x)$ 逼近 $f(x)$ 的截断误差界是

$$|R_n(x)|\leqslant\frac{M_{n+1}}{(n+1)!}|\omega_{n+1}(x)|. \tag{2.11}$$

例 **2.2**　设 $f(x)=\ln x$，并已知 $f(x)$ 的数据如表 2－1.

表 **2－1**

x	0.40	0.50	0.70	0.80
$\ln x$	－0.916 291	－0.693 147	－0.356 675	－0.223 144

试用 3 次 Lagrange 插值多项式 $L_3(x)$ 来计算 $\ln(0.6)$ 的近似值并估计误差.

解　用 $x_0=0.40$，$x_1=0.50$，$x_2=0.70$ 和 $x_3=0.80$ 作 3 次 Lagrange 插值多项式 $L_3(x)$，把 $x=0.6$ 代入 $L_3(x)$ 中，得

$$L_3(0.6)=-0.509\,975.$$

由于

$$\max_{0.4\leqslant x\leqslant 0.8}|f^{(4)}(x)|\leqslant 234.4,$$

利用余项估计(2.11)式可以得到

$$|R_3(x)|\leqslant 0.004.$$

$\ln(0.6)$ 的真值为 $-0.510\,826$，由此得到 $R_3(0.6)=-0.000\,85$. 这个例子

说明，估计式(2.11)给出了一个较好的估计.

利用(2.10)式估计截断误差实际上非常困难. 一是因为它要计算函数$f(x)$的高阶导数，当 $f(x)$很复杂时，计算量很大，而当 $f(x)$没有可用来计算的表达式时，导数无法准确计算；二是因为即使能得到高阶导数的解析式，但由于 ξ 的具体位置不知道，所以要估计高阶导数在插值区间上的界一般是非常困难的. 因此，(2.10)式并不实用. 不过，(2.10)式从理论上也说明了运用插值法时必须注意下列问题.

(1) 如果 $f(x)$本身是次数不超过 n 的多项式，那么满足 $n+1$ 个插值条件的插值多项式就是它自身. 这是因为 $f^{(n+1)}(x)\equiv 0$，从而 $R_n(x)\equiv 0$. 特别地，对 $f(x)\equiv 1$，由(2.10)式得到函数的一个性质

$$\sum_{k=0}^{n} L_k(x) = 1.$$

(2) 如果插值区间$[a,b]$很大，那么对给定的 x，$|\omega_{n+1}(x)|$的值一般会很大(因为这时许多因数都将大于 1). 因此，误差 $R_n(x)$可能很大. 反过来，如果插值区间$[a,b]$很小，比如 $b-a<1$，那么对给定的 x，$|\omega_{n+1}(x)|$的值一定很小(因为所有因数都将小于 1). 因而误差 $R_n(x)$就会很小.

(3) 由(2.10)式可见，当 $n\to\infty$时，$R_n(x)$未必趋于. 因此，依靠增多插值节点不一定能减少误差.

(4) 插值多项式一般仅用来估计插值区间内点的函数值(即内插). 用它计算插值区间外点的函数值(即外插)时，误差可能会较大.

2.2　逐次线性插值法

2.2.1　逐次线性插值思想

既然(2.10)式估计误差时不实用，那么实际中如何估计截断误差呢？

假设插值条件中包含 $n+2$ 组数据

$$f(x_i)=y_i,\quad i=0,1,\cdots,n,n+1,$$

那么利用前 $n+1$ 组数据我们可以构造一个 n 次 Lagrange 插值多项式 $L_n(x)$，利用后 $n+1$ 组数据我们可以构造另一个 n 次 Lagrange 插值多项式 $L_n^*(x)$. 利用(2.10)式知，它们各自的插值余项为

$$f(x)-L_n(x)=\frac{1}{(n+1)!}f^{(n+1)}(\xi)(x-x_0)(x-x_1)\cdots(x-x_n),$$

$$f(x)-L_n^*(x)=\frac{1}{(n+1)!}f^{(n+1)}(\xi^*)(x-x_1)(x-x_2)\cdots(x-x_{n+1}).$$

两式相减(假设 $f^{(n+1)}(\xi)\approx f^{(n+1)}(\xi^*)$)得

$$L_n^*(x)-L_n(x)\approx\frac{1}{(n+1)!}f^{(n+1)}(\xi)(x-x_1)\cdots(x-x_n)(x_{n+1}-x_0),$$

并可写成

$$\frac{1}{(n+1)!}f^{(n+1)}(\xi)(x-x_1)\cdots(x-x_n)\approx\frac{L_n^*(x)-L_n(x)}{x_{n+1}-x_0}.$$

由此可得

$$R_n(x)=f(x)-L_n(x)\approx\frac{L_n(x)-L_n^*(x)}{x_0-x_{n+1}}(x-x_0),\tag{2.12}$$

$$R_n^*(x)=f(x)-L_n^*(x)\approx\frac{L_n^*(x)-L_n(x)}{x_{n+1}-x_0}(x-x_{n+1}).\tag{2.13}$$

(2.12)式和(2.13)式分别给出了用 $L_n(x)$ 和 $L_n^*(x)$ 作近似计算时的实用误差估计式. 它不需要计算高阶导数,也不用估计插值区间上高阶导数的界.

例 **2.3** 已知

$$f(0)=2,\ f(1)=3,\ f(2)=12.$$

利用 Lagrange 插值法计算未知函数 $y=f(x)$ 在 $x=1.207\ 8$ 处的函数值 $f(1.207\ 8)$,并估计误差.

解 利用前两组数据可以构造一个 1 次 Lagrange 插值多项式

$$L_1(x)=\frac{x-1}{0-1}\cdot 2+\frac{x-0}{1-0}\cdot 3,$$

利用后两组数据可以构造另一个 1 次 Lagrange 插值多项式

$$L_1^*(x)=\frac{x-2}{1-2}\cdot 3+\frac{x-1}{2-1}\cdot 12.$$

因为 $1.207\ 8\in[1,2]$,所以

$$f(1.207\ 8)\approx L_1^*(1.207\ 8)=4.870\ 2,$$

其误差

$$R_1(1.207\ 8)\approx\frac{L_1^*(1.207\ 8)-L_1(1.207\ 8)}{2-0}(1.207\ 8-2)=-0.658\ 476\ 64.$$

基于上述分析,一种自然的想法是如果我们把 $L_n^*(x)$ 加上其截断误差 $R_n^*(x)$,那么所得的 $n+1$ 次多项式

$$P_{n+1}(x)=L_n^*(x)+\frac{L_n^*(x)-L_n(x)}{x_{n+1}-x_0}(x-x_{n+1})$$

应该是 $f(x)$ 更好的近似函数.

实际上,上述 $P_{n+1}(x)$ 满足插值条件

$$P_{n+1}(x_i)=y_i,\quad i=0,i,\cdots,n,n+1.$$

就是说,$P_{n+1}(x)$ 恰好是由已知 $n+2$ 个插值节点确定的 Lagrange 插值多项式

$L_{n+1}(x)$. 这意味着从任何 $n+1$ 个插值节点构造 n 次 Lagrange 插值多项式 $L_n(x)$，可以先选用合适的两个节点构造线性插值多项式，再利用线性插值多项式构造 2 次插值多项式，利用 2 次插值多项式又可以构造 3 次插值多项式……直到构造出 n 次插值多项式. 下面我们介绍具体的算法.

2.2.2　Aitken 算法

对求知函数或复杂函数 $f(x)$，假设已知如下信息

$$f(x_i)=y_i,\quad i=0,1,\cdots,n.$$

问题是利用以上信息计算 $f(x)$在任何一点 x 处的函数值 $f(x)$，且误差不超过上限 ε_0.

显然，为了尽快地计算出指定精度的函数值，我们应该尽量多地利用靠近 x 的节点信息. 为此，首先对所有节点排序，与 x 接近的点排在前面，比如说，这个顺序恰好是 $x_0,x_1,x_2,\cdots,x_n$. 接下来，我们计算 $f(x)$的近似值.

第一步：利用节点 x_0,x_1 构造线性插值多项式 $I_{0,1}(x)$，利用节点 x_0,x_2 构造另一个线性插值多项式 $I_{0,2}(x)$.

利用实用误差估计式估计 $I_{0,1}(x)$的误差

$$R_{0,1,2}(x)=\frac{I_{0,1}(x)-I_{0,2}(x)}{x_1-x_2}(x-x_1).$$

若 $|R_{0,1}(x)|<\varepsilon_0$，算法终止，记

$$I_{0,1,2}(x)=I_{0,1}(x)+R_{0,1,2}(x),$$

则得 $f(x)$的近似值 $I_{0,1,2}(x)\approx f(x)$.

一般地，设 $I_{0,1,\cdots,k}(x)$ 是关于节点 $x_0,x_1,\cdots,x_k$ 的 k 次插值多项式，$I_{0,1,\cdots,k,l}(x)$是关于节点 $x_0,\cdots,x_{k-1},x_l$ 的 k 次插值多项式，令

$$I_{0,1,\cdots,k,l}(x)=I_{0,1,\cdots,k}(x)+\frac{I_{0,1,\cdots,k}(x)-I_{0,1,\cdots,k-1,l}(x)}{x_k-x_l}(x-x_k),\tag{2.14}$$

则 $I_{0,1,\cdots,k,l}(x)$是关于节点 $x_0,x_1,\cdots,x_k,x_l$ 的 $k+1$ 次多项式. 我们称(2.14)式为 Aitken 逐次线性插值公式. 记

$$R_{0,1,\cdots,k,k+1}(x)=\frac{I_{0,1,\cdots,k}(x)-I_{0,1,\cdots,k-1,k+1}(x)}{x_k-x_{k+1}}(x-x_k),$$

如果计算过程是数值稳定的，则当 $|R_{0,1,\cdots,k,k+1}(x)|<\varepsilon_0$ 时，算法终止.

当 $k=0$ 时，为线性插值. 当 $k=1$ 时，插值节点为 x_0,x_1,x_l 的插值多项式为

$$I_{0,1,l}(x)=I_{0,1}(x)+\frac{I_{0,1}(x)-I_{0,l}(x)}{x_1-x_l}(x-x_1).$$

计算时，可由 $k=0$ 到 $k=n-1$ 逐次求得所需要的插值多项式. 计算过程可用三

角形表 2－2(Aitken 算法)表示.

表 2－2

i	x_i	$I_i(x)=y_i$	$I_{0,i}(x)$	$I_{0,1,i}(x)$	$I_{0,1,2,i}(x)$
0	x_0	$I_0(x)=y_0$			
1	x_1	$I_1(x)=y_1$			
2	x_2	$I_2(x)=y_2$			
3	x_3	$I_3(x)=y_3$			
…	…	…	…	…	…

从表 2－2 可看到,每增加一个节点就增加计算一行,斜线上是多项式的值.如果精度不满足要求,再增加一个节点,前面计算的结果完全有效.这个算法适用于计算机上计算,且具有自动选节点并逐步比较精度的特点,程序也较简单.

例 **2.4** 已知特殊角 0°,30°,45°,60°,90°的正弦函数值.问用多少个节点计算 sin50°可使近似值准确到 2,3,4 位小数(已知 sin50°=0.766 044)?

解 按离 50°角的距离由近到远顺序排列插值节点,再按 Aitken 算法完成三角形表 2－3 的计算.

表 2－3

i	x_i	$\sin(x_i)$	$I_{0,i}(x)$	$I_{0,1,i}(x)$	$I_{0,1,2,i}(x)$	$I_{0,1,2,3,i}(x)$
0	45°	$\frac{\sqrt{2}}{2}$				
1	60°	$\frac{\sqrt{3}}{2}$	$I_{0,1}=0.760\ 08$			
2	30°	0.5	$I_{0,2}=0.776\ 14$	$I_{0,1,2}=0.765\ 43$		
3	90°	1	$I_{0,3}=0.739\ 65$	$I_{0,1,3}=0.766\ 89$	$I_{0,1,2,3}=0.765\ 92$	
4	0°	0	$I_{0,4}=0.785\ 67$	$I_{0,1,4}=0.764\ 35$	$I_{0,1,2,4}=0.766\ 16$	$I_{0,1,2,3,4}=0.766\ 03$

同 sin50°=0.766 044 相比,近似值 $I_{0,1,2}=0.765\ 43$,$I_{0,1,2,3}=0.765\ 492$ 和 $I_{0,1,2,3,4}=0.766\ 03$分别准确到了 2,3,4 位小数,它们分别用了 3,4,5 个节点.

2.3 Newton 插值多项式

2.3.1 均差及其性质

Lagrange 插值公式结构紧凑,便于理论分析.利用插值基函数也容易得到插值多项式的值.Lagrange 插值公式的缺点是,当插值节点增加、减少或其位置

变化时，全部插值基函数均要随之变化，从而整个插值公式的结构也将发生变化，这在实际计算中是非常不利的. 逐次线性插值法能够有效地计算任何给定点的函数值，而不需要写出各步用到的插值多项式的表达式. 当解决某个问题时，需要插值多项式的表达式时，这个优点就成了缺点. 下面引入 的 Newton 插值公式可以克服上述不足.

当 $n=1$ 时，由点斜式直线方程知，过两点$(x_0, f(x_0))$和$(x_1, f(x_1))$的直线方程为

$$N_1(x)=f(x_0)+\frac{f(x_1)-f(x_0)}{x_1-x_0}(x-x_0).$$

若记

$$f[x_0, x_1]=\frac{f(x_1)-f(x_0)}{x_1-x_0},$$

则可把 $N_1(x)$写成

$$N_1(x)=f(x_0)+f[x_0, x_1](x-x_0).$$

显然，$N_1(x)$就是 1 次插值多项式 $L_1(x)$.

当 $n=2$ 时，进而记

$$f[x_1, x_2]=\frac{f(x_2)-f(x_1)}{x_2-x_1},$$

$$f[x_0, x_1, x_2]=\frac{f[x_1, x_2]-f[x_0, x_1]}{x_2-x_0},$$

类似地，构造次数不超过 2 次的多项式

$$N_2(x)=f(x_0)+f[x_0, x_1](x-x_0)+f[x_0, x_1, x_2](x-x_0)(x-x_1).$$

容易检验，这样的 $N_2(x)$满足插值条件

$$N_2(x_0)=f(x_0),\ N_2(x_1)=f(x_1),\ N_2(x_2)=f(x_2).$$

因此，$N_2(x)$就是 2 次插值多项式 $L_2(x)$.

为了构造更一般的插值多项式，先用递推方法定义均差.

定义 2.1　给定 $f(x)$在节点上的函数值，

$$f[x_0, x_k]=\frac{f(x_k)-f(x_0)}{x_k-x_0}$$

称为函数 $f(x)$关于节点 x_0, x_k 的一阶均差. 一般地

$$f[x_0, x_1, \cdots, x_k]=\frac{f[x_1, x_2, \cdots, x_k]-f[x_0, x_1, \cdots, x_{k-1}]}{x_k-x_0} \tag{2.15}$$

称为 $f(x)$在 $x_0, x_1, \cdots, x_k$ 上的 k 阶均差(或差商).

均差是数值分析的基本工具，它具有下列基本性质.

(1) k 阶均差(2.15)是函数值 $f(x_0), f(x_1), \cdots, f(x_k)$的线性组合，即有

$$f[x_0,x_1,\cdots,x_k]=\sum_{i=0}^{k}\frac{f(x_i)}{\omega'_{k+1}(x_i)}. \tag{2.16}$$

这个性质可用归纳法证明.这个性质也表明均差与节点的排列次序无关,称为均差的对称性.

(2) 由性质(1)和(2.16)式可得

$$f[x_0,x_1,\cdots,x_k]=\frac{f[x_0,\cdots,x_{k-2},x_k]-f[x_0,x_1,\cdots,x_{k-1}]}{x_k-x_{k-1}}. \tag{2.17}$$

(3) 设 $f(x)\in C^n[a,b]$,且 $x_i\in[a,b](i=0,1,\cdots,n)$为相异节点,那么 $f(x)$ 的 n 阶均差与其 n 阶导数有如下关系

$$f[x_0,x_1,\cdots,x_n]=\frac{1}{n!}f^{(n)}(\xi),\quad \xi\in(a,b). \tag{2.18}$$

这个公式可直接用 Rolle 定理证明.

在实际计算中,经常利用由表 2-4 给出的均差表.

表 **2-4**

x_k	$f(x_k)$	一阶均差	二阶均差	三阶均差
x_0	$\underline{f(x_0)}$			
x_1	$f(x_1)$	$\underline{f[x_0,x_1]}$		
x_2	$f(x_2)$	$f[x_1,x_2]$	$\underline{f[x_0,x_1,x_2]}$	
x_3	$f(x_3)$	$f[x_2,x_3]$	$f[x_1,x_2,x_3]$	$\underline{f[x_0,x_1,x_2,x_3]}$
…	…	…	…	…

2.3.2 Newton 插值公式

下面利用均差表 2-4 中加下划横线的均差值直接构造插值多项式.根据均差定义,把 x 看成$[a,b]$上一点,可得

$$f(x)=f(x_0)+f[x,x_0](x-x_0),$$
$$f[x,x_0]=f[x_0,x_1]+f[x,x_0,x_1](x-x_1),$$
$$\cdots\cdots$$
$$f[x,x_0,\cdots,x_{n-1}]=f[x_0,x_1,\cdots,x_n]+f[x,x_0,\cdots,x_n](x-x_n).$$

对上述各式,把后一式代入前一式得

$$\begin{aligned}f(x)=&f(x_0)+f[x_0,x_1](x-x_0)+f[x_0,x_1,x_2](x-x_0)(x-x_1)+\cdots\\&+f[x_0,x_1,\cdots,x_n](x-x_0)(x-x_1)\cdots(x-x_{n-1})\\&+f[x,x_0,\cdots,x_n]\omega_{n+1}(x).\end{aligned}$$

由此令

$$N_n(x)=f(x_0)+f[x_0,x_1](x-x_0)+f[x_0,x_1,x_2](x-x_0)(x-x_1)+\cdots$$

$$+f[x_0,x_1,\cdots,x_n](x-x_0)(x-x_1)\cdots(x-x_{n-1}), \tag{2.19}$$

则有

$$R_n(x)=f(x)-N_n(x)=f[x,x_0,\cdots,x_n]\omega_{n+1}(x). \tag{2.20}$$

称 $N_n(x)$为 n 次 Newton 插值多项式，(2.20)式为均差型余项.

显然，$N_n(x)$是次数不超过 n 次的多项式，并且由(2.20)可知，它满足插值条件

$$N_n(x_i)=f(x_i),\quad i=0,1,\cdots,n.$$

根据插值多项式的唯一性，$N_n(x)$就是 $L_n(x)$，并且(2,20)式与(2.10)式等价. 事实上，利用均差与导数的关系式(2.18)，可由(2.20)式推出(2.10)式. 不过，(2.20)式更有一般性，它对 $f(x)$是由离散点给出的情形或 $f(x)$的导数不存在的情形均适用.

例 2.5　设 $f(x)=\sqrt{x}$，并已知 $f(x)$的数据如表 2－5. 试用 2 次 Newton 插值多项式 $N_2(x)$计算 $f(2.15)$的近似值，并讨论其误差.

表 2－5

x	2.0	2.1	2.2
$\sqrt{x}$	1.414 214	1.449 138	1.483 240

解　先按表 2－4 构造均差表，具体数据如表 2－6.

表 2－6

x_k	$f(x_k)$	一阶均差	二阶均差
2.0	1.414 214		
2.1	1.449 138	0.349 24	
2.2	1.483 240	0.341 02	−0.041 10

利用 Newton 插值公式(2.19)有

$$N_2(x)=1.414\,214+0.349\,24(x-2.0)-0.041\,10(x-2.0)(x-2.1).$$

取 $x=2.15$ 得 $N_2(2.15)=1.466\,292$.

注意到

$$f^{(3)}(x)=\frac{3}{8x^2\sqrt{x}},\quad \max_{2.0\leqslant x\leqslant 2.2}|f^{(3)}(x)|=0.066\,29,$$

由(2.20)式可以得出

$$\max_{2\leqslant x\leqslant 2.2}|f(x)-N_2(x)|\leqslant 0.552\,417\times 10^{-5}.$$

事实上，$f(2.15)$的真值为 1.466 2 88，可得出 $R(2.15)=-0.4\times 10^{-5}$. 由此看出，所得结果是满意的.

利用 Newton 插值公式，还可以方便地导出某些带导数的插值公式，举例说明如下.

例 **2.6** 已知函数 $f(x)$的值如下

$$f(-1)=-2,\ f(0)=-1,\ f(1)=0,\ f'(0)=0.$$

求不超过 3 次的多项式 $P_3(x)$，使得满足插值条件：

$$P_3(-1)=f(-1),\ P_3(0)=f(0),\ P_3(1)=f(1),\ P_3'(0)=f'(0).$$

解 记 $x_0=-1, x_1=0, x_2=1$，构造不超过 3 次的多项式

$$P_3(x)=f(x_0)+f[x_0,x_1](x-x_0)+f[x_0,x_1,x_2](x-x_0)(x-x_1)$$
$$+\alpha(x-x_0)(x-x_1)(x-x_2),$$

其中，前 3 项是通过 3 个插值点的 2 次 Newton 插值 $N_2(x)$，从而 $P_3(x)$满足 3 个函数值的插值条件. α 是待定常数，由 $x_1=0$ 处的导数值条件确定.

易知，其中的均差

$$f[x_0,x_1]=f[x_1,x_2]=1,\ f[x_0,x_1,x_2]=0,$$

从而

$$P_3(x)=-2+(x+2)+\alpha x(x^2-1).$$

由 $P_3'(0)=0$，得 $\alpha=1$. 所以问题的解是 $P_3(x)=x^3-1$.

2.3.3 差分和等距节点插值公式

在实际计算中，经常遇到插值节点等距分布的情形. 引入差分作为工具，可使 Newton 插值公式得到简化.

设函数 $f(x)$在等距节点 $x_k=x_0+kh(k=0,1,\cdots,n)$上的值 $f_k=f(x_k)$为已知，这里 h 为常数，称为步长. 引入记号

$$\triangle f_k=f_{k+1}-f_k,$$
$$\nabla f_k=f_k-f_{k-1},$$
$$\delta f_k=f\left(x_k+\frac{h}{2}\right)-f\left(x_k-\frac{h}{2}\right)=f_{k+\frac{1}{2}}-f_{k-\frac{1}{2}}.$$

分别称之为 $f(x)$在 x_k 处以 h 为步长的向前差分、向后差分和中心差分. 符号 $\triangle$、∇ 和 δ 分别称为向前差分算子、向后差分算子和中心差分算子.

上面定义的差分为一阶差分. 一般地，m 阶差分可以递推地定义：

$$\triangle^m f_k=\triangle^{m-1}f_{k+1}-\triangle^{m-1}f_k,$$
$$\nabla^m f_k=\nabla^{m-1}f_k-\nabla^{m-1}f_{k-1},$$
$$\delta^m f_k=\delta^{m-1}f_{k+\frac{1}{2}}-\delta^{m-1}f_{k-\frac{1}{2}}.$$

并规定$\triangle^0 f_k=\nabla^0 f_k=\delta^0 f_k=f_k$，称其为零阶差分.

为了讨论差分的性质，再引入两个常用的算子符号.

$$Ef_k=f_{k+1},\ E^{-1}f_k=f_{k-1},\ If_k=f_k.$$

称 E 为步长 h 的移位算子，I 为单位算子(也称不变算子).

由差分定义并应用算子符号运算，可得下列差分的基本性质.

(1) 函数值与差分可以互相表示. 例如

$$f_{k+n}=E^nf_k=(I+\triangle)^nf_k=\sum_{i=0}^{n}\binom{n}{i}\Delta^if_k,$$

$$\triangle^nf_k=(E-I)^nf_k=\sum_{i=0}^{n}(-1)^i\binom{n}{i}f_{n+k-i},$$

$$\nabla^nf_k=(I-E^{-1})^nf_k=\sum_{i=0}^{n}(-1)^{n-i}\binom{n}{i}f_{k+i-n},$$

其中 $\binom{n}{i}=\dfrac{n!}{i!\ (n-i)!}$ 为二项式展开系数.

(2) 对于 $k\geqslant 0$ 有

$$f[x_0,x_1,\cdots,x_k]=\frac{1}{k!\ h^k}\Delta^kf_0. \tag{2.21}$$

该式可用归纳法证明.

(3) 设 $f\in C^k[x_0,x_0+kh]$，则有

$$\triangle^kf_0=h^kf^{(k)}(\xi),\quad \xi\in(x_0,x_k). \tag{2.22}$$

该式可由(2.18)式和(2.21)式得到.

下面利用差分构造等距节点插值公式. 在 Newton 插值公式(2.19)中，用差分代替均差就可以得到等距节点插值公式. 这里只推导常用的前插公式和后插公式.

设 $f_k=f(x_0+kh)(k=0,1,\cdots,N)$ 为已知，要计算 x_0 附近点 $x=x_0+th$ $(0\leqslant t\leqslant 1)$ 处 $f(x)$ 的近似值. 插值节点应取 $x_0,x_1,\cdots,x_n(n\leqslant N)$. 于是

$$\omega_{n+1}(x)=\prod_{i=0}^{n}(x-x_i)=t(t-1)\cdots(t-k)h^{n+1}.$$

将此式及(2.21)式代入 Newton 插值公式(2.19)，可以得到

$$N_n(x_0+th)=f_0+t\triangle f_0+\frac{1}{2!}t(t-1)\triangle^2f_0+\cdots$$

$$+\frac{1}{n!}t(t-1)\cdots(t-n+1)\triangle^nf_0. \tag{2.23}$$

此公式称为 Newton 向前插值公式. 利用二项式系数的记号，可以把(2.23)式写成

$$N_n(x_0+th)=\sum_{k=0}^{n}\binom{t}{k}\triangle^kf_0.$$

其余项可由(2.20)式直接得到

$$R_n(x)=\frac{t(t-1)\cdots(t-n)}{(n+1)!}h^{n+1}f^{(n+1)}(\xi),\quad \xi\in(x_0,x_n). \tag{2.24}$$

类似地,可以导出 Newton 向后插值公式. 设 $x=x_N+th(-1\leqslant t\leqslant 0)$,在Newton插值公式中用 x_N 代替 x_0,用 x_{N-1}代替 x_1,…,用 x_{N-k}代替 x_k,这样就可以得到

$$N_n(x_N+th)=f_N+t\nabla f_N+\frac{1}{2!}t(t+1)\nabla^2 f_N+\cdots$$
$$+\frac{1}{n!}t(t+1)\cdots(t+n-1)\nabla^n f_N. \tag{2.25}$$

此公式称为 Newton 向后插值公式. 把二项式系数扩大到包含负数的情形,记

$$\binom{-t}{k}=\frac{-t(-t-1)\cdots(-t-k+1)}{k!},$$

则有

$$\binom{-t}{k}=(-1)^k\frac{t(t+1)\cdots(t+k-1)}{k!}.$$

由此,(2.25)式可以表示为

$$N_N(x_N+th)=\sum_{k=0}^{n}(-1)^k\binom{-t}{k}\nabla^k f_N.$$

其余项为

$$R_n(x)=\frac{t(t+1)\cdots(t+n)}{(n+1)!}h^{n+1}f^{(n+1)}(\xi),\quad \xi\in(x_{N-n},x_N). \tag{2.26}$$

例 2.7 设 $x_0=1.0$,$h=0.05$,给出 $f(x)=\sqrt{x}$在 $x_k=x_0+kh(k=0,1,\cdots,6)$处的值如表 2-7 的第 3 列,试用 3 次等距节点插值公式求 $f(1.01)$和 $f(1.28)$的近似值.

表 2-7

k	x_k	f_k	$\triangle$	$\triangle^2$	$\triangle^3$
0	1.00	1.000 00	0.024 70		
1	1.05	1.024 70	0.024 11	−0.000 59	
2	1.10	1.048 81	0.023 57	−0.000 54	−0.000 05
3	1.15	1.072 38	……	……	……
4	1.20	1.095 44	0.023 07	−0.000 48	−0.000 03
5	1.25	1.118 03	0.022 59	−0.000 45	
6	1.30	1.140 17	0.022 14		
			∇	∇^2	∇^3

解 用 Newton 向前插值公式(2.23)来计算 $f(1.01)$的近似值. 先构造与均差表相似的差分表,见表 2-7 的上半部分. 由 $t=\dfrac{x-x_0}{h}=0.2$ 得

$$f(1.01)\approx N_3(1.01)=1.004\ 99.$$

用 Newton 向后插值公式(2.25)计算 $f(1.28)$的近似值,可利用表 2-7 中的下半部分. 由 $t=\dfrac{x-x_6}{h}=-0.4$,得

$$f(1.28)\approx N_3(1.28)=1.131\ 37.$$

事实上,$f(1.01)$和 $f(1.28)$的真值分别为 1.004 987 56 和 1.131 370 85. 由此看出,计算结果是相当精确的.

例 2.8 已知 $f(x)=\sin x$ 的数值如表 2-8 的前两列,分别用 Newton 向前、向后插值公式求 sin0.578 91 的近似值,并估计误差.

表 2-8

x	$\sin x$	$\triangle$	$\triangle^2$	$\triangle^3$
0.4	0.389 42			
0.5	0.479 43	0.090 01		
0.6	0.564 64	0.085 21	0.004 80	
0.7	0.644 22	0.079 58	−0.005 63	−0.000 83

解 作差分表如表 2-8,使用 Newton 向前插值公式,取 $x_0=0.5$,$x_1=0.6$,$x_2=0.7$,$x=0.578\ 91$,$h=0.1$,则 $t=\dfrac{x-x_0}{h}=0.789\ 1$,

$$\begin{aligned}N_2(0.578\ 91)&=f_0+t\triangle f_0+\frac{1}{2}t(t-1)\triangle^2 f_0\\&=0.479\ 43+0.085\ 21t+\frac{1}{2}t(t-1)\times(-0.005\ 63)\\&=0.547\ 14,\end{aligned}$$

即 $\sin 0.578\ 91\approx 0.547\ 14$. 误差为

$$R_2(x)=\frac{h^3}{3!}(t-1)(t-2)(-\cos\xi),\quad 0.5<\xi<0.7,$$

$$|R_2(x)|\leqslant 3.36\times 10^{-5}|\cos 0.5|=2.95\times 10^{-5}.$$

若用 Newton 向后插值公式,则可取 $x_0=0.4$,$x_1=0.5$,$x_2=0.6$,$x=0.578\ 91$,$h=0.1$,$t=\dfrac{x-x_2}{h}=-0.210\ 9$. 于是

$$N_2(0.578\ 91)=f_2+t\nabla f_2+\frac{1}{2}t(t+1)\nabla^2 f_2$$

$$=0.564\ 64+0.085\ 21t+\frac{1}{2}t(t+1)\times 0.004\ 80$$

$$=0.547\ 07,$$

即 $\sin 0.578\ 91\approx 0.547\ 07$. 误差为

$$R_2(x)=\frac{h^3}{3!}t(t+1)(t+2)(-\cos\xi),\quad 0.4<\xi<0.6,$$

$$|R_2(x)|\leqslant 4.57\times 10^{-5}.$$

2.4 Hermite 插值多项式

Hermite 插值是带导数的插值. 除了要求插值多项式与被插函数在插值节点上取值相等外,还要求在节点上它们的导数值也相等,甚至要求高阶导数值也相等. 下面只讨论在插值节点上函数值和函数的一阶导数值给定的情形.

设在 $n+1$ 个不同的插值节点 $x_0,x_1,\cdots,x_n$ 上,给定 $y_i=f(x_i),m_i=f'(x_i),i=0,1,\cdots,n$. 要求一个次数不超过 $2n+1$ 的多项式 $H_{2n+1}(x)$,使得满足插值条件

$$H_{2n+1}(x_i)=y_i,\ H'_{2n+1}(x_i)=m_i,\quad i=0,1,\cdots,n. \tag{2.27}$$

满足这种插值条件的多项式称为 Hermite 插值多项式.

Hermite 插值多项式可用类似于求 Lagrange 插值多项式的方法给出,这种插值多项式是唯一存在的.

先求出插值基函数 $\alpha_i(x),\beta_i(x),i=0,1,\cdots,n$,每个基函数为 $2n+1$ 次多项式,并且满足如下条件

$$\begin{cases}\alpha_i(x_k)=\delta_{ik},\ \alpha_i'(x_k)=0,\\ \beta_i(x_k)=0,\ \beta_i'(x_k)=\delta_{ik},\quad k=0,1,\cdots,n.\end{cases} \tag{2.28}$$

利用 $\alpha_i(x),\beta_i(x),i=0,1,\cdots,n$,构造多项式

$$H_{2n+1}(x)=\sum_{i=0}^{n}(y_i\alpha_i(x)+m_i\beta_i(x)). \tag{2.29}$$

这是一个次数不超过 $2n+1$ 的多项式,由条件(2.28)式知,$H_{2n+1}(x)$是满足插值条件(2.27)式的 Hermite 插值多项式.

下面来确定 $\alpha_i(x),\beta_i(x),i=0,1,\cdots,n$. 令

$$\alpha_i(x)=(ax+b)l_i^2(x),\quad i=0,1,\cdots,n,$$

其中 $l_i(x)$为 Lagrange 插值基函数,由(2.5)式给出. 由条件(2.28)式得

$$\begin{cases}ax_i+b=1,\\ a+2l_i'(x_i)=0.\end{cases}$$

由此得

$$\alpha_i(x)=(1-2(x-x_i)l_i'(x_i))l_i^2(x),\quad i=0,1,\cdots,n. \tag{2.30}$$

同理可得

$$\beta_i(x)=(x-x_i)l_i^2(x),\quad i=0,1,\cdots,n. \tag{2.31}$$

下面讨论唯一性问题,设还有一个次数不超过 $2n+1$ 的多项式 $G_{n+1}(x)$ 满足插值条件(2.27)式. 令 $R(x)=H_{2n+1}(x)-G_{2n+1}(x)$,则由(2.27)式有

$$R(x_i)=R'(x_i)=0,\quad i=0,1,\cdots,n.$$

因此,$R(x)$是一个次数不超过 $2n+1$ 的多项式,且它有 $n+1$ 个二重根 $x_0,x_1,\cdots,x_n$,即有 $2n+2$ 个根. 所以,根据多项式的性质,必有 $R(x)=0$,即 $H_{2n+1}(x)=G_{2n+1}(x)$.

仿照 Lagrange 插值余项的证明方法,可得下面的余项定理.

定理 2.2　设 $x_0,x_1,\cdots,x_n$ 为$[a,b]$上相异节点,$f(x)\in C^{2n+1}[a,b]$,并且 $f^{(2n+2)}(x)$在(a,b)内存在,$H_{n+1}(x)$是满足插值条件(2.27)的插值多项式,则对任何 $x\in[a,b]$,存在 $\xi(a,b)$,使得

$$R_{2n+1}(x)=f(x)-H_{2n+1}(x)=\frac{f^{(2n+2)}(\xi)}{(2n+2)!}\omega_{n+1}^2(x). \tag{2.32}$$

3 次 Hermite 插值多项式在应用上特别重要,现列出它的计算公式. 取节点 x_k 和 x_{k+1},3 次 Hermite 插值多项式 $H_3(x)$满足插值条件

$$H_3(x_i)=y_i,\ H_3'(x_i)=m_i,\quad i=k,k+1.$$

相应的插值基函数为

$$\begin{cases}\alpha_k(x)=\left(1+2\dfrac{x-x_k}{x_{k+1}-x_k}\right)\left(\dfrac{x-x_{k+1}}{x_k-x_{k+1}}\right)^2,\\ \alpha_{k+1}(x)=\left(1+2\dfrac{x-x_{k+1}}{x_k-x_{k+1}}\right)\left(\dfrac{x-x_k}{x_{k+1}-x_k}\right)^2,\end{cases} \tag{2.33}$$

$$\begin{cases}\beta_k(x)=(x-x_k)\left(\dfrac{x-x_{k+1}}{x_k-x_{k+1}}\right)^2,\\ \beta_{k+1}(x)=(x-x_{k+1})\left(\dfrac{x-x_k}{x_{k+1}-x_k}\right)^2.\end{cases} \tag{2.34}$$

于是,插值多项式为

$$H_3(x)=y_k\alpha_k(x)+y_{k+1}\alpha_{k+1}(x)+m_k\beta_k(x)+m_{k+1}\beta_{k+1}(x). \tag{2.35}$$

其余项 $R_3(x)=f(x)-H_3(x)$,由(2.32)式得

$$R_3(x)=\frac{1}{4!}f^{(4)}(\xi)(x-x_k)^2(x-x_{k+1})^2, \tag{2.36}$$

其中 ξ 位于 x_k,x_{k+1}和 x 所界定的范围内.

例 2.9　设 $f(x)=\ln x$,给定 $f(1)=0$,$f(2)=0.693\ 147$,$f'(1)=1$,$f'(2)=0.5$. 用 3 次 Hermite 插值多项式 $H_3(x)$计算 $f(1.5)$的近似值.

解　记 $x_0=1, x_1=2$，利用(2.33)式和(2.34)式可得

$$\alpha_0(x)=(2x-1)(2-x)^2,\quad \alpha_1(x)=(5-2x)(x-1)^2,$$

$$\beta_0(x)=(x-1)(2-x)^2,\quad \beta_1(x)=(x-2)(x-1)^2.$$

利用(2.35)式得 3 次 Hermite 插值多项式

$$H_3(x)=0.693\,147(5-2x)(x-1)^2+(x-1)(2-x)^2+0.5(x-2)(x-1)^2.$$

由此得 $f(1.5)$的近似值 $H_3(1.5)=0.409\,074$.

2.5　分段低次插值

2.5.1　多项式插值的问题

用插值多项式近似被插函数时，并不是插值多项式的次数越高越好. 下面是说明这种现象的一个典型例子.

例 **2.10**　给定函数

$$f(x)=\frac{1}{1+x^2},\quad -5\leqslant x\leqslant 5,$$

取等距插值节点 $x_k=-5+10\,\frac{k}{n}$，$k=0,1,\cdots,n$，构造 n 次 Lagrange 插值多项式

$$L_n(x)=\sum_{i=0}^{n}\frac{1}{1+x_i^2}\,\frac{\omega_{n+1}(x)}{(x-x_i)\omega_{n+1}'(x_i)}.$$

当 $n=10$ 时，10 次插值多项式 $L_{10}(x)$以及函数 $f(x)$的图形如图 2-1. 由此可见，$L_{10}(x)$的截断误差 $R_{10}(x)=f(x)-L_{10}(x)$在区间$[-5,5]$的两端非常大. 例如，$L_{10}(4.8)=1.804\,38$，而 $f(4.8)=0.041\,60$. 这种现象称为 Runge 现象. 不管 n 取多大，Runge 现象依然存在.

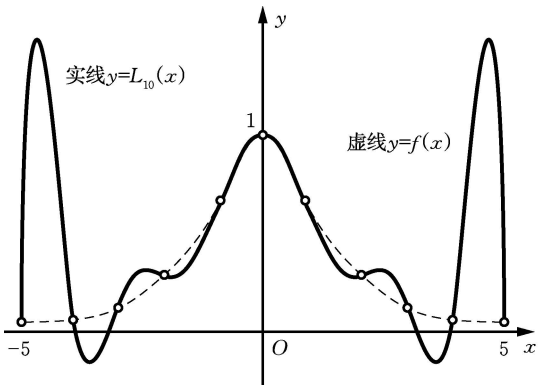

图 2-1

因此，对函数作插值多项式时，必须小心处理，不能认为插值多项式的次数越高，插值余项就越小. 此外，当节点增多时，舍入误差的影响不能低估. 为了克服高次插值的不足，采用分段低次插值将是理论和实际应用的一个良好的插值方法.

2.5.2　分段线性插值

分段线性插值就是通过相邻两个插值点作线性插值来构成的. 设已知节点 $a=x_0<x_1<\cdots<x_n=b$ 上的函数值 $f_k=f(x_k)$，$k=0,1,\cdots,n$. 记 $h_k=x_{k+1}-x_k$，$h=\max\limits_{0\leqslant k\leqslant n-1}h_k$. 若函数 $I_n(x)$ 满足条件：

(1) $I_n(x)\in C[a,b]$；

(2) $I_n(x_k)=f_k$，$k=0,1,\cdots,n$；

(3) 在每个小区间 $[x_k,x_{k+1}]$($k=0,1,\cdots,n-1$)上，$I_n(x)$ 是线性多项式，

则称 $I_n(x)$ 为分段线性插值函数.

分段线性插值函数 $I_n(x)$ 的几何意义是通过 $n+1$ 个点 (x_i,f_i)($i=0,1,\cdots,n$)的折线. 在每个小区间 $[x_k,x_{k+1}]$($k=0,1,\cdots,n-1$)上，$I_n(x)$ 的表示式为

$$I_n(x)=\frac{x-x_{k+1}}{x_k-x_{k+1}}f_k+\frac{x-x_k}{x_{k+1}-x_k}f_{k+1}. \tag{2.37}$$

若用插值基函数表示，则在整个区间 $[a,b]$ 上，$I_n(x)$ 的表示式为

$$I_n(x)=\sum_{i=0}^{n}f_i l_i(x),$$

插值基函数 $l_i(x)$($i=0,1,\cdots,n$)的形式是

$$l_i(x)=\begin{cases}\dfrac{x-x_{i-1}}{x_i-x_{i-1}}, & x\in[x_{i-1},x_i],\\[2ex] \dfrac{x-x_{i+1}}{x_i-x_{i+1}}, & x\in[x_i,x_{i+1}],\\[2ex] 0, & \text{其他},\end{cases}$$

其中，当 $i=0$ 时，没有第一式，当 $i=n$ 时，没有第二式. 显然，分段线性插值基函数 $l_i(x)$ 只在 x_i 的附近不为零，在其他地方均为零，这种性质称为局部非零性质.

例 2.11　对例 2.10 中的函数 $f(x)$，已知它在区间 $[0,5]$ 上取等距节点处的函数值为表 2-9，求区间 $[0,5]$ 上的分段线性插值函数，并利用它求出 $f(4.5)$ 的近似值.

表 2－9

x_i	0	1	2	3	4	5
y_i	1	0.5	0.2	0.1	0.058 82	0.038 46

解　在每个小区间$[i,i+1]$上，由(2.37)式得分段线性插值函数

$$I(x)=(i+1-x)y_i+(x-i)y_{i+1},\quad i=0,1,2,3,4.$$

于是

$$f(4.5)\approx I(4.5)=0.058\,82(5-4.5)+0.038\,46(4.5-4)=0.048\,64.$$

分段线性插值函数的余项可以通过线性插值多项式的余项来估计.

定理 2.3　如果 $f(x)\in C^2[a,b]$，记 $M_2=\max\limits_{a\leqslant x\leqslant b}|f''(x)|$，则对任意 $x\in[a,b]$，分段线性插值函数 $I_n(x)$有余项估计

$$|f(x)-I_n(x)|\leqslant\frac{h^2}{8}M_2. \tag{2.38}$$

证　根据(2.10)式，在每个小区间$[x_k,x_{k+1}]$$(k=0,1,\cdots,n-1)$上有

$$|f(x)-I_n(x)|\leqslant\frac{1}{8}(x_{k+1}-x_k)^2\max_{x_k\leqslant x\leqslant x_{k+1}}|f''(x)|.$$

因此，在整个区间$[a,b]$上有

$$|f(x)-I_n(x)|\leqslant\frac{h^2}{8}M_2.$$

该定理也说明分段线性插值函数 $I_n(x)$具有一致收敛性. 于是可以加密插值节点，缩小插值区间，使 h 减小，从而减小插值误差.

例 2.12　对平方根表作线性插值，已知 $10\leqslant x\leqslant 999$，步长 $h=1$. 试给出按插值方法求$\sqrt{x}$的近似值的误差界，并估计有效数字的位数，假定表上已给的函数值足够精确.

解　令 $f(x)=\sqrt{x}$，$M=\max|f''(x)|$，则由(2.37)式知截断误差

$$|R(x)|\leqslant\frac{1}{8}Mh^2.$$

分两段讨论$|R(x)|$.

(1) 当 $10\leqslant x\leqslant 100$ 时，

$$|f''(x)|=\frac{1}{4x^{\frac{3}{2}}}\leqslant\frac{1}{4\times 10^{\frac{3}{2}}}\approx 0.007\,9=M,$$

$$|R(x)|\leqslant\frac{1}{8}\times 0.007\,9\approx 0.000\,99.$$

由于 $3\leqslant\sqrt{x}<10$，故$\sqrt{x}$可以具有 3 位有效数字.

(2) 当 $100\leqslant x\leqslant 999$ 时

$$|f''(x)|=\frac{1}{4x^{\frac{3}{2}}}\leqslant\frac{1}{4\times100^{\frac{3}{2}}}\approx0.25\times10^{-3}=M,$$

$$|R(x)|\leqslant\frac{1}{8}\times0.25\times10^{-3}=0.000\ 031\ 3.$$

由于 $10\leqslant x\leqslant32$，故$\sqrt{x}$可以具有 6 位有效数字.

2.5.3　分段 3 次 Hermite 插值

分段线性插值函数具有良好的一致收敛性，但它是不光滑的，它在节点处的左右导数不相等. 为了克服这个缺陷，一个自然的想法是添加一阶导数的插值条件.

设已给节点 $a=x_o<x_1<\cdots<x_n=b$ 上的函数值和导数值

$$f_k=f(x_k),\ m_k=f'(x_k),\quad k=0,1,\cdots,n.$$

记 $h=\max\limits_{0\leqslant k\leqslant n-1}(x_{k+1}-x_k)$. 如果函数 $I_n(x)$满足条件：

(1) $I_n(x)\in C^1[a,b]$；

(2) $I_n(x_k)=f_k$，$I_n{}'(x_k)=m_k$，$k=0,1,\cdots,n$；

(3) 在每个小区间$[x_k,x_{k+1}]$($k=0,1,\cdots,n-1$)上，$I_n(x)$是 3 次多项式，则称 $I_n(x)$为分段 3 次 Hermite 插值函数.

显然，在每个小区间$[x_k,x_{k+1}]$($k=0,1,\cdots,n-1$)上，$I_n(x)$的表示式为(2.35). 可以直接用它进行数值计算.

若用插值基函数表示，则在整个区间$[a,b]$上，$I_n(x)$的表示式为

$$I_n(x)=\sum_{i=0}^{n}(f_i\alpha_i(x)+m_i\beta_i(x)). \tag{2.39}$$

插值基函数 $\alpha_i(x)$和 $\beta_i(x)$的形式分别是

$$\alpha_i(x)=\begin{cases}\left(1+2\dfrac{x-x_i}{x_{i-1}-x_i}\right)\left(\dfrac{x-x_{i-1}}{x_i-x_{i-1}}\right)^2, & x\in[x_{i-1},x_i],\\ \left(1+2\dfrac{x-x_i}{x_{i+1}-x_i}\right)\left(\dfrac{x-x_{i+1}}{x_i-x_{i+1}}\right)^2, & x\in[x_i,x_{i+1}],\\ 0, & \text{其他.}\end{cases} \tag{2.40}$$

$$\beta_i(x)=\begin{cases}(x-x_i)\left(\dfrac{x-x_{i-1}}{x_i-x_{i-1}}\right)^2, & x\in[x_{i-1},x_i],\\ (x-x_i)\left(\dfrac{x-x_{i+1}}{x_i-x_{i+1}}\right)^2, & x\in[x_i,x_{i+1}],\\ 0, & \text{其他.}\end{cases} \tag{2.41}$$

其中，当 $i=0$ 时，上述两个分段函数没有第一式；当 $i=n$ 时，上述两个分段函数没有第二式. 显然，(2.40)式和(2.41)式具有局部非零性质，这种性质使得

(2.39)式也可写成分段表示式(2.35)的形式.

例 2.13 已知函数 $f(x)=\dfrac{1}{1+x^2}$ 在区间[0,3]上取等距节点处的函数值如表 2-10,求区间[0,3]上的分段 3 次 Hermite 插值函数,并利用它求 $f(1.5)$ 的近似值.

表 2-10

x_i	0	1	2	3
y_i	1	0.5	0.2	0.1
m_i	0	−0.5	−0.16	−0.06

解 在每个小区间[i,$i+1$]上,由(2.35)式得

$$\begin{aligned}I(x)=&(1+2(x-i))(x-i-1)^2y_i\\&+(1-2(x-i-1))(x-i)^2y_{i+1}+(x-i)(x-i-1)^2m_i\\&+(x-i-1)(x-i)^2m_{i+1},\quad i=0,1,2.\end{aligned}$$

于是

$$f(1.5)\approx I(1.5)=0.312\,5.$$

分段 3 次 Hermite 插值函数的余项可以通过前面 3 次 Hermite 插值多项式的余项来估计.

定理 2.4 如果 $f(x)\in C^4[a,b]$,记 $M_4=\max\limits_{a\leqslant x\leqslant b}|f^{(4)}(x)|$,那么对于任意 $x\in[a,b]$,分段 3 次 Hermite 插值函数 $I_n(x)$ 有余项估计

$$|f(x)-I_n(x)|\leqslant\frac{h^4}{384}M_4. \tag{2.42}$$

证 根据(2.36)式,在每个小区间[x_k,x_{k+1}]($k=0,1,\cdots,n-1$)上有

$$|f(x)-I_n(x)|\leqslant\frac{1}{4!}\frac{1}{16}(x_{k+1}-x_k)^4\max_{x_k\leqslant x\leqslant x_{k+1}}|f^{(4)}(x)|.$$

因此,在整个区间[a,b]上有(2.42)式.

该定理除了可以用于误差估计之外,还说明分段 3 次 Hermite 插值函数具有一致收敛性.

2.6 3 次样条插值

2.6.1 3 次样条插值函数的概念

分段 3 次 Hermite 插值函数只有当被插函数在所有插值节点处的函数值和导数值都已知时才能使用,并且该插值函数在内节点处的 2 阶导数一般不连续,

因而插值曲线不是很光滑. 在一些实际问题中,我们不可能也没有必要已知被插值函数在内节点处的导数值. 本节将讨论在科学和工程计算中起到重要作用的一种分段 3 次插值,它只在插值区间的端点比 Lagrange 插值多两个边界条件,但却在内节点处 2 阶导数连续.

样条这一名词来源于工程制图. 以前,绘图员为了将一些指定点(称作样点)连接成一条光滑曲线,往往把富有弹性的细长木条(称为样条)固定在样点上,然后画下木条表示的曲线所形成的样条曲线. 下面用数学语言来描述 3 次样条插值函数的概念.

设在区间$[a,b]$上取 $n+1$ 个节点 $a=x_0<x_1<\cdots<x_n=b$,给定这些点的函数值 $f(x_i)=f_i(i=0,1,\cdots,n)$. 若 $S(x)$满足条件:

(1) $S(x)\in C^2[a,b]$;

(2) $S(x_i)=f_i,i=0,1,\cdots,n$;

(3) 在每个小区间$[x_i,x_{i+1}]$上,$S(x)$是一个 3 次多项式,

则称 $S(x)$为 $f(x)$在$[a,b]$上的 **3** 次样条插值函数.

3 次样条插值函数是分段 3 次多项式,在每个小区间$[x_i,x_{i+1}]$上可以写成

$$S(x)=a_ix^3+b_ix^2+c_ix+d_i,\quad i=0,1,\cdots,n-1,$$

其中 a_i,b_i,c_i 和 d_i 为待定系数. 所以,$S(x)$共有 $4n$ 个待定参数. 根据 $S(x)$在$[a,b]$上 2 阶导数连续的条件,在节点 $x_i(i=1,2,\cdots,n-1)$处应满足连续性条件

$$S^{(k)}(x_i-0)=S^{(k)}(x_i+0),\quad k=0,1,2,$$

共有 $3(n-1)$个条件. 再加上 $n+1$ 个插值条件,共有 $4n-2$ 个条件. 因此,还需要两个条件才能确定 $S(x)$. 通常在区间$[a,b]$端点 $a=x_0$ 和 $b=x_n$ 上各加一个条件(称为边界条件),可根据实际问题的要求给定. 常用的有以下 3 种:

(1) 已知两端点的一阶导数值,即

$$S'(x_0)=f_0',\ S'(x_n)=f_n'. \tag{2.43}$$

(2) 已知两端点的 2 阶导数值,即

$$S''(x_0)=f_0'',\ S''(x_n)=f_n''. \tag{2.44}$$

其特殊情况为

$$S''(x_0)=0,\ S''(x_n)=0. \tag{2.45}$$

条件(2.45)称为自然边界条件.

(3) 周期边界条件

$$S^{(k)}(x_0)=S^{(k)}(x_n),\quad k=0,1,2. \tag{2.46}$$

此时,对函数值有周期条件 $f_0=f_n$.

2.6.2　三弯矩算法

3 次样条插值函数 $S(x)$可以有多种表达方式,有时用二阶导数值 $S''(x_i)=$

$M_i(i=0,1,\cdots,n)$表示时，使用更方便. M_i 在力学上解释为细梁在 x_i 处的弯矩，并且得到的弯矩与相邻两个弯矩有关，故称用 M_i 表示 $S(x)$的算法为三弯矩算法.

由于 $S(x)$在区间$[x_i,x_{i+1}](i=0,1,\cdots,n-1)$上是 3 次多项式，故 $S''(x)$在$[x_i,x_{i+1}]$上是线性函数，可表示为

$$S''(x)=M_i\frac{x_{i+1}-x}{h_i}+M_{i+1}\frac{x-x_i}{h_i},$$

其中 $h_i=x_{i+1}-x_i$. 对 $S''(x)$积分两次，并利用插值条件 $S(x_i)=f_i$，$S(x_{i+1})=f_{i+1}$定出积分常数，可以得到

$$S(x)=M_i\frac{(x_{i+1}-x)^3}{6h_i}+M_{i+1}\frac{(x-x_i)^3}{6h_i}+\left(f_i-\frac{M_ih_i^2}{6}\right)\frac{x_{i+1}-x}{h_i}$$
$$+\left(f_{i+1}-\frac{M_{i+1}h_i^2}{6}\right)\frac{x-x_i}{h_i},\quad x\in[x_i,\quad x_{i+1}]. \tag{2.47}$$

这是 3 次样条插值函数的表达式，当求出 $M_i(i=0,1,\cdots,n)$后，$S(x)$就由(2.47)式完全确定.

下面推导 $M_i(i=0,1,\cdots,n)$所要满足的条件. 对 $S(x)$求导得

$$S'(x)=-M_i\frac{(x_{i+1}-x)^2}{2h_i}+M_{i+1}\frac{(x-x_i)^2}{2h_i}+f(x_i,x_{i+1})-\frac{h_i}{6}(M_{i+1}-M_i),$$

由此可得

$$S'(x_i+0)=f(x_i,x_{i+1})-\frac{h_i}{6}(2M_i+M_{i+1}),$$

$$S'(x_{i+1}-0)=f(x_i,x_{i+1})+\frac{h_i}{6}(M_i+2M_{i+1}).$$

当 $x\in[x_{i-1},x_i]$时，$S(x)$的表达式由(2.47)式平移下标可得，因此有

$$S'(x_i-0)=f(x_{i-1},x_i)+\frac{h_{i-1}}{6}(M_{i-1}+2M_i).$$

利用条件 $S'(x_i+0)=S'(x_i-0)$得

$$\mu_iM_{i-1}+2M_i+\lambda_iM_{i+1}=d_i,\quad i=1,2,\cdots,n-1, \tag{2.48}$$

其中

$$\mu_i=\frac{h_{i-1}}{h_{i-1}+h_i},\quad \lambda_i=\frac{h_i}{h_{i-1}+h_i}=1-\mu_i, \tag{2.49}$$

$$d_i=6f(x_{i-1},x_i,x_{i+1}). \tag{2.50}$$

方程组(2.48)是关于 M_i 的方程组，有 $n+1$ 个未知数，但只有 $n-1$ 个方程. 可由式(2.43)—(2.46)的任一种边界条件补充两个方程.

对于边界条件式(2.43)，由 $S'(x)$的表达式可以导出两个方程

$$\begin{cases} 2M_0+M_1=\dfrac{6}{h_0}(f(x_0,x_1)-f_0'), \\ M_{n-1}+2M_n=\dfrac{6}{h_{n-1}}(f_n'-f(x_{n-1},x_n)). \end{cases} \tag{2.51}$$

这样，由(2.48)式和(2.51)式可解出 $M_i(i=0,1,\cdots,n)$，从而得 $S(x)$ 的表达式(2.47)，若令 $\lambda_0=\mu_n=1$，$d_0=\dfrac{6}{h_0}(f(x_0,x_1)-f_0')$，$d_n=\dfrac{6}{h_{n-1}}(f_n'-f(x_{n-1},x_n))$，则(2.48)式和(2.51)式可以写成矩阵形式

$$\begin{pmatrix} 2 & \lambda_0 & & & \\ \mu_1 & 2 & \lambda_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_{n-1} & 2 & \lambda_{n-1} \\ & & & \mu_n & 2 \end{pmatrix} \begin{pmatrix} M_0 \\ M_1 \\ \vdots \\ M_{n-1} \\ M_n \end{pmatrix} = \begin{pmatrix} d_0 \\ d_1 \\ \vdots \\ d_{n-1} \\ d_n \end{pmatrix}. \tag{2.52}$$

对于边界条件(2.44)式，直接得

$$M_0=f_0'', \quad M_n=f_n''. \tag{2.53}$$

将(2.53)式代入(2.48)式可解出 $M_i(i=1,2,\cdots,n-1)$. 如果令 $\lambda_0=\mu_n=0$，$d_0=2f_0''$，$d_n=2f_n''$，则(2.48)式和(2.53)式也可写成(2.52)式的形式.

对于边界条件(2.46)，有

$$\begin{cases} M_0=M_n, \\ \lambda_n M_1+\mu_n M_{n-1}+2M_n=d_n, \end{cases} \tag{2.54}$$

其中　　$\lambda_n=h_0(h_{n-1}+h_0)^{-1}$，$\mu_n=1-\lambda_n=h_{n-1}(h_{n-1}+h_0)^{-1}$，

$d_n=6(f[x_0,x_1]-f[x_{n-1},x_n])(h_0+h_{n-1})^{-1}$.

由(2.48)式和(2.54)式可解出 $M_i(i=0,1,\cdots,n)$，方程组的矩阵形式为

$$\begin{pmatrix} 2 & \lambda_1 & & & \mu_1 \\ \mu_2 & 2 & \lambda_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_{n-1} & 2 & \lambda_{n-1} \\ \lambda_n & & & \mu_n & 2 \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ M_{n-1} \\ M_n \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-1} \\ d_n \end{pmatrix}. \tag{2.55}$$

实际上，方程组(2.52)和(2.55)的系数矩阵是一类特殊的矩阵，在后面线性方程组的解法中，将专门介绍这类方程组的解法和性质.

例 2.14　设在节点 $x_i=i(i=0,1,2,3)$ 上，函数 $f(x)$ 的值为 $f(x_0)=0$，$f(x_1)=0.5$，$f(x_2)=2$，$f(x_3)=1.5$. 试求 3 次样条插值函数 $S(x)$，满足条件

(1) $S'(x_0)=0.2$，$S'(x_3)=-1$；

(2) $S''(x_0)=-0.3$，$S''(x_3)=3.3$.

解 (1) 利用方程组(2.52)进行求解,可知 $h_i=1(i=0,1,2)$,$\lambda_0=1$,$\mu_3=1$,$\lambda_1=\lambda_2=\mu_1=\mu_2=0.5$.经简单计算有 $d_0=1.8$,$d_1=3$,$d_2=-6$,$d_3=-3$.由此得(2.52)形式的方程组

$$\begin{pmatrix} 2 & 1 & & \\ 0.5 & 2 & 0.5 & \\ & 0.5 & 2 & 0.5 \\ & & 1 & 2 \end{pmatrix}\begin{pmatrix} M_0 \\ M_1 \\ M_2 \\ M_3 \end{pmatrix}=\begin{pmatrix} 1.8 \\ 3 \\ -6 \\ -3 \end{pmatrix}.$$

先消去 M_0 和 M_3 得

$$\begin{bmatrix} 3.5 & 1 \\ 1 & 3.5 \end{bmatrix}\begin{bmatrix} M_1 \\ M_2 \end{bmatrix}=\begin{bmatrix} 5.1 \\ -10.5 \end{bmatrix}.$$

由此解得 $M_1=2.52$,$M_2=-3.72$.代回方程组得 $M_0=-0.36$,$M_3=0.36$.

用 M_0,M_1,M_2,M_3 的值代入 3 次样条插值函数的表达式(2.47)中,经化简有

$$S(x)=\begin{cases} 0.48x^3-0.18x^2+0.2x, & x\in[0,1], \\ -1.04(x-1)^3+1.26(x-1)^2+1.28(x-1)+0.5, & x\in[1,2], \\ 0.68(x-2)^3-1.86(x-2)^2+0.68(x-2)+2, & x\in[2,3]. \end{cases}$$

(2) 仍用方程组进行求解,不过要注意 λ_0,μ_3,d_0,d_3 的不同.由于 M_0 和 M_3 已知,故可以化简得

$$\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}\begin{bmatrix} M_1 \\ M_2 \end{bmatrix}=\begin{bmatrix} 6.3 \\ -15.3 \end{bmatrix}.$$

由此解得 $M_1=2.7$,$M_2=-4.5$.将 $M_i(i=0,1,2,3)$代入 3 次样条插值函数的表达式(2.47)中,经化简有

$$S(x)=\begin{cases} 0.5x^3-0.15x^2+0.15x, & x\in[0,1], \\ -1.2(x-1)^3+1.35(x-1)^2+1.35(x-1)+0.5, & x\in[1,2], \\ 1.3(x-2)^3-2.25(x-2)^2+0.45(x-2)+2, & x\in[2,3]. \end{cases}$$

2.6.3 三转角算法

下面构造用一阶导数值 $S'(x_i)=m_i(i=0,1,\cdots,n)$表示的 3 次样条插值函数.$m_i$ 在力学上解释为细梁在 x_i 截面处的转角,并且得到的转角与相邻两个转角有关,故称用 m_i 表示 $S(x)$的算法为三转角算法.

根据 Hermite 插值函数的唯一性和表达式(2.33)—(2.35),可设 $S(x)$在区间$[x_i,x_{i+1}](i=0,1,\cdots,n-1)$上的表达式为

$$S(x)=\frac{(h_i+2(x-x_i))(x-x_{i+1})^2}{h_i^3}f_i+\frac{(h_i+2(x_{i+1}-x))(x-x_i)^2}{h_i^3}f_{i+1}$$
$$+\frac{(x-x_i)(x-x_{i+1})^2}{h_i^2}m_i+\frac{(x-x_{i+1})(x-x_i)^2}{h_i^2}m_{i+1}. \tag{2.56}$$

对 $S(x)$求 2 次导数得

$$S''(x)=\frac{6x-2x_i-4x_{i+1}}{h_i^2}m_i+\frac{6x-4x_i-2x_{i+1}}{h_i^2}m_{i+1}$$
$$+\frac{6(x_i+x_{i+1}-2x)}{h_i^3}(f_{i+1}-f_i).$$

于是有

$$S''(x_i+0)=-\frac{4}{h_i}m_i-\frac{2}{h_i}m_{i+1}+\frac{6}{h_i^2}(f_{i+1}-f_i).$$

同理,考虑 $S(x)$在$[x_{i-1},x_i]$上的表达式,可以得到

$$S''(x_i-0)=\frac{2}{h_{i-1}}m_{i-1}+\frac{4}{h_{i-1}}m_i-\frac{6}{h_{i-1}^2}(f_i-f_{i-1}).$$

利用条件 $S''(x_i+0)=S''(x_i-0)$,得

$$\lambda_i m_{i-1}+2m_i+\mu_i m_{i+1}=e_i,\quad i=1,2,\cdots,n-1. \tag{2.57}$$

其中,λ_i,μ_i 由(2.49)式所示,而

$$e_i=3(\lambda_i f(x_{i-1},x_i)+\mu_i f(x_i,x_{i+1})). \tag{2.58}$$

方程组(2.57)是关于 m_i 的方程组,有 $n+1$ 个未知数,$n-1$ 个方程. 可由(2.43)—(2.46)式的任一种边界条件补充两个方程.

对于边界条件(2.43)式,则 $m_0=f_0'$,$m_n=f_n'$. 由(2.57)式,$m_1,m_2,\cdots,m_{n-1}$满足方程组

$$\begin{pmatrix}2&\mu_1&&&\\ \lambda_2&2&\mu_2&&\\ &\ddots&\ddots&\ddots&\\ &&\lambda_{n-2}&2&\mu_{n-2}\\ &&&\lambda_{n-1}&2\end{pmatrix}\begin{pmatrix}m_1\\ m_2\\ \vdots\\ m_{n-2}\\ m_{n-1}\end{pmatrix}=\begin{pmatrix}e_1-\lambda_1 f_0'\\ e_2\\ \vdots\\ e_{n-2}\\ e_{n-1}-\mu_{n-1}f_n'\end{pmatrix}. \tag{2.59}$$

由此可解得 $m_1,m_2,\cdots,m_{n-1}$,从而得 $S(x)$的表达式(2.56).

对于边界条件(2.44)式,则可导出两个方程

$$\begin{cases}2m_0+m_1=3f(x_0,x_1)-\dfrac{h_0}{2}f_0'',\\ m_{n-1}+2m_n=3f(x_{n-1},x_n)+\dfrac{h_{n-1}}{2}f_n''.\end{cases} \tag{2.60}$$

由(2.57)式和(2.60)式可解出 $m_i(i=0,1,\cdots,n)$. 若令 $e_0=3f(x_0,x_1)-\dfrac{h_0}{2}f_0''$,

$e_n=3f(x_{n-1},x_n)+\frac{h_{n-1}}{2}f_n''$，则(2.57)式和(2.60)式可合并成矩阵形式

$$\begin{pmatrix} 2 & 1 & & & \\ \lambda_1 & 2 & \mu_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \lambda_{n-1} & 2 & \mu_{n-1} \\ & & & 1 & 2 \end{pmatrix}\begin{pmatrix} m_0 \\ m_1 \\ \vdots \\ m_{n-1} \\ m_n \end{pmatrix}=\begin{pmatrix} e_0 \\ e_1 \\ \vdots \\ e_{n-1} \\ e_n \end{pmatrix}. \tag{2.61}$$

对于边界条件(2.46)，可得

$$\begin{cases} m_0=m_n, \\ \mu_0 m_1+\lambda_n m_{n-1}+2m_n=e_n, \end{cases} \tag{2.62}$$

其中 $\mu_n=h_{n-1}(h_0+h_{n-1})^{-1}$，$\lambda_n=h_0(h_0+h_{n-1})^{-1}$，$e_n=3(\mu_n f(x_0,x_1)+\lambda_n f(x_{n-1},x_n))$. 由(2.57)式和(2.62)式可解得 $m_i(i=0,1,\cdots,n)$，方程组的矩阵形式为

$$\begin{pmatrix} 2 & \mu_1 & & & \lambda_1 \\ \lambda_2 & 2 & \mu_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \lambda_{n-1} & 2 & \mu_{n-2} \\ \mu_n & & & \lambda_n & 2 \end{pmatrix}\begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_{n-1} \\ m_n \end{pmatrix}=\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{n-1} \\ e_n \end{pmatrix}. \tag{2.63}$$

例 2.15 给定数据如表 2－11. 求满足边界条件 $S'(0)=1,S'(3)=0$ 的3 次样条插值函数 $S(x)$.

表 2－11

x_i	0	1	2	3
f_i	0	0	0	0

解 取 x_i 处的一阶导数 $m_i(i=1,2)$作为参数，由

$$\lambda_i=\frac{h_i}{h_{i-1}+h_i}=\frac{1}{2},\quad \mu_i=1-\lambda_i=\frac{1}{2},$$

$$e_i=3(\lambda_i f(x_{i-1},x_i)+\mu_i f(x_i,x_{i+1}))=0,$$

以及(2.57)式，得

$$\begin{cases} \frac{1}{2}m_0+2m_1+\frac{1}{2}m_2=0, \\ \frac{1}{2}m_1+2m_2+\frac{1}{2}m_3=0. \end{cases}$$

将 $m_0=1,m_3=0$ 代入上面的方程组，得

$$\begin{cases} 4m_1+m_2=-1, \\ m_1+4m_2=0. \end{cases}$$

解得 $m_1=-\dfrac{4}{15}$，$m_2=\dfrac{1}{15}$.

利用表达式(2.56)得

$$S(x)=\begin{cases} \dfrac{1}{15}x(x-1)(15-11x), & x\in[0,1], \\ \dfrac{1}{15}(x-1)(x-2)(7-3x), & x\in[1,2], \\ \dfrac{1}{15}(x-3)^2(x-2), & x\in[2,3]. \end{cases}$$

2.6.4　3 次样条插值函数的误差估计

在实际应用中，如果不需要规定内节点处的一阶导数值，那么使用 3 次样条插值函数会得到很好的效果. 3 次样条插值函数 $S(x)$不仅在内节点处二阶导数是连续的，而且 $S(x)$逼近 $f(x)$具有很好的收敛性，也是数值稳定的. 由于误差估计与收敛性定理的证明较复杂，下面只给出误差估计的结论.

定理 2.5　设函数 $f(x)\in C^4[a,b]$，记 $M_4=\max\limits_{a\leqslant x\leqslant b}|f^{(4)}(x)|$，$h=\max\limits_{0\leqslant i\leqslant n-1}(x_{i+1}-x_i)$，则对任意 $x\in[a,b]$，满足边界条件(2.43)式或(2.44)式的 3 次样条插值函数 $S(x)$有估计式

$$|f^{(k)}(x)-S^{(k)}(x)|\leqslant C_k h^{4-k}M_4, \quad k=0,1,2, \tag{2.64}$$

其中 $C_0=\dfrac{5}{384}$，$C_4=\dfrac{1}{24}$，$C_2=\dfrac{1}{8}$.

误差估计式(2.64)除了可以用于误差估计之外，它进一步表明，当 $f(x)\in C^4[a,b]$时，在插值区间$[a,b]$上，对于满足边界条件(2.43)式或(2.44)式的插值函数 $S(x)$，不仅 $S(x)$一致收敛于 $f(x)$，而且 $S'(x)$一致收敛于 $f'(x)$，$S''(x)$一致收敛于 $f''(x)$.

评　　注

插值函数是数值分析的基本工具，是函数逼近、数值积分、数值微分和微分方程数值解的基础. 插值法的基本思想就是如何用一个多项式函数、分段多项式函数或者较简单的函数来逼近某个较复杂的. 也可能没有表达式的函数，使得在插值节点处满足给定的插值条件.

本章按插值函数的特征，分别介绍了多项式插值、分段多项式插值和 3 次样

条插值. 虽然满足插值条件的插值多项式是唯一存在的，但是构造和利用插值多项式的方式却有很大差别.

Lagrange 插值多项式虽然计算量大，但表示式简单明确，便于理论推导，理论上较重要. 逐次线性插值法不具体给出插值多项式的表达式，可以递推地计算插值点的值. Newton 插值多项式便于逐步增加节点，并且计算过程中能估计误差. 带导数的插值多项式适合于已知导数的情形. 当 n 较大时，这些方法都有数值不稳定的缺陷，所以实际中用得最广的插值方法是分段低次插值法.

分段低次插值具有良好的稳定性和收敛性. 当仅仅知道函数值时，可以采用分段线性插值或 3 次样条插值. 当同时知道函数值和导数值时，可以采用分段 3 次Hermite 插值. 3 次样条插值函数不仅在内节点处二阶导数是连续的，而且具有很好逼近性和收敛性. 至于 B 样条和一般样条函数本书未涉及，需要对样条函数作更深入地了解的读者可参看专门文献.

关于插值误差估计论述了微分形式和均差形式. 对于充分光滑的被插函数，采用微分形式的误差估计可以给出实用的误差界. 均差形式的误差估计虽不能给出确切的误差界，形式上更有一般性，并且在数值积分和数值微分的误差推导中将有重要应用.

习 题 2

2.1 已知 $f(1)=0, f(-1)=-3, f(2)=4$. 求函数 $f(x)$ 过这 3 点的 2 次 Lagrange 插值多项式 $L_2(x)$.

2.2 已知函数 $\ln x$ 的数据如表 2-12，分别用线性插值和 2 次插值求 $\ln 0.54$ 的近似值.

表 2-12

x	0.5	0.6	0.7
$f(x)$	−0.693 147	−0.510 826	−0.356 675

2.3 设 $\{x_i\}_{i=0}^n$ 为互异的插值节点，求证：

$$\sum_{i=0}^{n} x_i^k l_i(x) = x^k, \quad k = 0, 1, \cdots, n,$$

其中 $l_i(x)$ 为 n 次 Lagrange 插值基函数.

2.4 设 $f(x)\in C^2[a,b]$，且 $f(a)=f(b)=0$，求证：

$$\max_{a\leqslant x\leqslant b} |f(x)| \leqslant \frac{1}{8}(b-a)^2 \max_{a\leqslant x\leqslant b} |f''(x)|.$$

2.5　设 $f(x)=(3x-2)e^x$. 求 $f(x)$的关于节点 $x=1,1.05,1.07$ 的 2 次 Lagrange 插值多项式 $L_2(x)$,并估计误差 $R_2(1.03)$.

2.6　设 $f(x)=x^4$,试用 Lagrange 插值求 $f(x)$的以 $-1,0,1,2$ 为插值节点的 3 次插值多项式.

2.7　给定数据如表 2-13. 用 Newton 插值公式求 3 次插值多项式 $N_3(x)$.

表 2-13

x	1	1.5	0	2
$f(x)$	1.25	2.50	1.00	5.50

2.8　设 $f(x)=x^7+x^4+3x+1$,求 $f[2^0,2^1,\cdots,2^7]$和 $f[2^0,2^2,\cdots,2^8]$.

2.9　若 $f(x)=a_0+a_1x+\cdots+a_nx^n$ 有 n 个不同实根 $x_1,x_2,\cdots,x_n$,求证:

$$\sum_{i=1}^{n}\frac{{x_i}^k}{f'(x_i)}=\begin{cases}0, & 0\leqslant k\leqslant n-2,\\ {a_n}^{-1}, & k=n-1.\end{cases}$$

2.10　证明 $\sum_{i=0}^{n-1}\triangle^2 y_i=\triangle y_n-\triangle y_0$.

2.11　在 $-4\leqslant x\leqslant 4$ 上给出 $f(x)=e^x$ 的等距节点函数表. 若用 2 次插值求 e^x 的近似值,要使截断误差不超过 10^{-6},问使用函数表的步长 h 应取多少?

2.12　求不超过 4 次的多项式 $P(x)$,使它满足插值条件

$$P(0)=P'(0)=0,\ P(1)=P'(1)=1,\ P(2)=1.$$

2.13　求不超过 3 次的多项式 $H(x)$,使它满足插值条件.

$$H(-1)=-9,\ H'(-1)=15,\ H(1)=1,\ H'(1)=-1.$$

2.14　求 $f(x)=x^2$ 在 $[a,b]$上的分段线性插值函数 $I_n(x)$,并估计误差.

2.15　求 $f(x)=x^4$ 在 $[a,b]$上的分段 3 次 Hermite 插值函数,并估计误差.

2.16　已知函数的数据如表 2-14,用三弯矩算法在第一类边界条件下求 3 次样条插值多项式 $S(x)$.

表 2-14

x	-1	0	1
$f(x)$	$\frac{5}{3}$	0	1
$f'(x)$	-1		7

数值试验题 2

2.1 编写一个用 Newton 插值公式计算函数值的程序，要求先输出差分表，再计算 x 点的函数值，应用于表 2－15 给出的数据. 求 $x=21.4$ 时的 3 次插值多项式的值.

表 2－15

x_i	20	21	22	23	24
y_i	1.301 03	1.322 22	1.342 42	1.361 73	1.380 21

2.2 对区间$[-5,5]$作等距划分 $x_i=-5+ih$，$h=\dfrac{10}{n}$，$i=0,1,\cdots,n$. 对函数

$$y=\frac{1}{1+x^2},\quad y=\frac{x}{1+x^4},\quad y=\arctan x$$

按下列算法作插值，并分析数值结果.

算法 1：取 $n=10,20$ 作 Lagrange 插值.

算法 2：取 $n=10,20$ 作 3 次自然样条插值.

2.3 编制分段线性插值和分段 3 次 Hermite 插值程序，对函数 $f(x)=\dfrac{1}{1+x^2}$，插值区间$[-5,5]$，分成 10 等分，求分段插值函数在各节点间中点处的值，并画出分段插值函数和 $y=f(x)$的图形.

2.4 给定数据如表 2-16. 编制程序求 3 次样条插值函数在插值节点间中点处的样条函数值，并作点集$\{x_i,y_i\}$和样条插值函数的图形，满足的边界条件为

(1) $S'(0)=0.8$，$S'(10)=0.2$；　　(2) $S''(0)=S''(10)=0$.

表 2－16

x_i	0	1	2	3	4	5	6	7	8	9	10
y_i	0.0	0.79	1.53	2.19	2.71	3.03	3.27	2.89	3.06	3.19	3.29

第 3 章　函数的最佳逼近

在科学实验和统计研究中，往往要从大量的实验数据$(x_i,y_i)(i=0,1,\cdots,m)$中寻找其函数关系 $y=f(x)$的近似表达式 $y=\varphi(x)$. 插值法要求插值曲线严格通过每一个数据点，即在插值点处的误差为零. 考虑到数据不一定准确，并且对大量数据，高次插值多项式容易出现振荡现象，我们可以不必要求近似函数 $\varphi(x)$经过所有的数据点(x_i,y_i)，而只要求其误差 $\delta_i=y_i-\varphi(x_i)(i=0,1,\cdots,m)$按某种标准最小. 若记 $\delta=(\delta_0,\delta_1,\cdots,\delta_m)^{\mathrm{T}}$，就是要求向量的范数 $\|\delta\|$ 最小. 这就是函数的最佳逼近问题. 在介绍最佳逼近方法之前，我们先介绍具有重要作用的正交多项式.

3.1　正 交 多 项 式

正交多项式是数值计算中的重要工具，这里只介绍正交多项式的基本概念、某些性质和构造方法. 离散情形的正交多项式用于数据拟合，连续情形的正交多项式用于生成最佳平方逼近多项式和下章的高斯型求积公式的构造. 它们在数值分析的其他领域中也有不少应用.

3.1.1　离散点集上的正交多项式

设有点集$\{x_i\}_{i=0}^m$，函数 $f(x)$和 $g(x)$在离散意义上的内积定义为

$$(f,g)=\sum_{i=0}^{m}w_i f(x_i)g(x_i),\tag{3.1}$$

其中 $w_i>0$ 为给定的权数. 在离散意义下，函数 $f(x)$的 2 范数定义为

$$\|f\|_2=\sqrt{(f,f)}.\tag{3.2}$$

有了内积，就可以定义正交性. 若函数 $f(x)$和 $g(x)$的内积$(f,g)=0$，则称两者正交. 若多项式组$\{\varphi_k(x)\}_{k=0}^n$在离散意义上的内积满足

$$(\varphi_i,\varphi_j)=\begin{cases}0, & i\neq j,\\ a_i>0, & i=j,\end{cases}\tag{3.3}$$

则称多项式组$\{\varphi_k(x)\}_{k=0}^n$为在离散点集$\{x_i\}_{i=0}^m$上的带权$\{w_i\}_{i=0}^m$的正交多项式序列.

下面给出离散点集上正交多项式的构造方法.

给定点集$\{x_i\}_{i=0}^m$和权数$\{w_i\}_{i=0}^m$，并且点集$\{x_i\}_{i=0}^m$中至少有 $n+1$ 个互异，则由下列 3 项递推公式

$$\begin{cases} P_0(x)=1,\ P_1(x)=x-a_0, \\ P_{k+1}(x)=(x-a_k)P_k(x)-b_kP_{k-1}(x), \quad k=1,2,\cdots,n-1, \end{cases} \tag{3.4}$$

给出的多项式序列$\{P_k(x)\}_{k=0}^n (n<m)$是正交多项式序列，其中

$$a_k=\frac{(xP_k,P_k)}{(P_k,P_k)}, \quad b_k=\frac{(P_k,P_k)}{(P_{k-1},P_{k-1})}. \tag{3.5}$$

3 项递推公式(3.4)是构造正交多项式的简单公式，此外，还有其他的等价形式，这里，不进一步地讨论.

例 3.1 已知点集$\{x_i\}_{i=0}^4=\{0,0.25,0.5,0.75,1\}$和权数$\{w_i\}_{i=0}^4=\{1,1,1,1,1\}$. 试用 3 项递推公式求关于该点集的正交多项式 $P_0(x),P_1(x),P_2(x)$.

解 先令 $P_0(x)=1$，由此得

$$(P_0,P_0)=\sum_{i=0}^4 w_iP_0^2(x_i)=5,$$

$$(xP_0,P_0)=\sum_{i=0}^4 w_ix_iP_0^2(x_i)=2.5,$$

$$a_0=\frac{(xP_0,P_0)}{(P_0,P_0)}=0.5,$$

$$P_1(x)=x-a_0=x-0.5.$$

进一步地，

$$(P_1,P_1)=\sum_{i=0}^4 w_iP_1^2(x_i)=0.625,$$

$$(xP_1,P_1)=\sum_{i=0}^4 w_ix_iP_1^2(x_i)=0.312\,5.$$

从而有

$$a_1=\frac{(xP_1,P_1)}{(P_1,P_1)}=0.5, \quad b_1=\frac{(P_1,P_1)}{(P_0,P_0)}=0.125,$$

$$P_2(x)=(x-a_1)P_1(x)-b_1P_0(x)=(x-0.5)^2-0.125.$$

3.1.2 连续区间上的正交多项式

连续区间上的正交多项式的概念与离散点集上的正交多项式的概念类似，只要将内积的定义作相应的改变. 函数 $f(x)$和 $g(x)$在连续意义上的内积定义为

$$(f,g)=\int_a^b \rho(x)f(x)g(x)\mathrm{d}x, \quad f,g\in C[a,b], \tag{3.6}$$

其中的 $\rho(x)\geqslant 0$ 为给定的权函数. 按照连续意义上的内积，若多项式组

$\{\varphi_k(x)\}_{k=0}^{n}$满足条件(3.3)式,则称它为在区间$[a,b]$上的带权$\rho(x)$的正交多项式序列.

完全类似于离散情况下的正交多项式的构造方法,连续区间上的正交多项式序列同样可由 3 项递推公式(3.4)和(3.5)构造,但要注意,其中的内积要按(3.6)式计算.

例 3.2　求$[0,1]$上带权$\rho(x)=\ln\dfrac{1}{x}$的前 3 个正交多项式$P_0(x),P_1(x),P_3(x)$.

解　由 3 项递推公式有$P_0(x)=1,P_1(x)=x-a_0$.

$$P_2(x)=(x-a_1)P_1(x)-b_1P_0(x),$$

其中

$$a_0=\frac{(xP_0,P_0)}{(P_0,P_0)},\quad a_1=\frac{(xP_1,P_1)}{(P_1,P_1)},\quad b_1=\frac{(P_1,P_1)}{(P_0,P_0)}.$$

由内积的定义得

$$(P_0,P_0)=-\int_0^1\ln x\mathrm{d}x=1,\quad (xP_0,P_0)=-\int_0^1 x\ln x\mathrm{d}x=\frac{1}{4},$$

即得

$$a_0=\frac{1}{4},\quad P_1(x)=x-\frac{1}{4}.$$

再由

$$(P_1,P_1)=\int_0^1(-\ln x)\left(x-\frac{1}{4}\right)^2\mathrm{d}x=\frac{7}{144},$$

$$(xP_1,P_1)=\int_0^1(-\ln x)x\left(x-\frac{1}{4}\right)^2\mathrm{d}x=\frac{13}{576},$$

得到$a_1=\dfrac{13}{28},b_1=\dfrac{7}{144}$,于是

$$P_2(x)=\left(x-\frac{13}{28}\right)\left(x-\frac{1}{4}\right)-\frac{7}{144}=x^2-\frac{5}{7}x+\frac{17}{252}.$$

除了利用 3 项递推公式构造正交多项式外,还可以利用 Gramma-Smidt 方法由函数组$\{x^n\}$构造正交多项式序列,递推公式为

$$\begin{cases}\varphi_0(x)=1,\\ \varphi_n(x)=x^n-\displaystyle\sum_{i=0}^{n-1}\frac{(x^n,\varphi_i)}{(\varphi_i,\varphi_i)}\varphi_i(x),\quad n=1,2,\cdots.\end{cases}$$

下面给出几种常用的正交多项式.

(1) Legendre 多项式.

Legendre 多项式可由 3 项递推公式

$$\begin{cases} P_0(x)=1,\ P_1(x)=x, \\ (n+1)P_{n+1}(x)=(2n+1)xP_n(x)-nP_{n-1}(x),\ n=1,2,\cdots, \end{cases} \tag{3.7}$$

给出. 它们是在区间$[-1,1]$上带权$\rho(x)=1$的正交多项式. Legendre 多项式也可表示为

$$P_n(x)=\frac{1}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}[(x^2-1)^n],\quad n=1,2,\cdots.$$

它具有正交性质

$$\int_{-1}^{1} P_n(x)P_m(x)\mathrm{d}x=\begin{cases} 0, & n\neq m, \\ \dfrac{2}{2n+1}, & n=m. \end{cases}$$

前几个 Legendre 多项式如下：

$$P_2(x)=\frac{1}{2}(3x^2-1),$$

$$P_3(x)=\frac{1}{2}(5x^3-3x),$$

$$P_4(x)=\frac{1}{8}(35x^4-30x^2+3),$$

$$P_5(x)=\frac{1}{8}(63x^5-70x^3+15x).$$

它们的根都是在开区间$(-1,1)$上的单根，并且与原点对称.

(2) 第一类 Chebyshev 多项式.

第一类 Chebyshev 多项式可由 3 项递推公式

$$\begin{cases} T_0(x)=1,\ T_1(x)=x, \\ T_{n+1}(x)=2xT_n(x)-T_{n-1}(x),\quad n=1,2,\cdots, \end{cases} \tag{3.8}$$

给出. 它们是在区间$[-1,1]$带权$\rho(x)=\dfrac{1}{\sqrt{1-x^2}}$的正交多项式. 第一类 Chebyshev多项式也可表示为

$$T_n(x)=\cos(n\arccos x),\quad n=0,1,\cdots.$$

它具有正交性质

$$\int_{-1}^{1}\frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\mathrm{d}x=\begin{cases} 0, & n\neq m, \\ \dfrac{\pi}{2}, & n=m\neq 0, \\ \pi, & n=m=0. \end{cases}$$

前几个第一类 Chebyshev 多项式如下：

$$T_2(x)=2x^2-1,$$

$$T_3(x)=4x^3-3x,$$
$$T_4(x)=8x^4-8x^2+1,$$
$$T_5(x)=16x^5-20x^3+5x.$$

它们的根都是在开区间$(-1,1)$上的单根,并且与原点对称.

(3) Laguerre 多项式.

Laguerre 多项式可由 3 项递推公式

$$\begin{cases} L_0(x)=1,\ L_1(x)=1-x, \\ L_{n+1}(x)=(1+2n-x)L_n(x)-n^2L_{n-1}(x), \quad n=1,2,\cdots, \end{cases} \tag{3.9}$$

给出. 它们是在区间$[0,+\infty)$上带权 $\rho(x)=\mathrm{e}^{-x}$ 的正交多项式. Laguerre 多项式也可表示为

$$L_n(x)=\mathrm{e}^{x}\,\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^n\mathrm{e}^{-x}),\quad n=1,2,\cdots.$$

它具有正交性质

$$\int_0^{\infty}\mathrm{e}^{-x}L_n(x)L_m(x)\mathrm{d}x=\begin{cases}0, & n\neq m,\\ (n!)^2, & n=m.\end{cases}$$

前几个 Laguerre 多项式如下:

$$L_2(x)=x^2-4x+2,$$
$$L_3(x)=-x^3+9x^2-18x+6,$$
$$L_4(x)=x^4-16x^3+72x^2-96x+24,$$
$$L_5(x)=-x^5+25x^4-200x^3+600x^2-600x+120,$$

它们的根都是在区间$(0,+\infty)$上的单根.

(4) Hermite 多项式.

Hermite 多项式可由 3 项递推公式

$$\begin{cases} H_0(x)=1,\ H_1(x)=2x, \\ H_{n+1}(x)=2xH_n(x)-2nH_{n-1}(x), \quad n=1,2,\cdots, \end{cases} \tag{3.10}$$

给出. 它们是在区间$(-\infty,+\infty)$上带权 $\rho(x)=\mathrm{e}^{-x^2}$ 的正交多项式. Hermite 多项式也可表示为

$$H_n(x)=(-1)^n\mathrm{e}^{x^2}\,\frac{\mathrm{d}^n}{\mathrm{d}x^n}(\mathrm{e}^{-x^2}),\quad n=1,2,\cdots.$$

它具有正交性质

$$\int_{-\infty}^{+\infty}\mathrm{e}^{-x^2}H_n(x)H_m(x)\mathrm{d}x=\begin{cases}0, & n\neq m,\\ 2^n n!\sqrt{\pi}, & n=m.\end{cases}$$

前几个 Hermite 多项式如下:

$$H_2(x)=4x^2-2,$$
$$H_3(x)=8x^3-12x,$$
$$H_4(x)=16x^4-48x^2+12,$$
$$H_5(x)=32x^5-160x^3+120x.$$

它们的根都是在区间$(-\infty,+\infty)$上的单根,并且与原点对称.

3.2 连续函数的最佳逼近

连续函数空间$C[a,b]$上定义了内积(3.6)就形成了一个内积空间. 在$\mathbf{R}^n$空间中任一向量都可用它的一组线性无关的基表示,类似地,对内积空间的任一元素$f(x)\in C[a,b]$,也可用线性无关的基函数近似表示.

设$\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)$在$[a,b]$上连续,如果

$$a_0\varphi_0(x)+a_1\varphi_1(x)+\cdots+a_n\varphi_n(x)=0$$

当且仅当$a_0=a_1=\cdots=a_n=0$时成立,则称$\varphi_0,\varphi_1,\cdots,\varphi_n$在$[a,b]$上是线性无关的. 对于函数组$\{\varphi_k(x)\}_{k=0}^n$的线性无关性,有如下定理.

定理 3.1 $\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)$在$[a,b]$上线性无关的充分必要条件是它的 Gramer 行列式$G_n\neq 0$,其中

$$G_n=\begin{vmatrix}(\varphi_0,\varphi_0) & (\varphi_1,\varphi_0) & \cdots & (\varphi_n,\varphi_0)\\ (\varphi_0,\varphi_1) & (\varphi_1,\varphi_1) & \cdots & (\varphi_n,\varphi_1)\\ \vdots & \vdots & & \vdots\\ (\varphi_0,\varphi_n) & (\varphi_1,\varphi_n) & \cdots & (\varphi_n,\varphi_n)\end{vmatrix}.$$

函数的最佳逼近问题可叙述为:对函数类A中给定的函数$f(x)$,比如$A=C[a,b]$,要求在另一类较简单的便于计算的函数类B中,比如B为多项式函数类,求函数$\varphi(x)\in B\subset A$,使$\varphi(x)$与$f(x)$之差在某种度量意义上最小. 下面对具体的度量进行讨论.

3.2.1 连续函数的最佳平方逼近

我们先讨论在区间$[a,b]$上一般的最佳平方逼近问题. 设$\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)$是$C[a,b]$中的线性无关函数,记

$$\Phi=\operatorname{span}\{\varphi_0,\varphi_1,\cdots,\varphi_n\}=\{\varphi(x):\varphi(x)=\sum_{k=0}^{n}a_k\varphi_k(x),a_k\in\mathbf{R}\}.$$

对于$f(x)\in C[a,b]$,若存在$\varphi^*(x)\in\Phi$,使得

$$\|f-\varphi^*\|_2^2=\inf_{\varphi\in\Phi}\|f-\varphi\|_2^2=\inf_{\varphi\in\Phi}\int_a^b\rho(x)[f(x)-\varphi(x)]^2\mathrm{d}x,\qquad(3.11)$$

则称 $\varphi^*(x)$是 $f(x)$在子集 $\Phi\subset C[a,b]$中的最佳平方逼近函数.

求 $\varphi^*(x)$等价于求多元函数

$$I(a_0,a_1,\cdots,a_n)=\int_a^b\rho(x)\Big[\sum_{k=0}^{n}a_k\varphi_k(x)-f(x)\Big]^2\mathrm{d}x$$

的极小值. 利用多元函数求极值的必要条件有

$$\frac{\partial I}{\partial a_j}=2\int_a^b\rho(x)\Big[\sum_{k=0}^{n}a_k\varphi_k(x)-f(x)\Big]\varphi_j(x)\mathrm{d}x=0,\quad j=0,1,\cdots,n.$$

按内积的定义,上式可写为

$$\sum_{k=0}^{n}a_k(\varphi_k,\varphi_j)=(f,\varphi_j),\quad j=0,1,\cdots,n. \tag{3.12}$$

这是关于 $a_0,a_1,\cdots,a_n$ 的线性方程组,称为法方程.

由于 $\varphi_0,\varphi_1,\cdots,\varphi_n$ 线性无关,故(3.12)式的系数矩阵非奇异,于是(3.12)式有唯一解$a_k=a_k^*$,$k=0,1,\cdots,n$. 从而得到

$$\varphi^*(x)=a_0^*\varphi_0(x)+a_1^*\varphi_1(x)+\cdots+a_n^*\varphi_n(x). \tag{3.13}$$

该式满足(3.11)式,即对任意 $\varphi(x)\in\Phi$,有

$$\|f-\varphi^*\|_2\leqslant\|f-\varphi\|_2. \tag{3.14}$$

事实上,由(3.12)式知

$$\Big(\sum_{k=0}^{n}a_k^*\varphi_k-f,\varphi_j\Big)=0,\quad j=0,1,\cdots,n.$$

因此,对任意 $\varphi(x)\in\Phi$,有$(\varphi^*-f,\varphi)=0$,从而也有$(f-\varphi^*,\varphi^*-\varphi)=0$. 于是

$$\begin{aligned}\|f-\varphi\|_2^2&=\|f-\varphi^*+\varphi^*-\varphi\|_2^2\\&=\|f-\varphi^*\|_2^2+2(f-\varphi^*,\varphi^*-\varphi)+\|\varphi^*-\varphi\|_2^2\\&=\|f-\varphi^*\|_2^2+\|\varphi^*-\varphi\|_2^2\geqslant\|f-\varphi^*\|_2^2.\end{aligned}$$

这就证明了(3.14)式,从而也证明了 f 在 Φ 中最佳平方逼近的存在唯一性.

若令 $\delta(x)=f(x)-\varphi^*(x)$,则称 $\|\delta\|_2$ 为最佳逼近的误差,称

$$\begin{aligned}\|\delta\|_2^2&=(f-\varphi^*,f-\varphi^*)=(f,f)-(\varphi^*,f)\\&=\|f\|_2^2-\sum_{k=0}^{n}a_k^*(\varphi_k,f)\end{aligned} \tag{3.15}$$

为平方误差.

考虑特殊情形,设$[a,b]=[0,1]$,$\varphi_k(x)=x^k$,$k=0,1,\cdots,n$,$\rho(x)=1$. 对于$f\in C[0,1]$,在 $\Phi=\mathrm{span}\{1,x,\cdots,x^n\}$中的最佳平方逼近多项式可以表示为

$$P_n^*(x)=a_0^*+a_1^*+\cdots+a_n^*x^n,$$

相应于法方程(3.12)中的系数矩阵为

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

称之为 Hilbert 矩阵.

例 3.3 设 $f(x)=\sqrt{1+x^2}$,求$[0,1]$上的一次最佳平方逼近多项式.

解 由于 $\varphi_0(x)=1,\varphi_1(x)=x$,

$$(f,\varphi_0)=\int_0^1\sqrt{1+x^2}\mathrm{d}x=\frac{1}{2}\ln(1+\sqrt{2})+\frac{\sqrt{2}}{2}\approx 1.147,$$

$$(f,\varphi_1)=\int_0^1 x\sqrt{1+x^2}\mathrm{d}x=\frac{1}{3}(2\sqrt{2}-1)\approx 0.609,$$

得方程组

$$\begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix}\begin{bmatrix} a_0 \\ a_1 \end{bmatrix}=\begin{bmatrix} 1.147 \\ 0.609 \end{bmatrix},$$

解得 $a_0=0.934,a_1=0.426$. 从而最佳平方逼近为

$$P_1^*(x)=0.934+0.426x.$$

平方误差

$$\|\delta\|_2^2=\int_0^1(1+x^2)\mathrm{d}x-0.934(f,\varphi_0)-0.426(f,\varphi_1)=0.002\,6.$$

由于 Hilbert 矩阵是病态的(见第 5 章),用$\{1,x,\cdots,x^n\}$作基时,求法方程的解,舍入误差很大. 实用的办法是采用正交多项式作基.

若 $\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)$是 $C[a,b]$中的正交多项式组,则由(3.12)式得

$$a_k^*=\frac{(f,\varphi_k)}{(\varphi_k,\varphi_k)},\quad k=0,1,\cdots,n.$$

于是 $f(x)$的最佳平方逼近多项式为

$$\varphi^*(x)=\sum_{k=0}^{n}\frac{(f,\varphi_k)}{(\varphi_k,\varphi_k)}\varphi_k(x).$$

此时,由(3.15)式可得

$$\|\delta\|^2=\|f\|_2^2-\sum_{k=0}^{n}(a_k^*)^2(\varphi_k,\varphi_k)=\|f\|_2^2-\|\varphi^*\|_2^2.$$

例 3.4 设 $f(x)=\mathrm{e}^x$,在$[-1,1]$上用 Legendre 多项式作 f 的 3 次最佳

平方逼近多项式.

解　用 Legendre 多项式 $P_k(x)(k=0,1,2,3)$,可得

$$(f,P_0)=\int_{-1}^{1}\mathrm{e}^x\mathrm{d}x\approx 2.350\,4,$$

$$(f,P_1)=\int_{-1}^{1}x\mathrm{e}^x\mathrm{d}x\approx 0.735\,8,$$

$$(f,P_2)=\int_{-1}^{1}\frac{1}{2}(3x^2-1)\mathrm{e}^x\mathrm{d}x\approx 0.143\,1,$$

$$(f,P_3)=\int_{-1}^{1}\frac{1}{2}(5x^3-3x)\mathrm{e}^x\mathrm{d}x\approx 0.020\,13,$$

$$a_0^*=\frac{(f,P_0)}{(P_0,P_0)}=1.175\,2,\ a_1^*=\frac{(f,P_1)}{(P_1,P_1)}=1.103\,6,$$

$$a_2^*=\frac{(f,P_2)}{(P_2,P_2)}=0.357\,8,\ a_3^*=\frac{(f,P_3)}{(P_3,P_3)}=0.070\,46.$$

于是最佳平方逼近为

$$\varphi^*(x)=0.996\,3+0.997\,9x+0.536\,7x^2+0.176\,1x^3.$$

平方误差

$$\|\delta\|_2^2=\int_{-1}^{1}\mathrm{e}^{2x}\mathrm{d}x-\sum_{k=0}^{3}(a_k^*)^2(\varphi_k,\varphi_k)=0.000\,07.$$

3.2.2　连续函数的最佳一致逼近

类似于最佳平方逼近的思想,对给定的较复杂的函数 $f(x)\in C[a,b]$,我们考虑用多项式

$$p(x)=c_0+c_1x+\cdots+c_nx^n$$

来近似代替它.为了使 $p(x)$是 $f(x)$的最好近似,要求它们之间的另一种距离最小,即要求

$$\|f(x)-p(x)\|_\infty=\max_{a\leqslant x\leqslant b}|f(x)-p(x)|$$

取得最小值,可以看出,这等价于如下 $n+1$ 元函数

$$I(c_0,c_1,\cdots,c_n)=\max_{a\leqslant x\leqslant b}|f(x)-p(x)|=\max_{a\leqslant x\leqslant b}\left|f(x)-\sum_{i=0}^{n}c_ix^i\right|$$

取得最小值.与最佳平方逼近法不同的是,目标函数 $I(c_0,c_1,\cdots,c_n)$一般不是光滑函数.因此就不能用极值的必要条件来推导 $c_0,c_1,\cdots,c_n$ 满足的条件.

如果 $c_0,c_1,\cdots,c_n$ 是优化问题

$$\min_{c_0,c_1,\cdots,c_n}I(c_0,c_1,\cdots,c_n)$$

的解,那么就称多项式 $p(x)$为 $f(x)$在区间$[a,b]$上的最佳一致逼近多项式.把

上述确定 $f(x)$的近似函数的方法叫做最佳一致逼近法.

记

$$R(x)=f(x)-p(x),\quad E=\max_{a\leqslant x\leqslant b}|R(x)|.$$

通过比较复杂的推导可以证明下面反映最佳逼近多项式特征的 Chebyshev 定理.

定理 3.2 n 次多项式 $p(x)$是 $f(x)$在区间$[a,b]$上的最佳一致逼近多项式,当且仅当误差函数 $R(x)$在区间$[a,b]$上以正负交替的符号依次取值 E 的点(称为偏差点)的个数不少于 $n+2$.

如果记区间$[a,b]$上取值为$\pm E$ 的点为 x_j,j 属于某个下标集 J,那么根据上述定理知$|J|\geqslant n+2$,其中,$|J|$表示集合 J 中的元素个数.

实际计算时,求出满足定理 3.2 条件的 $c_0,c_1,\cdots,c_n,x_j,j\in J$ 和 E 是很难的.下面我们介绍一种近似方法,称为 Remes 算法.

第一步:选取近似偏差点(通常取区间的端点为偏差点)$x_j^{(0)}$,$j=1,2,\cdots,n+2$,满足条件

$$a=x_1^{(0)}<x_2^{(0)}<\cdots<x_{n+2}^{(0)}=b.$$

第二步:求解含 $n+2$ 个未知数 $c_0,c_1,\cdots,c_n$ 和 E 的线性方程组

$$\begin{cases}c_0+c_1x_1^{(0)}+c_2(x_1^{(0)})^2+\cdots+c_n(x_1^{(0)})^n+E=f(x_1^{(0)}),\\ c_0+c_1x_2^{(0)}+c_2(x_2^{(0)})^2+\cdots+c_n(x_2^{(0)})^n-E=f(x_2^{(0)}),\\ \cdots\cdots\\ c_2+c_1x_{n+2}^{(0)}+c_2(x_{n+2}^{(0)})^2+\cdots+c_n(x_{n+2}^{(0)})^n-(-1)^{n+2}E=f(x_{n+2}^{(0)}),\end{cases}$$

得初始逼近多项式 $P_n(x)=c_0+c_1x+\cdots+c_nx^n$ 和 E 的值.

第三步:利用某种优化方法计算 $P_n(x)-f(x)$的所有极值点,假设正好有 $n+2$个,分别记之为

$$a\leqslant x_1^{(1)}<x_2^{(1)}<\cdots<x_{n+2}^{(1)}\leqslant b.$$

第四步:将 $x_j^{(1)}$ 分别取代 $x_j^{(0)}$,$j=1,2,\cdots,n+2$,转向第二步.

上述过程进行到相邻两步得到的 c_i,$i=0,1,\cdots,n$ 相差很小时终止.可以证明上述算法所得的 c_i 将收敛到最佳逼近解.

可以看出,Remes 算法十分复杂,比如要求计算 $P_n(x)-f(x)$的所有极值点,这本身是一个很难的全局优化问题.因此,寻找函数的最佳一致逼近多项式是一个计算上很困难的问题.实际中常采用一些近似求法.

例 3.5 求区间$[0,1]$上函数 $y=\arctan x$ 的一次最佳一致逼近多项式.

解 根据 Chebyshev 定理,误差函数

$$R(x)=\arctan x-c_0-c_1x,$$

在区间$[0,1]$上至少有 3 个偏差点，不妨设之为 $x_0=0, x_1, x_2=1$，它们对应的最大(小)值的绝对值为 E. 则得

$$R(0)=-E,\ R(x_1)=E,\ R(1)=-E,\ R'(x_1)=0,$$

或

$$R(0)=E,\ R(x_1)=-E,\ R(1)=E,\ R'(x_1)=0,$$

即

$$\begin{cases} -c_0=-E, \\ \arctan(x_1)-c_0-c_1x_1=E, \\ \dfrac{\pi}{4}-c_0-c_1=-E, \\ \dfrac{1}{1+x_1^2}-c_1=0. \end{cases}$$

或

$$\begin{cases} -c_0=E, \\ \arctan(x_1)-c_0-c_1x_1=-E, \\ \dfrac{\pi}{4}-c_0-c_1=E, \\ \dfrac{1}{1+x_1^2}-c_1=0, \end{cases}$$

解得

$$c_1=\frac{\pi}{4}\approx 0.785\,4,\quad x_1=\sqrt{\frac{1}{c_1}-1}\approx 0.522\,7,$$

$$c_0=\frac{1}{2}(\arctan x_1-c_1x_1)\approx 0.035\,6,$$

即区间$[0,1]$上函数 $y=\arctan x$ 的一次最佳一致逼近多项为

$$p_1(x)=0.035\,6+0.785\,4x.$$

3.3　离散数据的曲线拟合

3.3.1　最小二乘拟合

对于已知的 $m+1$ 对离散数据$\{x_i,y_i\}_{i=0}^m$和权数$\{w_i\}_{i=0}^m$，记

$$a=\min_{0\leqslant i\leqslant m}x_i,\quad b=\max_{0\leqslant i\leqslant m}x_i.$$

在连续函数空间 $C[a,b]$中选定 $n+1$ 个线性无关的基函数$\{\varphi_k(x)\}_{k=0}^n$，并记由

它们生成的子空间为 $\Phi=\operatorname{span}\{\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)\}$. 如果存在 $\varphi^*(x)=\sum_{k=0}^{n}a_k^*\varphi_k(x)\in\Phi$，使得

$$\sum_{i=0}^{m}w_i[y_i-\varphi^*(x_i)]^2=\min_{\varphi(x)\in\Phi}\sum_{i=0}^{m}w_i[y_i-\varphi(x_i)]^2, \tag{3.16}$$

则称 $\varphi^*(x)$为离散数据$\{x_i,y_i\}_{i=0}^m$在子空间 Φ 中带权$\{w_i\}_{i=0}^m$的最小二乘拟合.

函数 $\varphi(x)$在离散点处的值为

$$\varphi(x_i)=\sum_{k=0}^{n}a_k\varphi_k(x_i),\quad i=0,1,\cdots,m.$$

因此，(3.16)右边的和式是参数 $a_0,a_1,\cdots,a_n$ 的函数，记作

$$I(a_0,a_1,\cdots,a_n)=\sum_{i=0}^{m}w_i\Big[y_i-\sum_{k=0}^{n}a_k\varphi_k(x_i)\Big]^2.$$

这样，求极小值问题(3.16)式的解 $\varphi^*(x)$，就是求多元二次函数 $I(a_0,a_1,\cdots,a_n)$ 的极小点$(a_0^*,a_1^*,\cdots,a_n^*)$，使得

$$I(a_0^*,a_1^*,\cdots,a_n^*)=\min_{a_0,a_1,\cdots,a_n\in\mathbf{R}}I(a_0,a_1,\cdots,a_n).$$

由求多元函数极值的必要条件有

$$\frac{\partial I}{\partial a_j}=-2\sum_{i=0}^{m}w_i\Big[y_i-\sum_{k=0}^{n}a_k\varphi_k(x_i)\Big]\varphi_j(x_i)=0,\quad j=0,1,\cdots,n.$$

按内积的定义，上式可写为

$$\sum_{k=0}^{n}a_k(\varphi_k,\varphi_j)=(y,\varphi_j),\quad j=0,1,\cdots,0. \tag{3.17}$$

这个方程称为法方程(或正规方程). 这里，$y(x_i)=y_i$，$i=0,1,\cdots,m$.

由于 $\varphi_0,\varphi_1,\cdots,\varphi_n$ 线性无关，故(3.17)式的系数矩阵非奇异，方程组(3.17)存在唯一的解 $a_k=a_k^*$，$k=0,1,\cdots,n$，从而得

$$\varphi^*(x)=\sum_{k=0}^{n}a_k^*\varphi_k(x)\in\Phi.$$

可以证明，这样得到的 $\varphi^*(x)$，对任何 $\varphi(x)\in\Phi$，都有

$$\sum_{i=0}^{m}w_i[y_i-\varphi^*(x_i)]^2\leqslant\sum_{i=0}^{m}w_i[y_i-\varphi(x_i)]^2,$$

故 $\varphi^*(x)$是所求的最小二乘拟合. 记 $\delta=y-\varphi^*(x)$，显然，平方误差 $\|\delta\|_2^2$ 或均方误差 $\|\delta\|_2$ 越小，拟合的效果越好. 平方误差有与(3.15)式相同形式的表达式.

3.3.2 多项式拟合

前面讨论了子空间 Φ 中的最小二乘拟合. 这是一种线性拟合模型. 在离散

数据$\{x_i,y_i\}_{i=0}^m$的最小二乘拟合中，最简单、最常用的数学模型是多项式

$$\varphi(x)=a_0+a_1x+\cdots+a_nx^n,$$

即在多项式空间$\Phi=\mathrm{span}\{a,x,\cdots,x^n\}$中作曲线拟合，称为多项式拟合. 这是一种特定的线性模型，因此可用上面讨论的方法求解. 子空间Φ的基函数为$\varphi_k(x)=x^k,k=0,1,\cdots,n$.

例 3.6　对某个长度测量n次，得到n个近似值$x_1,x_2,\cdots,x_n$. 试按最小二乘的意义给出该长度的一个估计值.

解　设所求的估计值为a，考虑误差为 0 的拟合函数$\varphi(x)=x-a$，即求a使函数

$$I(a)=\sum_{i=1}^{n}[\varphi(x_i)-0]^2=\sum_{i=1}^{n}(x_i-a)^2$$

取最小值. 由$\dfrac{\mathrm{d}z}{\mathrm{d}a}=0$可得

$$a=\frac{1}{n}\sum_{i=1}^{n}x_i.$$

例 3.7　用多项式拟合表 3－1 中的离散数据.

表 3－1

i	0	1	2	3	4
x_i	0.00	0.25	0.50	0.75	1.00
y_i	0.10	0.35	0.81	1.09	1.96

解　作数据点的图形如图 3－1，从图形中可看出用 2 次多项式拟合比较合适. 这时$n=2$，子空间Φ的基函数为$\varphi_0(x)=1,\varphi_1(x)=x,\varphi_2(x)=x^2$. 数据中没有给出权数，不妨都取为 1，即$w_i=1,i=0,1,\cdots,4$.

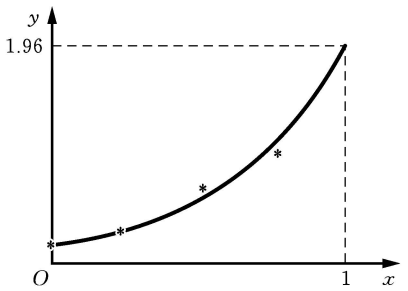

图 3－1

按(3.17)式有

$$\begin{pmatrix} 5 & 2.5 & 1.875 \\ 2.5 & 1.875 & 1.562\,5 \\ 1.875 & 1.562\,5 & 1.382\,8 \end{pmatrix}\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix}=\begin{pmatrix} 4.31 \\ 3.27 \\ 2.797\,5 \end{pmatrix}.$$

解此方程组得 $a_0^*=0.121\,4$，$a_1^*=0.572\,6$，$a_2^*=1.211\,4$. 从而，拟合多项式为

$$\varphi^*(x)=0.121\,4+0.572\,6x+1.211\,4x^2,$$

其平方误差 $\|\delta\|_2^2=0.033\,7$. 拟合曲线 $\varphi^*(x)$的图形见图 3－1.

在许多实际问题中，变量之间的关系不一定能用多项式很好地拟合. 如何找到更符合实际情况的数据拟合，一方面要根据专业知识和经验来确定拟合曲线的形式，另一方面要根据数据点的图形形状及特点来选择适当的曲线拟合这些数据.

例 3.8 已知函数 $y=f(x)$的数据如表 3－2. 试选择适当的数学模型进行拟合.

表 3－2

i	0	1	2	3	4	5	6	7	8	9
x_i	1	2	3	4	6	8	10	12	14	16
y_i	4.00	6.41	8.01	8.79	9.53	9.86	10.33	10.42	10.53	10.61

解 (1) 观察数据点的图形(见图 3－2)，选择二次多项式作为拟合模型. 取所有权数为 1，按(3.17)有

$$\begin{pmatrix} 10 & 76 & 826 \\ 76 & 826 & 103\,96 \\ 826 & 103\,96 & 140\,434 \end{pmatrix}\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix}=\begin{pmatrix} 88.49 \\ 757.59 \\ 8\,530.01 \end{pmatrix}.$$

解得 $a_0^*=4.149\,0$，$a_1^*=1.143\,6$，$a_2^*=-0.048\,32$，从而拟合函数为

$$\varphi^*(x)=4.149\,0+1.143\,6x-0.048\,32x^2,$$

平方误差 $\|\delta\|_2^2=3.948\,6$，$\varphi^*(x)$的图形见图 3－2. 由平方误差和 $\varphi^*(x)$的图形可见，拟合的效果不佳. 因此，不宜直接选用多项式作拟合.

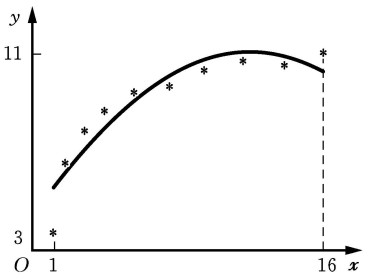

图 3－2

(2) 从数据的图形看，可以选用指数函数进行拟合. 设 $\varphi(x)=\alpha e^{\frac{\beta}{x}}$，其中 $\alpha>0$，$\beta<0$. 这是一个非线性模型，不能直接用上面讨论的方法求解. 对于一般的非线性最小二乘问题，用常规方法求解的难度较大. 这里的非线性模型比较简单，可以把它转化成线性模型，然后用上面讨论的方法求解.

对函数 $\varphi(x)=\alpha e^{\frac{\beta}{x}}$ 的两边取自然对数，得 $\ln\varphi(x)=\ln\alpha+\dfrac{\beta}{x}$. 若令 $t=\dfrac{1}{x}$，$z=\ln\varphi(x)$，$A=\ln\alpha$，则有 $z=A+\beta t$. 这是一个线性模型. 将本题离散数据作相应的转换，见表 3－3.

表 3－3

i	0	1	2	3	4	5	6	7	8	9
t_i	1.000 0	0.500 0	0.333 3	0.250 0	0.166 7	0.1250	0.1000	0.083 3	0.071 4	0.062 5
z_i	1.386 3	1.857 9	2.080 7	2.173 6	2.254 4	2.288 5	2.3 351	2.343 7	2.354 2	2.361 8

对表 3－3 中的数据，作线性拟合，这时 $n=1$，子空间 Φ 的基函数为 $\varphi_0(x)=1$，$\varphi_1(t)=t$. 易得法方程

$$\begin{bmatrix} 10 & 2.692\,3 \\ 2.692\,3 & 1.493\,0 \end{bmatrix}\begin{bmatrix} A \\ \beta \end{bmatrix}=\begin{bmatrix} 21.436\,2 \\ 4.958\,6 \end{bmatrix}.$$

解得 $A=2.428\,4$，$\beta=-1.057\,9$，从而 $\alpha=e^A=11.341\,1$. 于是，所求的拟合函数为

$$\varphi^*(x)=11.341\,1e^{-\frac{1.057\,9}{x}},$$

平方误差为 $\|\delta\|_2^2=0.110\,9$. 它比方法(1)的 $\|\delta\|_2^2=3.948\,6$ 小得多，拟合效果较好.

3.3.3　正交多项式拟合

一般地，用最小二乘法得到的方程组(3.17)，其系数矩阵是病态的. 实用的曲线拟合办法是采用正交函数作 Φ 的基.

若点集 $\{x_i\}_{i=0}^m$ 中至少有 $n+1$ 个互异，那么可用 3 项递推公式(3.4)和(3.5)求出正交多项式序列 $\{\varphi_k(x)\}_{k=0}^n$，它们可以作为子空间 $\Phi=\mathrm{span}\{1,x,\cdots,x^n\}$ 的一组基. 求出多项式序列 $\{\varphi_k(x)\}_{k=0}^n$ 后，可以建立拟合模型

$$\varphi(x)=\sum_{k=0}^{n}a_k\varphi_k(x).$$

此时，对应的法方程为

$$(\varphi_k,\varphi_k)a_k=(y,\varphi_k),\quad k=0,1,\cdots,n.$$

它的解为

$$a_k=\frac{(y,\varphi_k)}{(\varphi_k,\varphi_k)},\quad k=0,1,\cdots,n.$$

由于按法方程(3.17)有

$$(y,\varphi_j)=\sum_{k=0}^{n}a_k(\varphi_k,\varphi_j)=(\varphi,\varphi_j),$$

即$(y-\varphi,\varphi_j)=0,j=0,1,\cdots,n.$因而平方误差为

$$\begin{aligned}\|y-\varphi\|_2^2&=(y-\varphi,y-\varphi)=(y-\varphi,y)\\&=\|y\|_2^2-\sum_{k=0}^{n}a_k(\varphi_k,y)\\&=\|y\|_2^2-\sum_{k=0}^{n}a_k^2(\varphi_k,\varphi_k)=\|y\|_2^2-\|\varphi\|_2^2.\end{aligned}$$

按上述求离散数据$\{x_i,y_i\}_{i=0}^{m}$的拟合多项式$\varphi(x)$的方法,称为正交多项式拟合.根据唯一性,所得结果与用前面的方法所得的结果相同,但数值计算比前者稳定.

例 3.9 用正交化方法求例 3.7 中的离散数据的 2 次多项式拟合.

解 已知离散数据为

$$\{x_i\}_{i=0}^{4}=\{0,0.25,0.5,0.75,1\},$$
$$\{y_i\}_{i=0}^{4}=\{0.1,0.35,0.81,1.09,1.96\}.$$

对权数$\{w_i\}_{i=0}^{4}=\{1,1,1,1,1\}$,在例 3.1 中已求出了点集$\{x_i\}_{i=0}^{4}$上的正交多项式

$$\varphi_0(x)=1,\quad \varphi_1(x)=x-0.5,\quad \varphi_2(x)=(x-0.5)^2-0.125,$$

并且有

$$(\varphi_0,\varphi_0)=5,\quad (\varphi_1,\varphi_1)=0.625,\quad (\varphi_2,\varphi_2)=0.054\,687\,5.$$

进而有

$$\begin{aligned}&(y,\varphi_0)=4.31,\quad (y,\varphi_1)=1.115,\quad (y,\varphi_2)=0.066\,25,\\&a_0=0.862,\qquad a_1=1.784,\qquad a_2=1.211\,428\,571.\end{aligned}$$

最后得拟合多项式

$$\begin{aligned}\varphi(x)&=a_0\varphi_0(x)+a_1\varphi_1(x)+a_2\varphi_2(x)\\&=0.862+1.784(x-0.5)+1.211\,4[(x-0.5)^2-0.125]\\&=0.121\,4+0.572\,6x+1.211\,4x^2.\end{aligned}$$

所得结果与例 3.7 相同.

最后,我们说明可以利用最小二乘法的思想求方程组的近似解.对线性方程组

$$\sum_{j=1}^{n}a_{ij}x_j=b_i,\quad i=1,2,\cdots,m,$$

视 $\varphi_j=a_j$，拟合函数为

$$\varphi(x_1,x_2,\cdots,x_n)=\sum_{j=1}^{n}a_jx_j\ ,$$

则方程组的意义是当$(a_1,a_2,\cdots,a_n)$取值为$(a_{i1},a_{i2},\cdots,a_{in})$时，$\varphi$ 取值为 $b_i(i=1,2,\cdots,m)$. 于是，对应于(3.16)式的最小二乘法的目标函数是

$$\sum_{i=1}^{m}\Big(\sum_{j=1}^{n}a_{ij}x_j-b_i\Big)^2,$$

对应于(3.17)式的法方程是

$$A^TAx=A^Tb,$$

这里，$A=(a_{ij})_{m\times n}$，$x=(x_1,x_2,\cdots,x_n)^T$，$b=(b_1,b_2,\cdots,b_m)^T$. 可见，只要矩阵 A 是列满秩的，法方程就有唯一解.

例 **3.10**　对矛盾方程组

$$\begin{cases}x_1-x_2=1,\\-x_1+x_2=2,\\2x_1-2x_2=3,\\-3x_1+x_2=4,\end{cases}$$

求最小二乘解.

解　系数矩阵和右端向量分别为

$$A=\begin{pmatrix}1&-1\\-1&1\\2&-2\\-3&1\end{pmatrix},\quad b=\begin{pmatrix}1\\2\\3\\4\end{pmatrix}.$$

由此可得

$$A^{\mathrm{T}}A=\begin{bmatrix}15&-9\\-9&7\end{bmatrix},\quad A^{\mathrm{T}}b=\begin{bmatrix}-7\\-1\end{bmatrix},$$

即法方程为

$$\begin{bmatrix}15&-9\\-9&7\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}=\begin{bmatrix}-7\\-1\end{bmatrix}.$$

解此方程组得最小二乘解 $x_1=-\dfrac{29}{12}$，$x_2=-\dfrac{39}{12}$.

评　　注

本章分别就离散点集和连续区间讨论了正交多项式. 以误差平方的极小化

为准则，介绍了最佳平方逼近、最佳一致逼近和曲线拟合方法. 这些方法与插值法的不同之处是不需要知道被逼近函数在节点处的准确值，侧重于反映原函数整体的变化趋势，消除局部波动的影响.

正交多项式在数值分析中有广泛的应用，有离散型和连续型两种，它们的构造方法和基本性质类同. 正交多项式在 Gauss 积分方法中具有重要作用. 最小二乘法在应用科学中有重要应用. 最佳平方逼近和曲线拟合分别要求误差平方的积分和误差平方之和最小，因此个别点误差可能较大。它们的构造都要求解正规方程组，其正规方程组的构造方式相同，都需要计算内积，当正规方程组阶数较高时往往病态，所以最好选取正交多项式系作基函数，可以避免解方程组. 函数逼近中的另一类方法是最佳一致逼近，它要求最大误差最小. 由于难以求出其准确解，一般是求近似的最佳一致逼近. 本书仅对最佳一致逼近法作了简单介绍. 有兴趣的读者可参阅有关数值逼近的文献.

习 题 3

3.1 已知点集 $\{x_i\}_{i=0}^4=\{-2,-1,0,1,2\}$，权数 $\{w_i\}_{i=0}^4=\{0.5,1,1,1,1.5\}$. 试用 3 项递推公式构造对应的正交多项式 $\varphi_0(x),\varphi_1(x),\varphi_2(x)$.

3.2 设 $f(x)=|x|,x\in[-1,1]$. 求在 $\Phi=\mathrm{span}\{1,x^2,x^4\}$ 上的最佳平方逼近.

3.3 求参数 α 和 β，使积分值 $\int_0^{\frac{\pi}{2}}(\sin x-\alpha-\beta x)^2\,\mathrm{d}x$ 最小.

3.4 用 Legendre 多项式求 $f(x)=\sqrt{x}$ 在区间 $[0,1]$ 上的 1 次最佳平方逼近多项式.

3.5 用 Chebyshev 多项式求 e^x 在 $[-1,1]$ 上的 1 次和 3 次最佳平方逼近多项式.

3.6 求函数 $f(x)=\sqrt{1+x^2}$ 在 $[0,1]$ 上的 1 次最一致逼近多项式.

3.7 观察物体的直线运动，得出数据如表 3-4，求运动方程 $S=at+b$.

表 3-4

时间 t/s	0.0	0.9	1.9	3.0	3.9	5.0
距离 S/m	0	10	30	50	80	110

3.8 已知离散数据如表 3-5，用非线性模型 $\varphi(x)=ax+\beta\mathrm{e}^{-x}$ 作最小二乘拟合，并求平方误差 $\|\delta\|_2^2$.

表 3－5

x_i	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
y_i	4.000	2.927	2.470	2.393	2.540	2.829	3.198	3.621	4.072

3.9　对于例 3.8 给出的离散数据$\{x_i,y_i\}_{i=0}^{9}$，用非线性模型 $\varphi(x)=\dfrac{x}{\alpha x+\beta}$ 作曲线拟合，把它转化成线性模型，求 α,β 和平方误差 $\|\delta\|_2^2$.

数值试验题 3

3.1　对表 3－6 的数据作 3 次多项式拟合，取权数 $w_i=1$，给出拟合多项式的系数、平方误差，并作离散数据$\{x_i,y_i\}$和拟合多项式的图形.

表 3－6

x_i	－1.0	－0.5	0.0	0.5	1.0	1.5	2.0
y_i	－4.447	－0.452	0.551	0.048	－0.447	0.549	4.552

3.2　对上题给定的数据，用 3 次正交多项式求解最小二乘拟合问题.

3.3　考虑数值试验题 2.2 中的函数和节点，作函数数据的 2 次和 3 次拟合多项式，将拟合的结果与 Lagrange 插值及样条插值的结果比较.

3.4　用形如 $ae^x+b\sin x+c\ln x+d\cos x$ 的函数在最小二乘的意义下拟合表 3－7 数据.

表 3－7

x_i	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
y_i	1.284	1.648	2.117	2.718	3.427	2.798	3.534	4.456	5.465	5.894

第 4 章　数值积分和数值微分

积分与微分的计算，是具有广泛应用的古典问题. 然而，在微积分教材中，只对简单的或特殊的情况，提供了函数的积分或微分的解析表达式. 比如，对于在区间$[a,b]$上函数 $f(x)$的积分，只要能找到被积函数 $f(x)$的原函数 $F(x)$，在理论上可以使用 Newton-Leibniz 公式

$$\int_a^b f(x)\mathrm{d}x = F(b) - F(a)$$

计算. 但对很多实际问题，这种方法已无能为力，常常遇到的主要问题有：

(1) 找不到被积函数 $f(x)$的原函数 $F(x)$，如

$$f(x)=\frac{1}{\ln x},\ f(x)=\frac{\sin x}{x},\ f(x)=\mathrm{e}^{-x^2},$$

$$f(x)=\sqrt{1+\cos^2 x},\ f(x)=\frac{1}{1-k^2\sin^2 x}.$$

(2) 被积函数没有有限的解析表达式，而是由测量数据或数值计算给出的数据表示.

例如，一块铝合金薄板的横断面为正弦波，要求原材料铝合金板的长度，也就是 $f(x)=\sin x$ 从 $x=0$ 到 $x=b$ 的曲线弧长 L，可用积分表示为

$$L = \int_0^b \sqrt{1+(f'(x))^2}\mathrm{d}x = \int_0^b \sqrt{1+\cos^2 x}\mathrm{d}x,$$

这是一个椭圆积分计算问题.

因此，有必要研究积分的数值计算问题. 对函数的微分也一样，以表格形式给出的函数，要求出其导数时，还是要依靠数值微分的方法. 例如，已知一组实测数值 $y_i=y(x_i)$，$i=0,1,\cdots,n$，其数学模型是一个二阶常微分方程

$$xy''+ay'+(x-b)y=0,$$

需要确定模型中的待定参数 a 和 b. 如果我们能由实测数值得到 $y'(x_i)$和$y''(x_i)$的数值，代入模型中就可用最小二乘法确定 a 和 b. 这是一个计算数值微分值的问题.

所谓数值积分方法，就是对定积分

$$I[f] = \int_a^b f(x)\mathrm{d}x,$$

用被积函数 $f(x)$在区间$[a,b]$上的一些节点 x_k 处的函数值 $f(x_k)$的线性组合

$$I_n[f]=\sum_{k=0}^{n}A_kf(x_k)$$

来近似表示. 称 x_k 为求积节点，A_k 为相应的求积系数，$I_n[f]$为数值求积公式.

本章讨论常用的数值求积公式及它们的误差估计和代数精度，而对数值微分只作简单介绍.

4.1　Newton-Cotes 求积公式

4.1.1　插值型求积法

在积分区间$[a,b]$上给定 $n+1$ 个节点 $a\leqslant x_0<x_1<x\cdots<x_n\leqslant b$ 和相应的函数值$f(x_0),f(x_1),\cdots,f(x_n)$. 由此可以构造出 $f(x)$的 n 次 Lagrange 插值多项式

$$L_n(x)=\sum_{k=0}^{n}f(x_k)l_k(x).$$

从而有

$$I[f]=\int_a^b f(x)\mathrm{d}x\approx\sum_{k=0}^{n}A_kf(x_k),\tag{4.1}$$

其中求积系数

$$A_k=\int_a^b l_k(x)\mathrm{d}x,\quad k=0,1,\cdots,n.\tag{4.2}$$

称由(4.2)式给出求积系数的(4.1)式为插值型求积公式.

利用 Lagrange 插值多项式的余项可知插值型求积公式的余项为

$$R_n[f]=I-I_n=\frac{1}{(n+1)!}\int_a^b f^{(n+1)}(\xi)\omega_{n+1}(x)\mathrm{d}x,$$

其中 ξ 与变量 x 有关. 由此可知，对于次数小于或等于 n 的多项式 $f(x)$，其余项 $R_n[f]=0$.

如果 $f\in C^{n+1}[a,b]$，$\omega_{n+1}(x)$在$[a,b]$上不变号，则由第二积分中值定理知存在 $\eta\in[a,b]$，使得

$$R_n[f]=\frac{1}{(n+1)!}f^{(n+1)}(\eta)\int_a^b\omega_{n+1}(x)\mathrm{d}x.$$

例 4.1　给定求积节点 $x_0=\frac{1}{4}$，$x_1=\frac{3}{4}$，试推出计算积分 $\int_0^1 f(x)\mathrm{d}x$ 插值型求积公式，并写出它的余项.

解　因要求所构造的求积公式是插值型的，故其求积系数可表示为

$$A_0 = \int_0^1 l_0(x)\mathrm{d}x = \int_0^1 \frac{1}{2}(3-4x)\mathrm{d}x = \frac{1}{2},$$
$$A_1 = \int_0^1 l_1(x)\mathrm{d}x = \int_0^1 \frac{1}{2}(4x-1)\mathrm{d}x = \frac{1}{2}.$$

故求积公式为

$$\int_0^1 f(x)\mathrm{d}x \approx \frac{1}{2}\left(f\left(\frac{1}{4}\right)+f\left(\frac{3}{4}\right)\right).$$

若 $f''(x)$在$[0,1]$上存在,则该求积公式余项为

$$R[f] = \frac{1}{2}\int_0^1 f''(\xi)\left(x-\frac{1}{4}\right)\left(x-\frac{3}{4}\right)\mathrm{d}x\ ,$$

其中 $\xi\in(0,1)$并依赖于 x.

定义 4.1 如果定积分 $I[f]$的某个求积公式 $I_n[f]$对于一切次数不高于 m 次的代数多项式 $P_m(x)$准确成立,即 $I[P_m]=I_n[P_m]$,则称公式 $I_n[f]$至少具有 m 次代数精度.如果还有某个 $m+1$ 次多项式 $P_{m+1}(x)$使求积公式不准确成立,即 $I[P_{m+1}]\neq I_n[P_{m+1}]$,则称公式 $I_n[f]$恰有 m 次代数精度.

显然,插值型求积公式(4.1)至少具有 n 次代数精度.反之,如果一个形如(4.1)式的求积公式具有 n 次代数精度,那么它必是插值型的.事实上,由于此时该公式对插值基函数 $l_k(x)$是准确成立的,即

$$\int_a^b l_k(x)\mathrm{d}x = \sum_{j=0}^{n} A_j l_k(x_j), \quad k=0,1,\cdots,n,$$

由插值基函数的性质即得(4.2)式.

下面我们讨论便于使用的插值型求积公式.

4.1.2 Newton-Cotes 求积公式

将积分区间$[a,b]$划分为 n 等分,步长 $h=\dfrac{b-a}{n}$,节点 $x_k=a+kh$,$k=0,1,\cdots,n$.插值型求积公式(4.1)可以写成

$$I[f] \approx I_n[f] = (b-a)\sum_{k=0}^{n} C_k^{(n)} f(x_k), \tag{4.3}$$

其中

$$C_k^{(n)} = \frac{1}{b-a}\int_a^b l_k(x)\mathrm{d}x, \quad k=0,1,\cdots,n. \tag{4.4}$$

(4.3)式称为 n 阶 Newton-Cotes 求积公式,$C_k^{(n)}$ 称为 Cotes 系数.

利用节点的等分性,可以把 Cotes 系数的表达式化简.作变换 $x=a+th$,则有

$$C_k^{(n)} = \frac{h}{b-a}\int_0^n \prod_{\substack{j=0\\ j\neq k}}^{n} \frac{t-j}{k-j}\mathrm{d}t = \frac{(-1)^{n-k}}{k!(n-k)!}\frac{1}{n}\int_0^n \prod_{\substack{j=0\\ j\neq k}}^{n}(t-j)\mathrm{d}t. \tag{4.5}$$

可见，系数 $C_k^{(n)}$ 不但与被积函数无关，而且与积分区间也无关，并且，由(4.5)式可知，$C_k^{(n)}=C_{n-k}^{(n)}$，$k=0,1,\cdots,n$. 利用(4.5)式求出的部分 Cotes 系数见表 4－1.

表 4－1

n									
1	$\frac{1}{2}$	$\frac{1}{2}$							
2	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$						
3	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$					
4	$\frac{7}{90}$	$\frac{32}{90}$	$\frac{12}{90}$	$\frac{32}{90}$	$\frac{7}{90}$				
5	$\frac{19}{288}$	$\frac{75}{288}$	$\frac{50}{288}$	$\frac{50}{288}$	$\frac{75}{288}$	$\frac{19}{288}$			
6	$\frac{41}{840}$	$\frac{216}{840}$	$\frac{27}{840}$	$\frac{272}{840}$	$\frac{27}{840}$	$\frac{216}{840}$	$\frac{41}{840}$		
7	$\frac{751}{17\,280}$	$\frac{3\,577}{17\,280}$	$\frac{1\,323}{17\,280}$	$\frac{2\,989}{17\,280}$	$\frac{2\,989}{17\,280}$	$\frac{1\,323}{17\,280}$	$\frac{3\,577}{17\,280}$	$\frac{751}{17\,280}$	
8	$\frac{989}{28\,350}$	$\frac{5\,888}{28\,350}$	$-\frac{928}{28\,350}$	$\frac{10\,496}{28\,350}$	$-\frac{4\,540}{28\,350}$	$\frac{10\,496}{28\,350}$	$-\frac{928}{28\,350}$	$\frac{5\,888}{28\,350}$	$\frac{989}{28\,350}$

当 $n=1$ 时，Cotes 系数为

$$C_0^{(1)}=C_1^{(1)}=\frac{1}{2},$$

求积公式化为

$$\int_a^b f(x)\mathrm{d}x\approx\frac{b-a}{2}(f(a)+f(b)),\tag{4.6}$$

此公式称为梯形公式.

当 $n=2$ 时，Cotes 系数为

$$C_0^{(2)}=\frac{1}{6},C_1^{(2)}=\frac{4}{6},C_2^{(2)}=\frac{1}{6},$$

求积公式化为

$$\int_a^b f(x)\mathrm{d}x\approx\frac{b-a}{6}\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right),\tag{4.7}$$

此公式称为 Simpson 公式，也称抛物线求积分式.

同样，$n=3$ 时，由表 4－1 可写出公式

$$\int_a^b f(x)\mathrm{d}x\approx\frac{b-a}{8}(f(x_0)+3f(x_1)+3f(x_2)+f(x_3)),\tag{4.8}$$

称为 Newton 公式.

当 $n=4$ 时，由表 4－1 可写出公式

$$\int_a^b f(x)\mathrm{d}x \approx \frac{b-a}{90}(7f(x_0)+32f(x_1)+12f(x_2)+32f(x_3)+7f(x_4)), \tag{4.9}$$

称为 Cotes 公式.

例 **4.2** 用 Newton-Cotes 公式计算积分 $\int_0^{\frac{\pi}{4}} \sin x\mathrm{d}x = 1-\frac{\sqrt{2}}{2}$ 的近似值.

解 利用(4.3)式，计算结果列于表 4－2，其中误差采用积分精确值减去用 Newton-Cotes 公式的计算值.

表 **4－2**

n	1	2	3
计算值	0.277 680 18	0.292 932 64	0.292 910 70
误　差	0.015 213 03	0.000 039 42	0.000 017 48

4.1.3 Newton-Cotes 公式的误差分析

定理 4.1 设 $f(x)\in C^2[a,b]$，则对梯形公式(4.6)有

$$R_1[f]=I[f]-I_1[f]=-\frac{(b-a)^3}{12}f''(\eta),\quad \eta\in[a,b]. \tag{4.10}$$

证 设 $L_1(x)$ 是 $f(x)$ 以 $x_0=a, x_1=b$ 为节点的一次插值多项式，那么有

$$R_1(x)=f(x)-L_1(x)=\frac{1}{2}f''(\xi)\omega_2(x),\quad \xi\in[a,b].$$

两边积分得梯形公式的误差

$$R_1[f]=\frac{1}{2}\int_a^b f''(\xi)\omega_2(x)\mathrm{d}x.$$

由于 $\omega_2(x)=(x-a)(x-b)$ 在 $[a,b]$ 上不变号，$f(x)\in C^2[a,b]$，故 $f''(\xi)$ 在 $[a,b]$ 上是连续的. 由积分中值定理得

$$R_1[f]=\frac{1}{2}f''(\eta)\int_a^b \omega_2(x)\mathrm{d}x=\frac{1}{2}f''(\eta)\left(-\frac{1}{6}(b-a)^3\right),\quad \eta\in[a,b].$$

由此即得(4.10)式.

定理 4.2 设 $f(x)\in C^4[a,b]$，则对 Simpson 公式(4.7)有

$$R_2[f]=I[f]-I_2[f]=-\frac{1}{90}\left(\frac{b-a}{2}\right)^5 f^{(4)}(\eta),\quad \eta\in[a,b] \tag{4.11}$$

证 构造 3 次插值多项式 $H(x)$，使满足

$$H(a)=f(a),\ H\left(\frac{a+b}{2}\right)=f\left(\frac{a+b}{2}\right),$$

$$H(b)=f(b),\ H'\left(\frac{a+b}{2}\right)=f'\left(\frac{a+b}{2}\right).$$

由于 Simpson 公式(4.7)的代数精度是 3,它对于 3 次多项式 $H(x)$是准确的,即有

$$\begin{aligned}\int_a^b H(x)\mathrm{d}x &= \frac{b-a}{6}\left(H(a)+4H\left(\frac{a+b}{2}\right)+H(b)\right)\\ &=\frac{b-a}{6}\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right).\end{aligned}$$

因此,(4.7)式的积分余项为

$$R_2[f]=\int_a^b f(x)\mathrm{d}x-\int_a^b H(x)\mathrm{d}x=\int_a^b (f(x)-H(x))\mathrm{d}x.$$

对于插值多项式 $H(x)$,根据其满足的插值条件,不难证明

$$f(x)-H(x)=\frac{1}{4!}f^{(4)}(\xi)(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b),\quad \xi\in[a,b].$$

由于 $f^{(4)}(\xi)$在$[a,b]$上连续,$(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)$在$(a,b)$内不变号,应用积分中值定理有

$$\begin{aligned}R_2[f] &= \frac{1}{4!}f^{(4)}(\eta)\int_a^b (x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)\mathrm{d}x\\ &=-\frac{1}{90}\left(\frac{b-a}{2}\right)^5 f^{(4)}(\eta),\quad \eta\in[a,b].\end{aligned}$$

定理得证.

由(4.10)式可见,梯形公式的代数精度为 1.而(4.11)式表明,Simpson 公式的代数精度是 3.更一般地,我们有下述论断.

定理 4.3　当 n 为偶数时,n 阶 Newton-Cotes 公式(4.3)至少有 $n+1$ 次代数精度.

证　我们只要验证,当 n 为偶数时,Newton-Cotes 对 $f(x)=x^{n+1}$的余项为 0.此时,由于 $f^{(n+1)}(x)=(n+1)!$,我们有

$$R_n[f]=\int_a^b \omega_{n+1}(x)\mathrm{d}x=h^{n+2}\int_0^n\prod_{j=0}^n(t-j)\mathrm{d}t,$$

这里,$x=a+th$.再令 $t=u+\frac{n}{2}$,进一步有

$$R_n[f]=h^{n+2}\int_{-\frac{n}{2}}^{\frac{n}{2}}\prod_{j=0}^n\left(u+\frac{n}{2}-j\right)\mathrm{d}u.$$

显然,被积函数

$$\prod_{j=0}^n\left(u+\frac{n}{2}-j\right)=\prod_{j=-\frac{n}{2}}^{\frac{n}{2}}(u-j)$$

是个奇函数，因此 $R_n[f]=0$.

由定理可见，偶数阶的 Newton-Cotes 公式具有较高次的代数精度. 对 $n=4$，关于求积公式(4.9)，可以证明下面的结论.

定理 4.4 设 $f(x)\in C^6[a,b]$，则对 Cotes 公式(4.9)有

$$R_4[f]=I[f]-I_4[f]=-\frac{8}{945}\left(\frac{b-a}{4}\right)^7 f^{(6)}(\eta),\quad \eta\in[a,b] \tag{4.12}$$

下面我们讨论 Newton-Cotes 公式的计算稳定性问题.

在 Newton-Cotes 公式中，取 $f(x)=1$，此时 $R_n[f]=0$. 并且有

$$\sum_{k=0}^{n} C_k^{(n)} = 1. \tag{4.13}$$

一般地，假定初始数据 $f(x_k)$ 有舍入误差，设 $f(x_k)\approx f^*(x_k)$，$k=0,1,\cdots,n$，反映在计算中有

$$\sum_{k=0}^{n} C_k^{(n)} f(x_k) \approx \sum_{k=0}^{n} C_k^{(n)} f^*(x_k).$$

若记 $\delta = \max\limits_{0\leqslant k\leqslant n} | f(x_k)-f^*(x_k) |$，则有

$$\left| \sum_{k=0}^{n} C_k^{(n)} f(x_k) - \sum_{k=0}^{n} C_k^{(n)} f^*(x_k) \right| \leqslant \delta \sum_{k=0}^{n} | C_k^{(n)} | \tag{4.14}$$

当 $C_k^{(n)} > 0 (k=0,1,\cdots,n)$ 时，由(4.13)式和(4.14)式知计算是稳定的.

由表 4-1 知，当 $n\geqslant 8$ 时，Cotes 系数出现负值，那么

$$\sum_{k=0}^{n} | C_k^{(n)} | > \sum_{k=0}^{n} C_k^{(n)} = 1.$$

特别地，假定 $C_k^{(n)}(f(x_k)-f^*(x_k))>0$，并且 $| f(x_k)-f^*(x_k) |=\alpha$，那么有

$$\begin{aligned}
&\left| \sum_{k=0}^{n} C_k^{(n)} f(x_k) - \sum_{k=0}^{n} C_k^{(n)} f^*(x_k) \right| \\
&= \sum_{k=0}^{n} C_k^{(n)} (f(x_k)-f^*(x_k)) \\
&= \sum_{k=0}^{n} | C_k^{(n)} | | f(x_k)-f^*(x_k) | > \alpha.
\end{aligned}$$

此时，初始数据的误差引起计算结果的误差增大，即计算不稳定.

4.2 复化求积公式

对于定积分 $\int_1^{10} \frac{1}{x}\mathrm{d}x$，其精确值 $I = 2.302\,585$. 用梯形公式(4.6)计算有 $I_1 = 4.95$；用 Simpson 公式(4.7)计算有 $I_2 = 2.740\,909$. 可以看出，它们的误差

很大. 由上一节的讨论可知，高阶 Newton-Cotes 求积公式是不稳定的. 因此，通常不用高阶求积公式得到比较精确的积分值，而是将整个积分区间分段，在每一小段上用低阶求积公式. 这种方法称为复化求积方法. 本节讨论复化梯形公式和复化 Simpson 公式.

4.2.1　复化梯形求积公式

将积分区间$[a,b]$分为 n 等分，$h=\dfrac{b-a}{n}$，$x_k=a+kh$，$k=0,1,\cdots,n$. 在每个子区间$[x_k,x_{k+1}](k=0,1,\cdots,n-1)$上，用梯形公式，则有

$$I[f]=\int_a^b f(x)\mathrm{d}x=\sum_{k=0}^{n-1}\int_{x_k}^{x_{k+1}} f(x)\mathrm{d}x\approx\frac{h}{2}\sum_{k=0}^{n-1}(f(x_k)+f(x_{k+1}))$$

由此令

$$\begin{aligned}T_n[f]&=\frac{h}{2}\sum_{k=0}^{n-1}(f(x_k)+f(x_{k+1}))\\&=\frac{h}{2}\Big(f(a)+2\sum_{k=1}^{n-1}f(x_k)+f(b)\Big),\end{aligned}\tag{4.15}$$

称 $T_n[f]$ 为复化梯形公式.

设 $f(x)\in C^2[a,b]$，由梯形公式的误差有

$$R_{T_n}=I-T_n=\sum_{k=0}^{n-1}\Big(-\frac{h^3}{12}f''(\eta_k)\Big),\quad \eta_k\in[x_k,x_{k+1}].$$

因为

$$\min_{0\leqslant k\leqslant n-1}f''(\eta_k)\leqslant\frac{1}{n}\sum_{k=0}^{n-1}f''(\eta_k)\leqslant\max_{0\leqslant k\leqslant n-1}f''(\eta_k),$$

所以，由连续函数的介值定理知，存在 $\eta\in[\eta_0,\eta_{n-1}]\subset[a,b]$，使得

$$f''(\eta)=\frac{1}{n}\sum_{k=0}^{n-1}f''(\eta_k).$$

于是，复化梯形公式的余项为

$$R_{T_n}=-\frac{b-a}{12}h^2f''(\eta),\quad \eta\in(a,b).\tag{4.16}$$

可以看出，误差(4.16)式是 h^2 阶的. 当 $f(x)\in C^2[a,b]$ 时，$\lim\limits_{n\to\infty}R_{T_n}=0$，即复化梯形公式收敛到$\int_a^b f(x)\mathrm{d}x$.

值得指出的是，收敛性的结论，只要 $f(x)$ 在$[a,b]$上可积即可成立. 事实上，由定积分的定义可知，对$[a,b]$的任一划分 Δ 所作 Riemann 和的极限

$$\lim_{\max\Delta x_i\to0}\sum_{i=1}^{n}f(\xi_i)\Delta x_i=\int_a^b f(x)\mathrm{d}x$$

存在. 该积分对于等距划分和特殊的 ξ_i 当然成立，于是对复化梯形公式有

$$\lim_{n\to\infty} T_n = \frac{1}{2}\left(\lim_{n\to\infty}\sum_{k=0}^{n-1} f(a+kh)h + \lim_{n\to\infty}\sum_{k=0}^{n-1} f(a+(k+1)h)h\right)$$
$$= \frac{1}{2}\left(\int_a^b f(x)\mathrm{d}x + \int_a^b f(x)\mathrm{d}x\right) = \int_a^b f(x)\mathrm{d}x.$$

定义 4.2 如果一种求积公式 I_n 有

$$\lim_{h\to 0}\frac{I-I_n}{h^p} = c \neq 0,$$

则称求积公式 I_n 是 p 阶收敛的.

显然，复化梯形公式是二阶收敛的.

用复化梯形求积公式时，如果 T_n 不够精确，那么我们可以将每个子区间 $[x_k, x_{k+1}](k=0,1,\cdots,n-1)$ 对分，得到 $2n$ 个子区间，再用复化梯形公式计算. 此时，计算 T_n 的分点也是计算 T_{2n} 的分点. 因此，我们可以将复化梯形公式递推化，即有

$$T_{2n} = \frac{1}{2}T_n + \frac{h}{2}\sum_{k=0}^{n-1} f(x_{k+\frac{1}{2}}), \tag{4.17}$$

其中 $x_{k+\frac{1}{2}} = x_k + \frac{1}{2}h$. 这样，计算 T_{2n} 时，只需把新分点上的函数值算出，再利用 (4.17) 式即可.

4.2.2 复化 Simpson 求积公式

将积分区间 $[a,b]$ 分为 n 等份，$h = \frac{b-a}{n}$，$x_k = a+kh$，$k=0,1,\cdots,n$. 在每个子区间 $[x_k, x_{k+1}]$ 上用 Simpson 公式可得

$$\int_a^b f(x)\mathrm{d}x = \sum_{k=0}^{n-1}\int_{x_k}^{x_{k+1}} f(x)\mathrm{d}x \approx \frac{h}{6}\sum_{k=0}^{n-1}(f(x_k)+4f(x_{k+\frac{1}{2}})+f(x_{k+1})),$$

其中 $x_{k+\frac{1}{2}} = x_k + \frac{1}{2}h$. 令

$$S_n[f] = \frac{h}{6}\sum_{k=0}^{n-1}(f(x_k)+4f(x_{k+\frac{1}{2}})+f(x_k))$$
$$= \frac{h}{6}\left(f(a)+4\sum_{k=0}^{n-1} f(x_{k+\frac{1}{2}})+2\sum_{k=0}^{n-1} f(x_k)+f(b)\right), \tag{4.18}$$

称 $S_n[f]$ 为复化 Simpson 公式.

设 $f(x)\in C^4[a,b]$，由 Simpson 公式的误差有

$$R_{S_n} = I - S_n = \sum_{k=0}^{n-1}\left[-\frac{1}{90}\left(\frac{h}{2}\right)^5 f^{(4)}(\eta_k)\right], \quad \eta_k \in [x_k, x_{k+1}].$$

类似于复化梯形公式的推导，复化 Simpson 公式的余项为

$$R_{S_n} = -\frac{b-a}{2\,880}h^4 f^{(4)}(\eta), \quad \eta \in (a,b). \tag{4.19}$$

由此可见，复化 Simpson 公式是 4 阶收敛的.

例 4.3　分别用复分梯形公式和复化 Simpson 公式计算 $\int_0^{\pi}\sin x\mathrm{d}x$ 时，要使误差不超过 2×10^{-5}，问各取多少个节点？

解　由(4.16)式，令

$$|R_{T_n}| = \left|-\frac{b-a}{12}h^2 f''(\eta)\right| \leqslant \frac{\pi}{12}\left(\frac{\pi}{n}\right)^2 \max_{0\leqslant x\leqslant\pi}|\sin x| \leqslant 2\times10^{-5},$$

由此解得 $n^2 \geqslant \frac{\pi^3}{24}\times10^5$，$n \geqslant 360$.

由(4.19)式，令

$$|R_{S_n}| = \left|-\frac{b-a}{2\,880}h^4 f^{(4)}(\eta)\right| \leqslant \frac{\pi}{2\,880}\left(\frac{\pi}{n}\right)^4 \max_{0\leqslant k\leqslant\pi}|\sin x| \leqslant 2\times10^{-5},$$

由此解得 $n^4 \geqslant \frac{\pi^5}{5\,760}\times10^5$，$n \geqslant 9$.

因此，复化梯形公式取 361 个节点，复化 Simpson 公式取 19(即 $9\times2+1$)个节点. 可见，复化 Simpson 公式明显优于复化梯形公式.

例 4.4　把区间[1,2]分为 5 等份，用复化 Simpson 公式计算积分 $\int_1^2 \mathrm{e}^{\frac{1}{x}}\mathrm{d}x$ 的近似值，并估计误差.

解　此处 $h = 0.2$，$f(x) = \mathrm{e}^{\frac{1}{x}}$，节点

$$x_k = 1 + 0.2k, \quad k = 0,1,\cdots,5.$$

先算出节点 x_k 和 $x_{k+\frac{1}{2}}$ 处的函数值，见表 4-3(表中节点步长为 0.1).

表 4-3

k	x_k	$f(x_k)$	k	x_k	$f(x_k)$
0	1.0	2.718 282	6	1.6	1.868 246
1	1.1	2.482 065	7	1.7	1.800 808
2	1.2	2.300 976	8	1.8	1.742 909
3	1.3	2.158 106	9	1.9	1.692 685
4	1.4	2.042 727	10	2.0	1.648 721
5	1.5	1.947 734			

由(4.18)式得

$$\int_1^2 \mathrm{e}^{\frac{1}{x}}\mathrm{d}x \approx \frac{0.2}{6}\left(f(1)+f(2)+4\sum_{k=0}^{4}f(x_{k+\frac{1}{2}})+2\sum_{k=0}^{4}f(x_k)\right)=2.020\,077.$$

又由(4.19)式有

$$\begin{aligned}|R_s| &\leqslant \frac{1}{2\,880}(0.2)^4 \max_{1\leqslant x\leqslant 2}|f^{(4)}(x)| \\ &= \frac{(0.2)^4}{2\,880}\times 198.43 = 0.000\,110\,2.\end{aligned}$$

4.2.3 变步长求积法

复化求积公式的截断误差随 n 的增大而减小,但对于一个给定的积分,如何确定适当的 n,使得计算结果达到预先给定的精度要求呢?若用前面的误差估计式来求 n,则要用到高阶导数,一般是比较困难的. 在实际计算中,常采用自动选择积分步长的方法. 具体地说,就是在求积过程中,将步长逐次折半,反复利用复化求积公式,直到相邻两次的计算结果之差的绝对值小于允许误差为止. 这实际上是一种事后估计误差的方法.

对于复化梯形公式,由(4.16)式可知

$$I-T_n=-\frac{b-a}{12}\left(\frac{b-a}{n}\right)^2 f''(\eta_n),\quad \eta_n\in[a,b],$$

$$I-T_{2n}=-\frac{b-a}{12}\left(\frac{b-a}{2n}\right)^2 f''(\eta_{2n}),\quad \eta_{2n}\in[a,b].$$

当 $f''(x)$ 在区间$[a,b]$上连续,且其函数值变化不大时,即有 $f''(\eta_n)\approx f''(\eta_{2n})$,则有

$$\frac{I-T_n}{I-T_{2n}}\approx 4,$$

由此可得

$$I-T_{2n}\approx\frac{1}{3}(T_{2n}-T_n). \tag{4.20}$$

可见,对允许误差 ε,可用 $|T_{2n}-T_n|<\varepsilon$ 来判断近似值 T_{2n} 是否已满足精度要求. 若满足要求,则以 T_{2n} 为近似值,停止计算. 若不满足要求,则继续按形如(4.17)式的递推关系计算新的近似值 . 这就是基于复化梯形公式的变步长积分法.

对于复化 Simpson 公式,若 $f^{(4)}(x)$ 在$[a,b]$上连续且函数值变化不大,则类似复化梯形公式,由(4.19)式可推得

$$\frac{I-S_n}{I-S_{2n}}\approx 4^2,$$

由此即得

$$I - S_{2n} \approx \frac{1}{4^2 - 1}(S_{2n} - S_n). \tag{4.21}$$

若 $|S_{2n} - S_n| < \varepsilon$，则 S_{2n} 就是要求的近似值，否则，再将每个小区间分半进行计算，直到满足要求为止.

对于 $n = 4$ 的 Cotes 公式，将积分区间 $[a,b]$ 分为几等分，可得复化 Cotes 公式，积分近似值记为 C_n. 假设 $f^{(6)}(x)$ 在 $[a,b]$ 上连续且函数值变化不大，则可推得

$$\frac{I - C_n}{I - C_{2n}} \approx 4^3,$$

由此即得

$$I - C_{2n} \approx \frac{1}{4^3 - 1}(C_{2n} - C_n). \tag{4.22}$$

若 $|C_{2n} - C_n| < \varepsilon$，则 C_{2n} 就是要求的近似值，否则，再将每个小区间分半进行计算，直到满足要求为止.

例 **4.5**　利用变步长的复化梯形法计算

$$\int_0^1 \frac{\sin x}{x} \mathrm{d}x,$$

使截断误差不超过 0.5×10^{-3}.

解　按梯形公式和(4.17) 式有

$$T_1 = \frac{1}{2}(1 + 0.841\,471\,0) = 0.920\,735\,5,$$

$$T_2 = \frac{1}{2}T_1 + \frac{1}{2} \times 0.958\,851\,1 = 0.939\,793\,3.$$

计算近似误差

$$R_1 = \frac{1}{3}(T_2 - T_1) = 0.006\,352\,6,$$

不满足要求，由(4.17) 式有

$$T_4 = \frac{1}{2}T_2 + \frac{1}{4}(0.989\,615\,8 + 0.908\,851\,7) = 0.944\,513\,5.$$

计算近似误差

$$R_2 = \frac{1}{3}(T_4 - T_2) = 0.001\,573\,4,$$

不满足要求，由(4.17) 式有

$$T_8 = \frac{1}{2}T_4 + \frac{1}{8}(0.997\,397\,9 + 0.976\,726\,7 + 0.936\,155\,6 + 0.877\,192\,6)$$

$$= 0.945\,690\,9.$$

计算近似误差

$$R_3 = \frac{1}{3}(T_8 - T_4) = 0.000\ 392\ 5 \leqslant 0.5 \times 10^{-3},$$

满足要求，T_8 是满足要求的积分近似值.

4.3 外推原理与 Romberg 求积法

4.3.1 外推原理

在科学与工程计算中，很多算法与步长 h 有关，特别是数值积分、数值微分和微分方程数值解的问题. 对于这些算法，我们可以通过外推技巧提高计算精度. 先看一个计算 π 的近似值的例子，由函数 $\sin x$ 的 Taylor 展开式有

$$n\sin\frac{\pi}{n} = \pi - \frac{\pi^3}{3!n^2} + \frac{\pi^5}{5!n^4} - \cdots.$$

若记 $h = \frac{\pi}{6}$，$F(h) = 6\sin\frac{\pi}{6}$，则有

$$F(h) = \pi - \frac{\pi}{6}h^2 + \frac{\pi}{120}h^4 - \cdots,$$

$$F\left(\frac{h}{2}\right) = \pi - \frac{\pi}{6}\,\frac{1}{4}h^2 + \frac{\pi}{120}\,\frac{1}{16}h^4 - \cdots.$$

由此构造新的表达式：

$$F_1(h) = \frac{4F\left(\frac{h}{2}\right) - F(h)}{3} = \pi - \frac{\pi}{120}\,\frac{1}{4}h^4 + \cdots.$$

可见，计算 π 的近似值的算法 $F(h)$ 的截断误差是 $O(h^2)$，而算法 $F_1(h)$ 的截断误差是 $O(h^4)$. 外推一次，精度就提高了. 这就是外推法的基本思想. 若重复以上过程，不断外推，即不断折半步长 h，得到计算 π 的算法序列 $\{F_k(h)\}$. 随着 k 的增加，算法的截断误差阶越来越高，计算精度越来越好.

可将上述外推思想推广到一般情况. 设 $F(h)$ 是计算 $F(0)$ 的一种近似算式，带截断误差的表示式为

$$F(h) = F(0) + a_p h^p + O(h^s), \quad s > p,$$

其中，a_p 与 h 无关. 如果我们用 h 和 $\frac{h}{q}(q > 1)$ 两种步长分别计算 $F(h)$ 和 $F\left(\frac{h}{q}\right)$，则有

$$F\left(\frac{h}{q}\right) = F(0) + a_p\left(\frac{h}{q}\right)^p + O(h^s).$$

消去截断误差的主项，得新的算法

$$F_1(h)=\frac{q^pF\left(\frac{h}{q}\right)-F(h)}{q^p-1}=F(0)+O(h^s).$$

我们称这个计算过程为 Richardson 外推法. 这里，$F_1(h)$ 逼近 $F(0)$ 的截断误差是 $O(h^s)$.

只要知道 $F(h)$ 的更加完整的关于 h 幂的展开式，而无需知道展开式中各个系数的具体数值，就能重复使用 Richardson 外推法，直到截断误差达到容许误差. 用归纳法可以证明下面更一般的定理.

定理 4.4　假设 $F(h)$ 逼近 $F(0)$ 的余项为

$$F(h)-F(0)=a_1h^{p_1}+a_2h^{p_2}+a_3h^{p_3}+\cdots,$$

其中，$p_1<p_2<p_3<\cdots$，$a_k(k=1,2,\cdots)$ 是与 h 无关的非零常数，则由

$$F_0(h)=F(h),\quad F_{k+1}(h)=\frac{q^{p_k}F_k\left(\frac{h}{q}\right)-F_k(h)}{q^{p_k}-1},\quad k=0,1,\cdots,\tag{4.23}$$

定义的序列$\{F_n(h)\}$ 有

$$F_n(h)-F(0)=a_{n+1}^{(n)}h^{p_{n+1}}+a_{n+2}^{(n)}h^{p_{n+2}}+\cdots,$$

其中，$a_{n+k}^{(n)}(k=1,2,\cdots)$ 与 h 无关，$q>1$.

Richardson 外推法应用非常广泛和有效，下面应用于数值积分.

4.3.2　Romberg 求积法

先给出 Romberg 求积法的基础，即对于计算定积分 $I=I[f]$ 的复化梯形公式 $T(h)$，其余项为

$$I-T(h)=\sum_{k=1}^{m}\frac{\mathrm{B}_{2k}}{(2k)!}(f^{(2k-1)}(a)-f^{(2k-1)}(b))h^{2k}+r_{m+1},\tag{4.24}$$

其中，B_{2k} 为 Bernoulli 常数，

$$r_{m+1}=-\frac{\mathrm{B}_{2m+2}}{(2m+2)!}(b-a)f^{(2m+2)}(\eta)h^{2m+2},\quad \eta\in(a,b).$$

在外推算法(4.23) 式中，取 $q=2,p_k=2k$，由余项(4.24) 式可得著名的 Romberg 求积方法

$$\begin{cases}T_1^{(0)}=\dfrac{b-a}{2}(f(a)+f(b)),\\[2ex] T_1^{(i)}=\dfrac{1}{2}T_1^{(i-1)}+\dfrac{b-a}{2^i}\displaystyle\sum_{j=1}^{2^{i-1}}f\left(a+\frac{2j-1}{2^i}(b-a)\right),\quad i=1,2,\cdots,\\[2ex] T_{m+1}^{(k-1)}=\dfrac{4^mT_m^{(k)}-T_m^{(k-1)}}{4^m-1},\quad m=1,2,\cdots,\quad k=1,2,\cdots,i.\end{cases}$$

其中，$T_1^{(i)}$ 表示将积分区间$[a,b]$作 2^i 等分相应的复化梯形公式，求和项包括了每次等分后新增加点的函数值. $T_{m+1}^{(k)}$ 表示第 m 次外推所得的计算值. 可以验证，$m=1$ 时，所得外推值就是复化 Simpson 公式的计算值. 当 $m=1,2,3$ 时，Romberg 积分法所得的结果分别是将(4.20)、(4.21) 和(4.22) 式右边的值补充到相应的近似值上所得的结果.

对给定的精度标准 ε，我们可由

$$|T_m^{(0)}-T_{m-1}^{(0)}|<\varepsilon \text{ 或 } \left|\frac{T_m^{(0)}-T_{m-1}^{(0)}}{T_m^{(0)}}\right|<\varepsilon$$

作为计算终止的标准. 表 4－4 给出了计算过程，① 表示第 i 步计算.

表 4－4

k	T_1^k	T_2^k	T_3^k	T_4^k	T_5^k	…
0	① T_1^0	③ T_2^0	⑥ T_3^0	⑩ T_4^0	⑮ T_5^0	…
1	② T_1^1	⑤ T_2^1	⑨ T_3^1	⑭ T_4^1	⋮	
2	④ T_1^2	⑧ T_2^2	⑬ T_3^2	⋮		
3	⑦ T_1^3	⑫ T_2^3	⋮			
4	⑪ T_1^4	⋮				
⋮	⋮					

值得注意的是，若对某个 k，被积函数有性质 $f^{(2k-1)}(a)=f^{(2k-1)}(b)$，说明余项(4.24) 式中 h^{2k} 的系数为 0，则对 Romberg 求积法要作相应的修改，否则外推结果可能会差些.

例 4.6 用 Romberg 求积法计算定积分 $\int_0^1 \frac{\sin x}{x}\mathrm{d}x$，使计算值的误差不超过 $\varepsilon=0.5\times10^{-6}$.

解 $f(x)=\frac{\sin x}{x}$，$T_1^{(0)}=\frac{1}{2}(f(0)+f(1))=0.920\,735\,5.$

$$T_1^{(1)}=\frac{1}{2}T_1^{(0)}+\frac{1}{2}f\left(\frac{1}{2}\right)=0.939\,793\,3,$$

$$T_2^{(0)}=\frac{4}{3}T_1^{(1)}-\frac{1}{3}T_1^{(0)}=0.946\,145\,9,$$

$$T_1^{(2)}=\frac{1}{2}T_1^{(1)}+\frac{1}{4}\left(f\left(\frac{1}{4}\right)+f\left(\frac{3}{4}\right)\right)=0.944\,513\,5,$$

$$T_2^{(1)}=\frac{4}{3}T_1^{(2)}-\frac{1}{3}T_1^{(1)}=0.946\,086\,9,$$

$$T_3^{(0)} = \frac{16}{15}T_2^{(1)} - \frac{1}{15}T_2^{(0)} = 0.946\,083\,0,$$

$$| T_3^{(0)} - T_2^{(0)} | > 0.5 \times 10^{-6}.$$

此结果还没有满足精度要求，需继续进行外推，接着再计算 $T_1^{(3)}$，$T_2^{(2)}$，$T_3^{(1)}$ 和 $T_4^{(0)}$，于是得到计算结果如表 4－5. 由此看出，步长折半 3 次，复化梯形公式只达到 2 位有效数字，而经 3 次外推后达到 6 位有效数字.

表 4－5

k	$T_1^{(k)}$	$T_2^{(k)}$	$T_3^{(k)}$	$T_4^{(k)}$
0	0.920 735 5	0.946 145 9	0.946 083 0	0.946 083 1
1	0.939 793 3	0.946 086 9	0.946 083 1	
2	0.944 513 5	0.946 083 3		
3	0.945 690 9			

4.4　Gauss 求积公式

4.4.1　Gauss 求积公式的基本理论

在 Newton-Gotes 求积公式中，节点是等距的，从而限制了求积公式的代数精度. 下面的讨论将取消这个限制条件，使求积公式的代数精度尽可能高. 首先以简单情形论证这样做是可行的，然后给出概念和一般理论.

例 4.7　确定下列求积公式

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx A_0 f(x_0) + A_1 f(x_1)$$

中的待定参数，使其代数精度尽量高.

解　按代数精度的概念，分别令 $f(x) = 1, x, x^2, x^3$ 时上式左边与右边分别相等，有

$$\begin{aligned} A_0 + A_1 &= 2, \\ A_0 x_0 + A_1 x_1 &= 0, \\ A_0 x_0^2 + A_1 x_1^2 &= \frac{2}{3}, \\ A_0 x_0^3 + A_1 x_1^3 &= 0. \end{aligned}$$

由第二式和第四式可得 $x_0^2 = x_1^2$，结合第一式和第三式得 $x_0^2 = x_1^2 = \frac{1}{3}$. 取 $x_0 =$

$-\frac{1}{\sqrt{3}}, x_1 = \frac{1}{\sqrt{3}}$ 得 $A_0 = A_1 = 1$. 于是得到求积公式

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

它有 3 次代数精度,而以两个端点为节点的梯形公式只有 1 次代数精度.

一般地,考虑带权求积公式

$$\int_a^b \rho(x) f(x)\mathrm{d}x \approx \sum_{k=0}^{n} A_k f(x_k) \tag{4.25}$$

其中 $x_k, A_k (k = 0,1,\cdots,n)$ 为 $2n+2$ 个待定参数,适当选择这些参数,有可能使求积公式具有 $2n+1$ 次代数精度.

定义 4.3 如果求积公式(4.25)具有 $2n+1$ 次代数精度,则称该公式为 Gauss 型公式,称其节点 $x_k (k = 0,1,\cdots,n)$ 为 Gauss 点.

如果像例 4.7 那样,直接利用代数精度的概念去求 $n+1$ 个 Gauss 点和 $n+1$ 个求积系数,则要联立求解 $2n+2$ 个非线性方程组. 方程组是可解的,但当 n 稍大时,解析地求解就很难,数值求解非线性方程组也不容易. 下面从分析 Gauss 点的特性着手研究 Gauss 公式的构造问题.

定理 4.5 对于插值型求积公式(4.25),其节点 $x_k (k = 0,1,\cdots,n)$ 是 Gauss 点的充分必要条件是多项式 $\omega_{n+1}(x) = (x-x_0)(x-x_1)\cdots(x-x_n)$ 与任意不超过 n 次的多项式 $P(x)$ 带权正交,即

$$\int_a^b \rho(x) P(x) \omega_{n+1}(x)\mathrm{d}x = 0. \tag{4.26}$$

证 先证必要性. 设 $P(x)$ 是任意次数不超过 n 的多项式,则 $P(x)\omega_{n+1}(x)$ 的次数不超过 $2n+1$. 因此,如果 $x_0, x_1, \cdots, x_n$ 是 Gauss 点,则求积公式(4.25)对于 $P(x)\omega_{n+1}(x)$ 是准确成立的,即有

$$\int_a^b \rho(x) P(x) \omega_{n+1}(x)\mathrm{d}x = \sum_{k=0}^{n} A_k P(x_k) \omega_{n+1}(x_k).$$

但 $\omega_{n+1}(x_k) = 0 (k = 0,1,2,\cdots,n)$,故(4.26)式成立.

再证充分性. 设 $f(x)$ 是任意一个次数不超过 $2n+1$ 的多项式,用 $\omega_{n+1}(x)$ 除 $f(x)$,记商为 $P(x)$,余式为 $Q(x)$,即

$$f(x) = P(x)\omega_{n+1}(x) + Q(x),$$

其中 $P(x)$ 与 $Q(x)$ 都是次数不超过 n 的多项式. 利用(4.26)式有

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

由于(4.25)式是插值型的,它对于 $Q(x)$ 能准确成立,即

$$\int_a^b \rho(x)Q(x)\mathrm{d}x = \sum_{k=0}^{n} A_k Q(x_k).$$

注意到 $\omega_{n+1}(x_k) = 0$，知 $Q(x_k) = f(x_k)$，从而有

$$\int_a^b \rho(x)f(x)\mathrm{d}x = \sum_{k=0}^{n} A_k f(x_k).$$

由此可见，(4.25) 式对于一切次数不超过 $2n+1$ 的多项式均能准确成立. 因此，$x_k(k = 0,1,\cdots,n)$ 是 Gauss 点，定理得证.

由于 $n+1$ 次正交多项式与比它次数低的任意多项式正交，并且 $n+1$ 次正交多项式恰好有 $n+1$ 个互异的实的单根，我们有下面的推论.

推论　$n+1$ 次正交多项式的零点是 $n+1$ 点 Guass 公式的 Guass 点.

利用正交多项式得出 Guass 点 $x_0, x_1, \cdots, x_n$ 后，利用插值原理可得 Guass 公式的求积系数为

$$A_k = \int_a^b \rho(x)l_k(x)\mathrm{d}x, \quad k = 0,1,\cdots,n,$$

其中 $l_k(x)$ 是关于 Guass 点的 Lagrange 插值基函数.

例 4.8　确定 x_0, x_1, A_0, A_1，使下列公式为 Guass 公式

$$\int_0^1 \sqrt{1-x}f(x)\mathrm{d}x \approx A_0 f(x_0) + A_1 f(x_1).$$

解　我们可以像例 4.7 一样，直接由代数精度的概念构造该 Gauss 公式. 这里，我们用正交多项式的零点作为 Gauss 点的办法构造该 Gauss 公式.

先构造区间 $[0,1]$ 上关于权函数 $\rho(x) = \sqrt{1-x}$ 的正交多项式 $\{\varphi_j(x)\}$. 可用 3 项递推关系求出 $\{\varphi_j(x)\}$，这里我们直接利用正交性求解. 设 $\varphi_0(x) = 1, \varphi_1(x) = x + a, \varphi_2(x) = x^2 + bx + c$，则由

$$(\varphi_0, \varphi_1) = \int_0^1 \sqrt{1-x}(x+a)\mathrm{d}x = 0$$

得 $a = -\dfrac{2}{5}$. 由

$$(\varphi_0, \varphi_2) = \int_0^1 \sqrt{1-x}(x^2 + bx + c)\mathrm{d}x = 0$$

得 $\dfrac{2}{5}b + c + \dfrac{8}{35} = 0$. 由

$$(\varphi_1, \varphi_2) = \int_0^1 \sqrt{1-x}\left(x - \frac{2}{5}\right)(x^2 + bx + c)\mathrm{d}x = 0$$

得 $b = -\dfrac{8}{9}$，从而得 $c = \dfrac{8}{63}$. 由方程 $\varphi_2(x) = 0$ 得 $\varphi_2(x)$ 的零点 $x_0 = 0.1788$，$x_1 = 0.7101$.

按代数精度的概念，分别令 $f(x)=1, x$ 时公式准确成立，得

$$\begin{cases} A_0 + A_1 = \dfrac{2}{3}, \\ 0.178\,8A_0 + 0.710\,1A_1 = \dfrac{4}{15}. \end{cases}$$

由此解得 $A_0 = 0.389\,1, A_1 = 0.277\,6$. 从而得到 Gauss 求积公式

$$\int_0^1 \sqrt{1-x} f(x)\mathrm{d}x \approx 0.389\,1 f(0.178\,8) + 0.277\,6 f(0.710\,1).$$

4.4.2 常用 Gauss 求积公式

(1) Gauss-Legendre 求积公式.

在区间$[-1,1]$上取权函数$\rho(x)=1$，那么相应的正交多项式有 Legendre 多项式. 以 Legendre 多项式的零点为 Gauss 点的求积公式为

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \sum_{k=0}^{n} A_k f(x_k), \tag{4.27}$$

称之为 Gauss-Legendre 求积公式.

当 $n=1$ 时，2 次 Legendre 多项式 $P_2(x) = \dfrac{1}{2}(3x^2-1)$，零点为 $x_0 = -\dfrac{1}{\sqrt{3}}$，$x_1 = \dfrac{1}{\sqrt{3}}$. 此时，(4.27) 式即为例 4.7 所给的公式.

当 $n=2$ 时，3 次 Legendre 多项式 $P_3(x) = \dfrac{1}{2}(5x^3-3x)$，零点为 $x_0 = -\dfrac{\sqrt{15}}{5}$，$x_1 = 0, x_2 = \dfrac{\sqrt{15}}{5}$. 以此为 Gauss 点，仿两点 Gauss-Legendre 求积公式，求出相应的求积系数，可构造出具有 5 次代数精度的 3 点 Gauss-Legendre 求积公式

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \frac{5}{9} f\left(-\frac{\sqrt{15}}{5}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\frac{\sqrt{15}}{5}\right).$$

部分 Gauss-Legendre 求积公式中的 Gauss 点 x_k 和求积系数 A_k 见表 4－6.

表 4－6

n	x_k	A_k	n	x_k	A_k
0	0	2	3	±0.861 136 311 6	0.347 854 845 1
1	±0.577 350 269 2	1		±0.339 981 043 6	0.652 145 154 9
2	±0.774 596 669 2	0.555 555 555 5	4	±0.906 179 845 9	0.236 926 885 1
	0	0.888 888 888 8		±0.538 469 310 1	0.478 628 670 5
				0	0.568 888 889

对于一般区间$[a,b]$上的求积，如果用 Gauss-Legendre 求积公式，那么必须作变量替换

$$x=\frac{1}{2}(a+b)+\frac{1}{2}(b-a)t,$$

使 $x\in[a,b]$ 时，$t\in[-1,1]$，并有

$$\int_a^b f(x)\mathrm{d}x=\frac{b-a}{2}\int_{-1}^{1} f\left(\frac{1}{2}(a+b)+\frac{1}{2}(b-a)t\right)\mathrm{d}t.$$

对于上式右边的积分可以应用 Gauss-Legendre 求积公式.

例 4.9　用 Gauss-Legendre 求积公式($n=1,2$) 计算积分 $I=\int_0^1 x^2\mathrm{e}^x\mathrm{d}x$.

解　由于区间为$[0,1]$，所以先作变量替换 $x=\frac{1+t}{2}$，得

$$I=\int_0^1 x^2\mathrm{e}^x\mathrm{d}x=\frac{1}{8}\int_{-1}^{1}(t+1)^2\mathrm{e}^{\frac{1+t}{2}}\mathrm{d}t.$$

令 $f(t)=(1+t)^2\mathrm{e}^{\frac{1+t}{2}}$，对于 $n=1$，由两点 Gauss-Legendre 公式有

$$I\approx\frac{1}{8}\left(f\left(-\frac{1}{\sqrt{3}}\right)+f\left(\frac{1}{\sqrt{3}}\right)\right)=0.711\ 941\ 774.$$

对于 $n=2$，由 3 点 Gauss-Legendre 公式有

$$I\approx\frac{1}{8}\left(\frac{5}{9}f\left(-\frac{\sqrt{15}}{5}\right)+\frac{8}{9}f(0)+\frac{5}{9}f\left(\frac{\sqrt{15}}{5}\right)\right)=0.718\ 251\ 799.$$

容易求出定积分的精确值为

$$I=\mathrm{e}-2=0.718\ 281\ 828.$$

由此可见，$n=1$ 时的实际误差为 0.006 340 054，$n=2$ 时的实际误差为 0.000 030 049.

例 4.10　试证明求积公式

$$\int_1^3 f(x)\mathrm{d}x\approx\frac{5}{9}f\left(2-\sqrt{\frac{3}{5}}\right)+\frac{8}{9}f(2)+\frac{5}{9}f\left(2+\sqrt{\frac{3}{5}}\right)$$

具有 5 次代数精度.

证　此为 3 点公式，而 3 点 Gauss-Legendre 求积公式具有 5 次代数精度. 因此，只要证明它就是 3 点 Gauss-Legendre 求积公式即可.

令 $x=t+2$，则

$$\int_1^3 f(x)\mathrm{d}x=\int_{-1}^{1} f(t+2)\mathrm{d}t=\int_{-1}^{1} g(t)\mathrm{d}t$$

Legendre 多项式 $p_3(t)=\frac{1}{2}(5t^3-t)$ 的零点为

$$t_0 = -\sqrt{\frac{3}{5}}, t_1 = 0, t_2 = \sqrt{\frac{3}{5}}.$$

若公式

$$\int_{-1}^{1} g(t)\mathrm{d}t \approx \sum_{k=0}^{2} A_k g(t_k)$$

中的节点取 $p_3(t)$ 的零点，系数 $A_k = \int_{-1}^{1} l_k(x)\mathrm{d}x$，$l_k(x)$ 是关于节点 t_0, t_1, t_2 的 Lagrange 基函数，则该公式是 Gauss-Legendre 求积公式，具有 5 次代数精度.

经计算得 $A_1 = \frac{8}{9}$，$A_0 = A_2 = \frac{5}{9}$，结论得证.

(2) Gauss-Chebyshev 求积公式.

在区间$[-1,1]$上取权函数 $\rho(x) = \frac{1}{\sqrt{1-x^2}}$ 的正交多项式有 Chebyshev 正交多项式. $n+1$ 次 Chebyshev 多项式 $T_{n+1}(x) = \cos[(n+1)\arccos x]$ 的零点为

$$x_k = \cos\frac{2k+1}{2n+2}\pi, \quad k = 0,1,\cdots,n.$$

以此为 Gauss 点，利用 Chebyshev 多项式的性质可得相应的求积系数为

$$A_k = \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} l_k(x)\mathrm{d}x = \frac{\pi}{n+1}, \quad k = 0,1,\cdots,n,$$

其中 $l_k(x)$ 是关于 Gauss 点的 Lagrange 插值基函数. 从而有 Gauss-Chebyshev 求积公式

$$\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} f(x)\mathrm{d}x \approx \frac{\pi}{n+1}\sum_{k=0}^{n} f(x_k).$$

对于 $n = 1,2$ 点 Gauss-Chebyshev 求积公式为

$$\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} f(x)\mathrm{d}x \approx \frac{\pi}{2}\left(f\left(-\frac{1}{\sqrt{2}}\right) + f\left(\frac{1}{\sqrt{2}}\right)\right).$$

对于 $n = 2,3$ 点 Gauss-Chebyshev 求积公式为

$$\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} f(x)\mathrm{d}x \approx \frac{\pi}{3}\left(f\left(-\frac{\sqrt{3}}{2}\right) + f(0) + f\left(\frac{\sqrt{3}}{2}\right)\right).$$

例 4.11 计算积分 $\int_{-1}^{1} \sqrt{\frac{2+x}{1-x^2}}\mathrm{d}x$.

解 选用 $n = 2$ 的 Gauss-Chebyshev 求积公式计算，这时，$f(x) = \sqrt{2+x}$. 于是有

$$\int_{-1}^{1} \sqrt{\frac{2+x}{1-x^2}}\mathrm{d}x \approx \frac{\pi}{3}\left(\sqrt{2-\frac{\sqrt{3}}{2}} + \sqrt{2} + \sqrt{2+\frac{\sqrt{3}}{2}}\right) = 4.368\,939\,556.$$

4.4.3　Gauss 求积公式的余项与稳定性

定理 4.6　设 $f(x)\in C^{2n+2}[a,b]$，则 Gauss 公式(4.25) 的余项是

$$\begin{aligned}R_G&=\int_a^b\rho(x)f(x)\mathrm{d}x-\sum_{k=0}^{n}A_kf(x_k)\\&=\frac{1}{(2n+2)!}f^{(2n+1)}(\eta)\int_a^b\rho(x)\omega_{n+1}^2(x)\mathrm{d}x,\quad \eta\in(a,b).\end{aligned}$$

证　由 Gauss 点 $x_k(k=0,1,\cdots,n)$ 构造次数不超过 $2n+1$ 的 Hermite 插值多项式 $H(x)$，满足条件

$$H(x_k)=f(x_k),\quad H'(x_k)=f'(x_k),\quad k=0,1,\cdots,n.$$

由于 Gauss 公式具有 $2n+1$ 次代数精度，它对于 $H(x)$ 能准确成立，即

$$\int_a^b\rho(x)H(x)\mathrm{d}x=\sum_{k=0}^{n}A_kH(x_k)=\sum_{k=0}^{n}A_kf(x_k).$$

由 Hermite 插值多项式的插值余项有

$$\begin{aligned}R_G&=\int_a^b\rho(x)f(x)\mathrm{d}x-\sum_{k=0}^{n}A_kH(x_k)\\&=\int_a^b\rho(x)f(x)\mathrm{d}x-\int_a^b\rho(x)H(x)\mathrm{d}x\\&=\int_a^b\rho(x)\,\frac{f^{(2n+2)}(\xi)}{(2n+2)!}\omega_{n+1}^2(x)\mathrm{d}x.\end{aligned}$$

再考虑到 $\omega_{n+1}^2(x)$ 在$[a,b]$ 上保号，应用积分中值定理，定理得证.

对于两点 Gauss-Legendre 求积公式有

$$\begin{aligned}R_{G\text{-}L}&=\frac{f^{(4)}(\eta)}{4!}\int_{-1}^{1}\left(x+\frac{1}{\sqrt{3}}\right)^2\left(x-\frac{1}{\sqrt{3}}\right)^2\mathrm{d}x\\&=\frac{f^{(4)}(\eta)}{135},\quad \eta\in(-1,1).\end{aligned}$$

对于两点 Gauss-Chebyshev 求积公式有

$$\begin{aligned}R_{G\text{-}C}&=\frac{f^{(4)}(\eta)}{4!}\int_{-1}^{1}\frac{1}{\sqrt{1-x^2}}\left(x+\frac{1}{\sqrt{2}}\right)^2\left(x-\frac{1}{\sqrt{2}}\right)^2\mathrm{d}x\\&=\frac{\pi f^{(4)}(\eta)}{192},\quad \eta\in(-1,1).\end{aligned}$$

对比 Newton-Cotes 求积公式，Gauss 求积公式不但具有高精度，而且是数值稳定的. Gauss 公式的稳定性之所以能够得到保证，是由于它的求积系数具有非负性.

引理　Gauss 求积公式(4.25) 中的系数 $A_k(k=0,1,\cdots,n)$ 全部为正.

证　对于以 Gauss 点 $x_k(k=0,1,\cdots,n)$ 为节点的插值基函数 $l_i(x)(i=0,1,$

$\cdots,n)$,$l_i^2(x)$ 是 $2n$ 次多项式,故 Gauss 公式(4.25) 对于它能准确成立,即有

$$\int_a^b \rho(x) l_i^2(x)\mathrm{d}x = \sum_{k=0}^{n} A_k l_i^2(x_k) = A_i.$$

由于上式左端大于零,所以有 $A_i > 0(i = 0,1,\cdots,n)$.

在实际计算积分的近似值

$$I_n = \sum_{k=0}^{n} A_k f(x_k)$$

时,$f(x_k)$ 不能精确地取到,一般只能是近似值,设 $f^*(x_k) \approx f(x_k)(k = 0,1,\cdots,n)$,实际求得的积分值为

$$I_n^* = \sum_{k=0}^{n} A_k f^*(x_k).$$

定理 4.7 对于函数值的变化所引起的求积公式的误差有

$$|I_n^* - I_n| \leqslant \max_{0\leqslant k\leqslant n} |f^*(x_k) - f(x_k)| \int_a^b \rho(x)\mathrm{d}x.$$

证 由于求积系数 $A_k > 0$,$k = 0,1,\cdots,n$,因此有

$$\begin{aligned} |I_n^* - I_n| &= \left| \sum_{k=0}^{n} A_k f^*(x_k) - \sum_{k=0}^{n} A_k f(x_k) \right| \\ &\leqslant \sum_{k=0}^{n} A_k |f^*(x_k) - f(x_k)| \\ &\leqslant \max_{0\leqslant k\leqslant n} |f^*(x_k) - f(x_k)| \sum_{k=0}^{n} A_k. \end{aligned}$$

在 Gauss 求积公式中,取 $f(x) = 1$,此时求积公式精确成立,即得

$$\int_a^b \rho(x)\mathrm{d}x = \sum_{k=0}^{n} A_k.$$

因此,定理得证.

由定理 4.7 可知,数据误差对于求积公式计算值的影响是可以控制的,即 Gauss 求积公式在数值计算中是稳定的.

4.5 数 值 微 分

数值微分就是用离散方法近似地求出函数在某点的导数值. 按照 Taylor 展开原理可得

$$f'(x) = \frac{f(x+h) - f(x)}{h} + O(h),$$

$$f'(x) = \frac{f(x) - f(x-h)}{h} + O(h),$$

$$f'(x)=\frac{f(x+h)-f(x-h)}{2h}+O(h^2),$$

其中 h 为一增量. 上面几个公式是很实用的,下面我们再讨论一些常用方法.

4.5.1　插值型求导公式

设 $f(x)$ 是定义在$[a,b]$上的函数,并给定区间$[a,b]$上的 $n+1$ 个节点 x_k 处的函数值 $f(x_k)$,$k=0,1,\cdots,n$. 这样,我们可以建立函数 $f(x)$ 的 n 次插值多项式 $P_n(x)$. 多项式的求导是容易的,称

$$f'(x)\approx P_n'(x) \tag{4.28}$$

为插值型求导公式.

应当指出,即使 $f(x)$ 与 $P_n(x)$ 的值相差不多,导数的近似值 $P_n'(x)$ 与导数的值 $f'(x)$ 仍然可能相差很大. 因而在使用求导公式(4.28) 时,应注意进行误差分析.

依据插值余项定理,求导公式(4.28) 的余项为

$$f'(x)-P_n'(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}\omega'_{n+1}(x)+\frac{\omega_{n+1}(x)}{(n+1)!}\frac{\mathrm{d}}{\mathrm{d}x}f^{(n+1)}(\xi),$$

式中 $\omega_{n+1}(x)=(x-x_0)(x-x_1)\cdots(x-x_n)$.

在上述余项公式中,由于 ξ 是 x 的未知函数,我们无法对右端的第二项作出进一步的说明. 因此,对于随意给出的点 x,求导公式的余项是很难估计的. 然而,如果我们限定求节点的导数值,那么有余项公式

$$f'(x_k)-P_n'(x_k)=\frac{f^{(n+1)}(\xi)}{(n+1)!}\omega'_{n+1}(x_k). \tag{4.29}$$

下面我们考察节点处的导数值. 为简化讨论,假定所给的节点是等距的,h 是步长.

(1) 两点公式.

当 $n=1$ 时,由(4.29) 式得带余项的两点公式.

$$f'(x_0)=\frac{1}{h}(f(x_1)-f(x_0))-\frac{h}{2}f''(\xi), \tag{4.30}$$

$$f'(x_1)=\frac{1}{h}(f(x_1)-f(x_0))+\frac{h}{2}f''(\xi). \tag{4.31}$$

(2) 三点公式.

当 $n=2$ 时,由(4.29) 式得带余项的三点公式.

$$f'(x_0)=\frac{1}{2h}(-3f(x_0)+4f(x_1)-f(x_2))+\frac{h^2}{5}f''(\xi), \tag{4.32}$$

$$f'(x_1)=\frac{1}{2h}(-f(x_0)+f(x_2))-\frac{h^2}{6}f''(\xi) \tag{4.33}$$

$$f'(x_2)=\frac{1}{2h}(f(x_0)-4f(x_1)+3f(x_2))+\frac{h^2}{3}f''(\xi). \tag{4.34}$$

(3) 五点公式.

当 $n=4$ 时,由(4.29) 式不难导出带余项的五点求导公式. 这里,给出其中常用的五点公式.

$$f'(x_2)=\frac{1}{12h}(f(x_0)-8f(x_1)+8f(x_3)-f(x_4))+\frac{h^2}{30}f^{(5)}(\xi) \tag{4.35}$$

例 4.12 设 $f(x)=e^x$,对 $h=0.01$,计算 $f'(1.8)$ 的近似值.

解 由(4.32) 式有

$$f'(1.8)\approx\frac{1}{2h}(-3f(1.8)+4f(1.81)-f(1.82))=6.049\,4.$$

由(4.33) 式有

$$f'(1.8)\approx\frac{1}{2h}(f(1.81)-f(1.79))=6.049\,7$$

由(4.34) 式有

$$f'(1.8)\approx\frac{1}{2h}(f(1.78)-4f(1.79)+3f(1.8))=6.049\,4.$$

由(4.35) 式有

$$\begin{aligned}f'(1.8)&\approx\frac{1}{12h}(f(1.78)-8f(1.79)+8f(1.81)-f(1.82))\\&=6.049\,6.\end{aligned}$$

精确值 $e^{1.8}=6.049\,6$. 计算结果显然与它们的余项相一致,由(4.35) 式计算所得的结果最精确.

用插值多项式 $P_n(x)$ 作为 $f(x)$ 的近似函数,还可以建立高阶数值微分公式

$$f^{(k)}(x)\approx P_n^{(k)}(x),\quad k=1,2,\cdots.$$

然而,对于用插值法建立的数值求导公式,通常导数值的精确度比用插值公式求得的函数值的精确度差,高阶导数值的精度比低阶导数值的精度差. 所以,不宜用此方法建立高阶数值求导公式.

4.5.2 3 次样条求导

我们知道,3 次样条函数 $S(x)$ 作为 $f(x)$ 的近似函数,不但彼此的函数值很接近,导数值也很接近. 因此,用样条函数建立数值微分公式是很自然的.

设在区间$[a,b]$ 上,给定一种划分 $a=x_0<x_1<\cdots<x_n=b$,$h_k=x_{k+1}-x_k$ 及相应的函数值 $y_k=f(x_k)$,$k=0,1,\cdots,n$. 再给定适当的边界条件,按 3 次样条函数的算法,建立关于节点上的一阶导数 m_k 或 2 阶导数 M_k 的样条方程组,求得 m_k

或 $M_k, k=0,1,\cdots,n$，从而得到 3 次样条插值函数 $S(x)$ 的表达式. 这样，可得数值微分的公式

$$f^{(i)}(x) \approx S^{(i)}(x), i=0,1,2. \tag{4.36}$$

与前面插值型数值微分公式不同，样条数值微分公式(4.36) 式可以用来计算插值范围内任何一点(不仅是节点) 上的导数值. 误差估计由(2.64) 式给出.

对于节点上的导数值，若求得的是 $m_k, k=0,1,\cdots,n$，则由 $S(x)$ 的表达式有

$$f'(x_k) \approx m_k,$$

$$f''(x_k) \approx S''(x_k) = -\frac{2}{h_k}(2m_k + m_{k+1}) + \frac{6}{h_k} f(x_k, x_{k+1}).$$

若求得的是 $M_k, k=0,1,\cdots,n$，则由 $S(x)$ 的表达式有

$$f'(x_k) \approx S'(x_k) = -\frac{h_k}{6}(2M_k + M_{k+1}) + f(x_k, x_{k+1}),$$

$$f''(x_k) \approx M_k.$$

4.5.3　数值微分的外推算法

先看一个简单的例子. 求 $f(x) = -\cot x$ 在 $x=0.004$ 处的一阶导数值. 采用中点微分公式(4.33)，即

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h},$$

取 $h = 0.001\ 6$，那么得 $f'(0.004) \approx 626.335\ 043\ 8$，实际上 $f'(0.004) = 625.333\ 440\ 02$. 由此看出，仅有 2 位有效数字. 利用 Richardson 外推法可以提高计算精度.

对于中心差商，记

$$f'(x) \approx G(h) = \frac{1}{2h}(f(x+h) - f(x-h)).$$

由 Taylor 级数展开有

$$f'(x) - G(h) = f'(x_n) + \frac{h^2}{6} f'''(x_n) + \frac{h^4}{120} f^{(5)}(x_n) + \cdots.$$

利用 Richardson 外推公式，取 $q=2, P_k = 2k$，则有

$$\begin{cases} G_1(h) = G(h), \\ G_{m+1}(h) = \dfrac{4^m G_m\left(\dfrac{h}{2}\right) - G_m(h)}{4^m - 1}, \quad m = 1,2,\cdots. \end{cases} \tag{4.37}$$

外推公式(4.37) 的终止标准是 $|G_{m+1}(h) - G_m(\frac{h}{2})| < \varepsilon$，$\varepsilon$ 是预先给定的误差小量.

例 4.13　设 $f(x) = x^2 \mathrm{e}^{-x}$，当 h 分别取 0.1，0.05，0.025 时，求出 $x=0.5$ 处

的一阶导数的中心差商，进行外推，并与精确值进行比较.

解　先分别取 $h=0.1,0.05,0.025$，求出节点 $x=0.5$ 处的中心差商值，见表 4－7，再按(4.37) 式进行外推，外推两次，结果列于表 4－7 中.

表 4－7

h	$G_1(h)$	$G_2(h)$	$G_3(h)$	$f'(0.5)$
0.1	0.451 604 908 1	0.454 899 923 1	0.454 897 994	0.454 897 994
0.05	0.454 076 169 3	0.454 898 115 2		
0.025	0.454 692 628 8			

从表 4－7 可见，$h=0.025$ 时的中心差商值只有 3 位有效数字，外推一次达到 5 位有效数字，外推两次达到 9 位有效数字.

评　注

本章介绍积分和微分的数值计算方法，着重论述了 Newton-Cotes 求积公式、Romberg 求积公式和 Gauss 求积公式. 我们知道，积分和微分是两种分析运算，它们都是用极限来定义的. 数值积分和数值微分则归结为函数值的四则运算，从而使计算过程可以在计算机上完成. 处理数值积分和数值微分的基本方法是逼近法. 本章基于插值原理推导了数值积分和数值微分的基本公式.

Newton-Cotes 求积公式和 Gauss 求积公式都是插值型求积公式. 前者取等距节点，算法简单而容易编制程序. Gauss 求积公式采用正交多项式的零点作为节点，从而具有较高的精度，但节点没有规律. 运用带权的 Gauss 公式，能把复杂的求积问题化简，还可以直接计算奇异积分. 由于高阶 Newton-Cotes 公式的不稳定性，所以实际计算采用复化求积公式为宜. Gauss 求积公式是稳定的，但高阶求积方法的准备工作较繁杂. 因此，复化 Gauss 求积方法也是一个良好的方法.

基于数值积分的误差估计式，变步长求积公式不需要一开始就确定步长，且在随后的每一步都能估计出近似积分值的误差. 因此，变步长求积方法是一种易于执行的方法. 变步长积分法的不足是收敛慢.

Romberg 求积方法，通过利用误差不断修正近似积分值，有效地加快了收敛速度，并且程序简单，精度较高，因而是一个可选取的方法. 当节点加密提高积分近似程度时，前面计算的结果可以为后面的计算时使用，因此，对减少计算量很有好处. 该方法有比较简单的误差估计方法，能同时得到若干积分序列. 如果在作收敛性控制时，同时检验主对角线序列、梯形求积序列和抛物线求积序列，那么对不同形态的函数，可以用其中最快的收敛序列来逼近积分值.

外推原理是提高计算精度的一种重要技巧，应用很广泛，特别适用于数值微分、数值积分、常微分方程和偏微分方程数值解等问题.

奇异积分、振荡积分、二维和多维积分的求积方法，特别是 Monte Carlo 求积方法，都是数值积分的重要课题，限于篇幅，本章未作介绍，可参考有关专著.

习　题　4

4.1　确定下列求积公式的待定参数，使其代数精度尽量高，并指出其代数精度的次数.

(1) $\int_{-h}^{h} f(x)\mathrm{d}x \approx A_0 f(-h) + A_1 f(0) + A_2 f(h)$；

(2) $\int_{-2h}^{2h} f(x)\mathrm{d}x \approx A_0 f(-h) + A_1 f(0) + A_2 f(h)$；

(3) $\int_{0}^{1} f(x)\mathrm{d}x \approx A_0 f(0) + A_1 f(1) + A_2 f'(0)$.

4.2　证明求积公式

$$\int_{x_0}^{x_1} f(x)\mathrm{d}x \approx \frac{h}{2}(f(x_0) + f(x_1)) - \frac{h^2}{12}(f'(x_1) - f'(x_0))$$

具有 3 次代数精度，其中 $h = x_1 - x_0$.

4.3　用 Simpson 公式计算积分 $\int_0^1 \mathrm{e}^{-x}\mathrm{d}x$，并估计误差.

4.4　给定数据表 4－8 所示.

表 4－8

x	1.8	2.0	2.2	2.4	2.6
$f(x)$	3.120 14	4.425 69	6.042 41	8.030 14	10.466 75

分别用复化梯形公式和复化 Simpson 公式计算积分 $\int_{1.8}^{2.6} f(x)\mathrm{d}x$ 的近似值.

4.5　使用复化梯形公式和复化 Simpson 公式计算积分 $\int_1^3 \mathrm{e}^x \sin x\mathrm{d}x$，要求误差不超过 10^{-4}，不计舍入误差，问各需计算多少个节点上的函数值？

4.6　设 $f(x)$ 在 $[a,b]$ 上可积，证明当 $n \to \infty$ 时，复化 Simpson 公式趋于所计算的积分值.

4.7　用 Romberg 方法计算积分 $\frac{2}{\sqrt{\pi}}\int_0^1 \mathrm{e}^{-x}\mathrm{d}x$，要求误差不超过 10^{-5}.

4.8　用 3 点 Gauss-Legendre 求积公式计算积分 $\int_0^1 \frac{4}{1+x^2}\mathrm{d}x (= \pi)$ 的近似值.

4.9 用两点 Gauss-Chebyshev 求积公式计算积分$\int_{-1}^{1}\frac{1-x^2}{\sqrt{1-x^2}}\mathrm{d}x$的近似值.

4.10 用三点公式求 $f(x)=\frac{1}{(1+x)^2}$ 在 $x=1.0,1.1$ 和 1.2 处的导数值,并估计误差. $f(x)$ 的值由表 4-9 给出.

表 4-9

x	1.0	1.1	1.2
$f(x)$	0.250 0	0.226 8	0.206 6

4.11 给定 $f(x)=\sqrt{x}$ 在节点 $x_k=100+kh(h=1,k=0,1,2,3)$ 上的函数值和两个端点的导数值 $f'(100)$ 和 $f'(103)$. 用 3 次样条求导法,计算 $f'(101)$, $f'(101.5)$, $f'(102)$ 和 $f''(101.5)$ 的近似值.

数值试验题 4

4.1 用复化梯形公式、复化 Simpson 公式、Romberg 方法和复化 Gauss-Legendre公式计算下列积分的近似值,使绝对误差限为 0.5×10^{-7},并将计算结果与精确作比较以及比较各种算法的计算量.

(1) $\int_1^2\frac{1}{x}\mathrm{d}x=\ln 2$; (2) $\int_0^1\frac{1}{1+x^2}\mathrm{d}x=\frac{\pi}{4}$.

4.2 用外推方法计算下列积分值,并对计算结果进行比较. 如果所得结果不满意,对算法进行适当修改.

(1) $\int_0^1\left(\frac{x}{1+x^2}+\frac{x^2}{2}\right)\mathrm{d}x=0.513\,240\,25\cdots$;

(2) $\int_0^{\pi}\sin^2 x\mathrm{d}x=\frac{\pi}{2}$.

4.3 用样条函数方法和外推法求下列函数的一阶和二阶导数,并结合函数的图形说明精度与步长 h 的关系:

(1) $f(x)=\frac{1}{16}x^6-\frac{3}{10}x^2$, $-2\leqslant x\leqslant 2$;

(2) $f(x)=\mathrm{e}^{-x^2}\cos 20x$, $0\leqslant x\leqslant 2$.

4.4 设计自适应的 Simpson 方法求积分 $\int_0^1 x\sqrt{x}\mathrm{d}x(=0.4)$ 的近似值,即对不同的子区间分别按精度标准确定各自适当的步长,计算各子区间上的积分近似值,然后将各个近似值相加,要求近似值的绝对误差限为 0.5×10^{-7}.

4.5　编写 Gauss 求积法计算积分的程序(Gauss 点数取 1,2,3,4,5 即可),并用于计算积分 $I=\int_0^1 \frac{\sin x}{x}\mathrm{d}x$.

4.6　构造 Guass 型求积公式

$$\int_0^1 \frac{f(x)}{1+x^2}\mathrm{d}x \approx A_1 f(x_1)+A_2 f(x_2)+A_3 f(x_3),$$

并计算积分

$$\int_0^1 \frac{x^4}{1+x^2}\mathrm{d}x;\quad \int_0^1 \frac{\mathrm{e}^{-x}}{1+x^2}\mathrm{d}x;\quad \int_0^1 \frac{x^4}{\sqrt{1+x^2}}\mathrm{d}x;\quad \int_0^1 \frac{\mathrm{e}^{-x}}{\sqrt{1+x^2}}\mathrm{d}x.$$

第 5 章　线性方程组的直接解法

在科学与工程计算中，大量的问题归结为求解线性代数方程组 $\boldsymbol{Ax}=\boldsymbol{b}$，其中 $\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{n\times n}$，$\boldsymbol{b}=(b_1,b_2,\cdots,b_n)^{\mathrm{T}}\in\mathbf{R}^n$，$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}\in\mathbf{R}^n$ 分别称为方程组的系数矩阵、右端向量和解向量. 若 $\boldsymbol{A}$ 可逆，则方程组存在唯一解.

对于中小型方程组，常用直接解法. 从本质上说，直接方法的原理是找到一个可逆矩阵 $\boldsymbol{M}$，使得 $\boldsymbol{MA}$ 是一个上三角阵，这一过程一般称为“消元”过程，消元之后再进行“回代”，即求解 $\boldsymbol{MAx}=\boldsymbol{Mb}$. 这类直接方法中最基本和简单的就是 Gauss 消元法. 本章讨论 Gauss 消去法及其变形，以及一些情况下的特殊方法，最后进行误差分析.

5.1　Gauss 消 去 法

5.1.1　Gauss 消去法的计算过程

我们把方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 写成

$$\begin{cases} a_{11}x_1+a_{12}x_2+\cdots+a_{1n}x_n=b_1=a_{1,n+1}, \\ a_{21}x_1+a_{22}x_2+\cdots+a_{2n}x_n=b_2=a_{2,n+1}, \\ \qquad\qquad\cdots\cdots \\ a_{n1}x_1+a_{n2}x_2+\cdots+a_{nn}x_n=b_n=a_{n,n+1}. \end{cases} \tag{5.1}$$

设方程组(5.1)的系数矩阵 $\boldsymbol{A}$ 非奇异，记 $a_{ij}^{(1)}=a_{ij}\ (i=1,2,\cdots,n;j=1,2,\cdots,n+1)$，$(\boldsymbol{A}^{(1)},\boldsymbol{b}^{(1)})=(\boldsymbol{A},\boldsymbol{b})$. 这样，方程组(5.1)又可写成 $\boldsymbol{A}^{(1)}\boldsymbol{x}=\boldsymbol{b}^{(1)}$. 消元过程就是要按确定的计算过程对方程组进行初等行变换，将方程组化为上三角方程组.

第一步消元：假设 $a_{11}^{(1)}\neq0$，作初等行变换运算，即保留第 1 个方程并利用它分别与其余方程消去第 1 个未知量，这时，第 2 个至第 n 个方程的未知量的系数和常数项一般都有改变，分别记之为 $a_{ij}^{(2)}$，$i=2,3,\cdots,n;j=2,3,\cdots,n+1$. 即令

$$l_{i1}=\frac{a_{i1}^{(1)}}{a_{11}^{(1)}},\quad i=2,3,\cdots,n.$$

则有

$$a_{ij}^{(2)}=a_{ij}^{(1)}-l_{i1}a_{1j}^{(1)},\quad i=2,3,\cdots,n;\quad j=2,3,\cdots,n+1.$$

用增广矩阵表示，就是将 $(\boldsymbol{A}^{(1)},\boldsymbol{b}^{(1)})$ 变换为

$$(\boldsymbol{A}^{(2)},\boldsymbol{b}^{(2)})=\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} & a_{1,n+1}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & a_{2,n+1}^{(2)} \\ & \vdots & & \vdots & \vdots \\ & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} & a_{n,n+1}^{(2)} \end{pmatrix}.$$

它对应的方程组 $\boldsymbol{A}^{(2)}\boldsymbol{x}=\boldsymbol{b}^{(2)}$ 与(5.1)式等价，而在第 2 个至第 n 个方程中，含 x_1 的项已经消去.

第 k 步消元：设消去法已进行 $k-1$ 步，得到方程组 $\boldsymbol{A}^{(k)}\boldsymbol{x}=\boldsymbol{b}^{(k)}$，此时对应的增广矩阵是

$$(\boldsymbol{A}^{(k)},\boldsymbol{b}^{(k)})=\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & \cdots & \cdots & a_{1n}^{(1)} & a_{1,n+1}^{(1)} \\ & a_{22}^{(2)} & \cdots & \cdots & \cdots & a_{2n}^{(2)} & a_{2,n+1}^{(2)} \\ & & \ddots & & \vdots & \vdots & \vdots \\ & & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} & a_{k,n+1}^{(k)} \\ & & & \vdots & & \vdots & \vdots \\ & & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} & a_{n,n+1}^{(k)} \end{pmatrix}. \tag{5.2}$$

假设 $a_{kk}^{(k)}\neq 0$，则保留第 1 个至第 k 个方程不变，利用第 k 个方程分别消去其余方程第 k 个未知量. 这时，第 $k+1$ 个至第 n 个方程的未知量的系数和常数项一般都有改变. 令

$$l_{ik}=\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}},\quad a_{ij}^{(k+1)}=a_{ij}^{(k)}-l_{ik}a_{kj}^{(k)}, \tag{5.3}$$

$$i=k+1,k+2,\cdots,n;\quad j=k+1,k+2,\cdots,n+1,$$

则将 $(\boldsymbol{A}^{(k)},\boldsymbol{b}^{(k)})$ 变换为 $(\boldsymbol{A}^{(k+1)},\boldsymbol{b}^{(k+1)})$，$(\boldsymbol{A}^{(k+1)},\boldsymbol{b}^{(k+1)})$ 中第 1 行至第 k 行的元素与(5.2)式中的对应元素相同，第 $k+1$ 行至第 n 行的元素由(5.3)式给出. 它对应的方程组 $\boldsymbol{A}^{(k+1)}\boldsymbol{x}=\boldsymbol{b}^{k+1}$ 与(5.1)式等价，而在第 $k+1$ 个至第 n 个方程中，含 $x_1,x_2,\cdots,x_k$ 的项已经消去.

上述过程可进行 $n-1$ 步，得到方程组

$$\boldsymbol{A}^{(n)}\boldsymbol{x}=\boldsymbol{b}^{(n)}, \tag{5.4}$$

其中 $\boldsymbol{A}^{(n)}$ 是一个上三角阵，

$$(\boldsymbol{A}^{(n)},\boldsymbol{b}^{(n)})=\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} & a_{1,n+1}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & a_{2,n+1}^{(2)} \\ & & \ddots & \vdots & \vdots \\ & & & a_{nn}^{(n)} & a_{n,n+1}^{(n)} \end{pmatrix}.$$

这就完成了消元过程.

因为 $\boldsymbol{A}$ 非奇异，所以 $a_{nn}^{(n)}\neq 0$，可求解上三角方程组(5.4)，通过逐次代入计算可得方程组的解，其计算公式为

$$\begin{cases} x_n = \dfrac{a_{n,n+1}^{(n)}}{a_{nn}^{(n)}}, \\ x_i = \dfrac{1}{a_{ii}^{(i)}}\left(a_{i,n+1}^{(i)} - \sum\limits_{j=i+1}^{n} a_{ij}^{(i)} x_j\right), \quad i = n-1, n-2, \cdots, 1. \end{cases} \tag{5.5}$$

求解上式的过程称为回代过程.

以上由消去过程和回代过程合起来求解(5.1)式的过程就称为 Gauss 消去法，或称为顺序 Gauss 消去法.

由(5.3)式可知，消元过程的第 k 步需除法运算 $n-k$ 次，乘法和减法运算各需 $(n-k)(n+1-k)$ 次，所以消元过程共需乘除法运算的次数为

$$\sum_{k=1}^{n-1}(n-k) + \sum_{k=1}^{n-1}(n-k)(n+1-k) = \frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6},$$

需加减法运算的次数

$$\sum_{k=1}^{n-1}(n-k)(n+1-k) = \frac{n^3}{3} - \frac{n}{3}.$$

回代过程需乘除法运算 $\dfrac{n(n+1)}{2}$ 次，加减法运算 $\dfrac{n(n-1)}{2}$ 次. 所以，Gauss 消去法总共需乘除运算的次数为

$$\frac{n^3}{3} + n^2 - \frac{n}{3} \approx \frac{n^3}{3},$$

需加减法运算的次数为

$$\frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6} \approx \frac{n^3}{3}.$$

如果我们用 Cramer 法则计算(5.1)式的解，要计算 $n+1$ 阶行列式，并作 n 次除法. 如果用子式展开的方法计算行列式，则计算每个行列式有 $n!$ 次乘法. 所以用 Cramer 法则大约需要 $(n+1)!$ 次乘除法运算. 例如，当 $n=10$ 时，约需 4×10^7 次乘除法运算，而用 Gauss 消去法只需 430 次乘除法运算.

例 5.1 用 Gauss 消去法解方程组

$$\begin{cases} x_1 + \dfrac{2}{3}x_2 + \dfrac{1}{3}x_3 = 2, \\ \dfrac{9}{20}x_1 + x_2 + \dfrac{11}{20}x_3 = 2, \\ \dfrac{2}{3}x_1 + \dfrac{1}{3}x_2 + x_3 = 2. \end{cases}$$

解 第一步消元，令 $l_{21} = \dfrac{9}{20}$，$l_{31} = \dfrac{2}{3}$，得增广矩阵

$$\begin{pmatrix} 1 & \frac{2}{3} & \frac{1}{3} & 2 \\ 0 & \frac{7}{10} & \frac{2}{5} & \frac{11}{10} \\ 0 & -\frac{1}{9} & \frac{7}{9} & \frac{2}{3} \end{pmatrix}.$$

第二步消元，令 $l_{32}=-\frac{10}{63}$，得增广矩阵

$$\begin{pmatrix} 1 & \frac{2}{3} & \frac{1}{3} & 2 \\ 0 & \frac{7}{10} & \frac{2}{5} & \frac{11}{10} \\ 0 & 0 & \frac{53}{63} & \frac{53}{63} \end{pmatrix}$$

利用回代公式(5.5)依次得到 $x_3=1, x_2=1, x_1=1$.

在这个例子中，我们写出的是分数运算的结果. 如果在计算机上进行计算，系数矩阵和中间结果都用经过舍入的机器数表示，中间结果和方程组的解都会有误差.

5.1.2　矩阵的三角分解

从上面的消元过程看出，消元过程能顺序进行的重要条件是主元素 $a_{kk}^{(k)}\neq 0, k=1,2,\cdots,n-1$. 若用 A_k 表示矩阵 A 的 k 阶顺序主子阵，则有下面的定理.

定理 5.1　$a_{ii}^{(i)}(i=1,2,\cdots,k)$全不为零的充分必要条件是 A 的顺序主子式 $D_i=\det A_i\neq 0, i=1,2,\cdots,k$，其中 $k\leqslant n$.

证　先证必要性. 设 $a_{ii}^{(i)}\neq 0, i=1,2,\cdots,k$，则可进行 k 步消元过程. 显然 $a_{11}^{(1)}=D_1\neq 0$. 对 $i\geqslant 2$，由于每步进行的初等变换不改变顺序主子式的值，所以第 i 步消元后有

$$D_i=\begin{vmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1i}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2i}^{(2)} \\ & & \ddots & \vdots \\ & & & a_{ii}^{(i)} \end{vmatrix}=a_{11}^{(1)}a_{22}^{(2)}\cdots a_{ii}^{(i)}\neq 0.$$

用归纳法证充分性. $k=1$ 时，命题显然成立. 设命题对 $m-1$ 对立. 现设 $D_i\neq 0$，$i=1,2,\cdots,m$. 由归纳假设有 $a_{ii}^{(i)}\neq 0, i=1,2,\cdots,m-1$，Gauss 消去法可进行第 $m-1$ 步，矩阵 $\boldsymbol{A}$ 变换为

$$\boldsymbol{A}^{(m)}=\begin{pmatrix}\boldsymbol{A}_{11}^{(m)} & \boldsymbol{A}_{12}^{(m)}\\ 0 & \boldsymbol{A}_{22}^{(m)}\end{pmatrix},$$

其中 $\boldsymbol{A}_{11}^{(m)}$ 是对角元为 $a_{11}^{(1)},a_{22}^{(2)},\cdots,a_{m-1,m-1}^{(m-1)}$ 的上三角阵. 因 $\boldsymbol{A}^{(m)}$ 是通过消元过程由 $\boldsymbol{A}$ 逐步经初等变换得到的,$\boldsymbol{A}$ 的 m 阶顺序主子式等于 $\boldsymbol{A}^{(m)}$ 的 m 阶顺序主子式,即

$$D_m=a_{11}^{(1)}a_{22}^{(2)}\cdots a_{m-1,m-1}^{(m-1)}a_{mm}^{(m)}.$$

由 $D_m\neq 0$ 可推出 $a_{mm}^{(m)}\neq 0$,定理得证.

定理 5.2 在方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 中,若 $\boldsymbol{A}$ 非奇异,则当 $\boldsymbol{A}$ 的所有顺序主子式均不为零时,可用 Gauss 消去法求出方程组的解.

特别地,若 $\boldsymbol{A}$ 为对称正定矩阵,则由对称正定矩阵的性质可知,对原方程组不必作任何处理,可直接用 Gauss 消去法求解方程组.

下面将消元过程用矩阵运算表示. 对第 k 步,利用(5.3)式给出的乘数 l_{ik},记 $\boldsymbol{l}^{(k)}=(0,\cdots,0,l_{k+1,k},\cdots,l_{nk})^{\mathrm{T}}$,又记 $\boldsymbol{e}_k=(0,\cdots,0,1,0,\cdots,0)^{\mathrm{T}}$ 为第 k 个分量为 1 的单位向量,令

$$L_k=I-\boldsymbol{l}^{(k)}\boldsymbol{e}_k^{\mathrm{T}}=\begin{pmatrix}1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -l_{k+1,k} & 1 & & \\ & & \vdots & & \ddots & \\ & & -l_{nk} & & & 1\end{pmatrix} \tag{5.6}$$

不难验证

$$(\boldsymbol{I}-\boldsymbol{l}^{(k)}\boldsymbol{e}_k^{\mathrm{T}})(\boldsymbol{I}+\boldsymbol{l}^{(k)}\boldsymbol{e}_k^{\mathrm{T}})=(\boldsymbol{I}+\boldsymbol{l}^{(k)}\boldsymbol{e}_k^{\mathrm{T}})(\boldsymbol{I}-\boldsymbol{l}^{(k)}\boldsymbol{e}_k^{\mathrm{T}})=\boldsymbol{I},$$

即 $\boldsymbol{L}_k^{-1}=\boldsymbol{I}+\boldsymbol{l}^{(k)}\boldsymbol{e}_k^{\mathrm{T}}$.

利用矩阵(5.6),第 k 步消元过程相当于

$$\boldsymbol{L}_k(\boldsymbol{A}^{(k)},\boldsymbol{b}^{(k)})=(\boldsymbol{A}^{(k+1)},\boldsymbol{b}^{(k+1)}).$$

这样,经过 $n-1$ 步消元过程得到

$$\boldsymbol{L}_{n-1}\boldsymbol{L}_{n-2}\cdots\boldsymbol{L}_1\boldsymbol{A}^{(1)}=\boldsymbol{A}^{(n)},$$

$$\boldsymbol{L}_{n-1}\boldsymbol{L}_{n-2}\cdots\boldsymbol{L}_1\boldsymbol{b}^{(1)}=\boldsymbol{b}^{(n)},$$

这里,$\boldsymbol{A}^{(n)}$ 是上三角阵. 记 $\boldsymbol{U}=\boldsymbol{A}^{(n)}$,又记 $\boldsymbol{L}=\boldsymbol{L}_1^{-1}\boldsymbol{L}_2^{-1}\cdots\boldsymbol{L}_{n-1}^{-1}$,则有

$$\boldsymbol{L}=\begin{pmatrix}1 & & & & \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{n,n-1} & 1\end{pmatrix},$$

这种矩阵称为单位下三角阵. L 的对角线以下各元素就是各步消元过程的乘数. 最后我们得到

$$\boldsymbol{A}=\boldsymbol{L}\boldsymbol{U}, \tag{5.7}$$

称该式为 $\boldsymbol{A}$ 的 LU 分解.

定理 5.3　矩阵 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$，若其顺序主子式 $D_i(i=1,2,\cdots,n)$ 皆非零，则存在唯一的单位下三角阵 $\boldsymbol{L}$ 和上三角阵 $\boldsymbol{U}$，使 $\boldsymbol{A}=\boldsymbol{L}\boldsymbol{U}$.

证　以上的分析已证明了 $\boldsymbol{A}$ 可作 LU 分解，下面证明分解的唯一性. 设 $\boldsymbol{A}$ 有两个分解式

$$\boldsymbol{A}=\boldsymbol{L}\boldsymbol{U}=\tilde{\boldsymbol{L}}\tilde{\boldsymbol{U}}$$

其中，$\boldsymbol{L},\tilde{\boldsymbol{L}}$ 都是单位下三角阵，$\boldsymbol{U},\tilde{\boldsymbol{U}}$ 都是上三角阵. 因 $\boldsymbol{A}$ 非奇异，则 $\boldsymbol{L},\tilde{\boldsymbol{L}},\boldsymbol{U},\tilde{\boldsymbol{U}}$ 都可逆. $\boldsymbol{A}$ 左乘 $\boldsymbol{L}^{-1}$，右乘 $\tilde{\boldsymbol{U}}^{-1}$ 即得

$$\boldsymbol{U}\tilde{\boldsymbol{U}}^{-1}=\boldsymbol{L}^{-1}\tilde{\boldsymbol{L}}.$$

因 $\tilde{\boldsymbol{U}}^{-1}$ 仍为上三角阵，$\boldsymbol{U}\tilde{\boldsymbol{U}}^{-1}$ 也是上三角阵，同理 $\boldsymbol{L}^{-1}\tilde{\boldsymbol{L}}$ 是单位下三角阵，所以只能有

$$\boldsymbol{U}\tilde{\boldsymbol{U}}^{-1}=\boldsymbol{L}^{-1}\tilde{\boldsymbol{L}}=\boldsymbol{I},$$

即 $\boldsymbol{U}=\tilde{\boldsymbol{U}},\boldsymbol{L}=\tilde{\boldsymbol{L}}$. 定理得证.

分解式(5.7)也称为 Doolittle 分解. 由(5.7)式可求出 $\boldsymbol{A}$ 的行列式，即

$$\det\boldsymbol{A}=a_{11}^{(1)}a_{22}^{(2)}\cdots a_{nn}^{(n)}.$$

若将上三角阵 $\boldsymbol{U}$ 写成 $\boldsymbol{U}=\boldsymbol{D}\bar{\boldsymbol{U}}$，其中 $\bar{\boldsymbol{U}}$ 是单位上三角阵，则有

$$\boldsymbol{A}=\boldsymbol{L}\boldsymbol{D}\bar{\boldsymbol{U}}, \tag{5.8}$$

称该式为 $\boldsymbol{A}$ 的 $\mathrm{LD}\bar{\mathrm{U}}$ 分解. 显然，这种分解式具有唯一性.

例 5.2　求矩阵

$$\boldsymbol{A}=\begin{pmatrix}1 & 1 & 1\\ 0 & 4 & -1\\ 2 & -2 & -1\end{pmatrix}$$

的 LU 分解和 $\mathrm{LD}\bar{\mathrm{U}}$ 分解.

解　由 Gauss 消去法 $l_{21}=0,l_{31}=2$，

$$\boldsymbol{A}=\boldsymbol{A}^{(1)}\to\boldsymbol{A}^{(2)}=\begin{pmatrix}1 & 1 & 1\\ 0 & 4 & -1\\ 0 & -4 & -3\end{pmatrix}.$$

进一步，有 $l_{32}=-1$，

$$\boldsymbol{A}^{(2)}\to\boldsymbol{A}^{(3)}=\begin{pmatrix}1 & 1 & 1\\ 0 & 4 & -1\\ 0 & 0 & -4\end{pmatrix}.$$

所以，$\boldsymbol{A}=\boldsymbol{LU}$，其中

$$\boldsymbol{L}=\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 2 & -1 & 1\end{pmatrix},\quad \boldsymbol{U}=\boldsymbol{A}^{(3)}=\begin{pmatrix}1 & 1 & 1\\ 0 & 4 & -1\\ 0 & 0 & -4\end{pmatrix}.$$

又因为

$$\boldsymbol{U}=\begin{pmatrix}1 & 0 & 0\\ 0 & 4 & 0\\ 0 & 0 & -4\end{pmatrix}\begin{pmatrix}1 & 1 & 1\\ 0 & 1 & -\frac{1}{4}\\ 0 & 0 & 1\end{pmatrix},$$

所以有

$$\boldsymbol{A}=\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 2 & -1 & 1\end{pmatrix}\begin{pmatrix}1 & 0 & 0\\ 0 & 4 & 0\\ 0 & 0 & -4\end{pmatrix}\begin{pmatrix}1 & 1 & 1\\ 0 & 1 & -\frac{1}{4}\\ 0 & 0 & 1\end{pmatrix}=\boldsymbol{LD}\bar{\boldsymbol{U}}$$

5.1.3 主元素消去法

在以上的 Gauss 消去法中，消元过程能进行的条件是主元素 $a_{ii}^{(i)}\neq 0$，$i=1,2,\cdots,n-1$. 例如，若 $a_{11}=0$，消去过程的第 1 步就不能进行. 有时虽然 $a_{ii}^{(i)}\neq 0$，但是 $|a_{ii}^{(i)}|$ 很小，这时计算过程的舍入误差会导致消去法数值不稳定，以致结果不可靠.

例 5.3 用 3 位十进制浮点运算求解

$$\begin{cases}1.00\times 10^{-5}x_1+1.00x_2=1.00,\\ 1.00x_1+1.00x_2=2.00.\end{cases}$$

解 这个方程组的准确解显然应接近 $(1.00,1.00)^{\mathrm{T}}$. 但是系数 a_{11} 是个小主元，如果用 Gauss 消去法求解，则有

$$\begin{aligned}&l_{21}=\frac{a_{21}}{a_{11}}\times 1.00\times 10^{5},\\ &a_{22}^{(2)}=a_{22}-l_{21}a_{12}=1.00-1.00\times 10^{5},\\ &a_{23}^{(2)}=a_{23}-l_{21}a_{13}=2.00-1.00\times 10^{5}.\end{aligned}$$

在 3 位十进制运算的限制下，得到 $x_2=\dfrac{a_{23}^{(2)}}{a_{22}^{(2)}}=1.00$. 代回第 1 个方程得 $x_1=0$，这显然是不正确的解. 因为用小主元 a_{11} 做除法，使乘数 l_{21} 是个大数，在 $a_{22}^{(2)}$ 的计算中，a_{22} 的值完全被掩盖了.

如果先把两个方程的次序交换，再用 Gauss 消去法，就不会出现上述问题，解得 $x_1=1.00$，$x_2=1.00$. 这就是列主元素消去法的思想.

列主元素消去法也称按列部分选主元的消去法. 一般地，在完成了第 $k-1$ 步

消元运算后，在$(\boldsymbol{A}^{(k)},\boldsymbol{b}^{(k)})$的第 k 列元素 $a_{kk}^{(k)}$ 之下的所有元素中选一个绝对值最大的元素作为主元素，即若

$$|a_{i_k,k}^{(k)}|=\max_{k\leqslant i\leqslant n}|a_{ik}^{(k)}|,$$

则以 $a_{i_k,k}^{(k)}$ 为主元素，这里 $i_k\geqslant k$，且由于 $\boldsymbol{A}^{(k)}$ 非奇异，有 $a_{i_k,k}^{(k)}\neq 0$. 这样，有 $|l_{ik}|=\dfrac{|a_{ik}^{(k)}|}{|a_{i_k,k}^{(k)}|}\leqslant 1$，达到控制舍入误差的作用.

选出主元素后，若 $i_k=k$，则进行顺序 Gauss 消去法的第 k 步. 若 $i_k>k$，则将 $(\boldsymbol{A}^{(k)},\boldsymbol{b}^{(k)})$的第 i_k 行与第 k 行交换，然后进行消元运算.

完成了 $n-1$ 步选主元、换行与消元运算后，得到 $\boldsymbol{A}^{(n)}\boldsymbol{x}=\boldsymbol{b}^{(n)}$，这是与原方程组等价的方程组，$\boldsymbol{A}^{(n)}$ 是一个上三角阵，再回代求解. 这就是列主元素消去法的计算过程.

除了列主元素消去法外，还有一种完全主元素消去法. 在其过程的第 k 步($k\geqslant 1$)，不是按列来选主元，而是在 $\boldsymbol{A}^{(k)}$ 右下角的 $n-k+1$ 阶子阵中选主元 $a_{i_k,j_k}^{(k)}$，即

$$|a_{i_k,j_k}^{(k)}|=\max_{k\leqslant i,j\leqslant n}|a_{ij}^{(k)}|.$$

然后将$(\boldsymbol{A}^{(k)},\boldsymbol{b}^{(k)})$的第 i_k 行与第 k 行交换，将第 j_k 列与第 k 列交换，同时将自变量 x_k 与 x_{j_k} 的位置交换并记录自变量的排列次序. 直到消去法完成后，再按记录恢复自变量为自然次序. 完全主元法比列主元法运算量大得多，由于列主元法的舍入误差一般已较小，所以在实际计算中多用列主元法.

例 5.4　用列主元素消去法解方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$，计算过程中取 5 位有效数字进行运算，其中

$$(\boldsymbol{A},\boldsymbol{b})=\begin{pmatrix}-0.002 & 2 & 2 & 0.4\\ 1 & 0.781\,25 & 0 & 1.381\,6\\ 3.996 & 5.562\,5 & 4 & 7.417\,8\end{pmatrix}.$$

解　记$(\boldsymbol{A}^{(1)},\boldsymbol{b}^{(1)})=(\boldsymbol{A},\boldsymbol{b})$. 第一步选列主元为 $a_{31}^{(1)}=3.996$. 交换第 1 行与第 3 行，再消元计算得

$$(\boldsymbol{A}^{(2)},\boldsymbol{b}^{(2)})=\begin{pmatrix}3.996 & 5.562\,5 & 4 & 7.417\,8\\ 0 & -0.610\,77 & -1.001\,0 & -0.474\,71\\ 0 & 2.002\,8 & 2.002\,0 & 0.403\,71\end{pmatrix}.$$

第二步选列主元为 $a_{32}^{(2)}=2.0028$. 交换第 2 行与第 3 行，再消元计算得

$$(\boldsymbol{A}^{(3)},\boldsymbol{b}^{(3)})=\begin{pmatrix}3.996 & 5.562\,5 & 4 & 7.417\,8\\ 0 & 2.002\,8 & 2.002\,0 & 0.403\,71\\ 0 & 0 & -0.390\,47 & -0.351\,59\end{pmatrix}.$$

消去过程至此结束. 回代计算依次得到解

$$x_3=0.900\,43,\quad x_2=-0.698\,50,\quad x_1=1.927\,3.$$

这个例题的精确解是

$$\boldsymbol{x}=(1.927\,30,-0.698\,496,0.900\,423)^{\mathrm{T}},$$

而用不选主元的顺序 Gauss 消去法，则解得

$$\boldsymbol{x}=(1.930\,0,-0.686\,95,0.888\,88)^{\mathrm{T}}.$$

这个结果误差较大，这是因为消去法的第 1 步中，$a_{11}^{(1)}$ 按绝对值比其他元素小很多所引起的. 从此例看到主元素消去法是有效的方法.

下面讨论矩阵的含换行变换的三角分解，即列主元法中消去过程的矩阵表示. 一般地，将矩阵 $\boldsymbol{A}$ 的第 i 行与第 j 行交换，其结果相当于矩阵 $\boldsymbol{A}$ 左乘一个初等排列矩阵 $\boldsymbol{I}_{ij}$，即 $\boldsymbol{I}_{ij}\boldsymbol{A}$，这里 $\boldsymbol{I}_{ij}$ 是单位阵 $\boldsymbol{I}$ 交换第 i 行与第 j 行后所得的矩阵，不难验证

$$\boldsymbol{I}_{ij}=\boldsymbol{I}_{ji},\quad \boldsymbol{I}_{ij}^{-1}=\boldsymbol{I}_{ij},\quad \det\boldsymbol{I}_{ij}=-1.$$

若矩阵 $\boldsymbol{A}$ 右乘 $\boldsymbol{I}_{ij}$ 得 $\boldsymbol{A}\boldsymbol{I}_{ij}$，其结果是将 $\boldsymbol{A}$ 的第 i 列与第 j 列交换后所得的矩阵.

我们把若干个初等排列矩阵的乘积称作排列矩阵，其结果是将单位矩阵经过若干次行交换所得的矩阵.

列主元素消去法的每一步，一般是先按列选主元再交换行，然后进行消元计算，所以有

$$\boldsymbol{A}^{(k+1)}=\boldsymbol{L}_k\boldsymbol{I}_{k,i_k}\boldsymbol{A}^{(k)},$$

其中 $\boldsymbol{L}_k$ 为(5.6)式所示，$\boldsymbol{I}_{k,i_k}$ 是初等排列阵，i_k 是第 k 步列选主元所在的行号. 如果第 k 步不需换行，则 $i_k=k$，$\boldsymbol{I}_{kk}=\boldsymbol{I}$.

列主元素消去法的消元过程进行 $n-1$ 步之后，得到上三角阵 $\boldsymbol{A}^{(n)}$，记

$$\boldsymbol{U}=\boldsymbol{A}^{(n)}=\boldsymbol{L}_{n-1}\boldsymbol{I}_{n-1,i_{n-1}}\cdots\boldsymbol{L}_2\boldsymbol{I}_{2,i_2}\boldsymbol{L}_1\boldsymbol{I}_{1,i_1}\boldsymbol{A}. \tag{5.9}$$

这就是列主元法消去过程的矩阵表示. 由于列主元的选取，我们可知 $\boldsymbol{L}_k$ 及 $\boldsymbol{L}_k^{-1}$ 的元素的绝对值不大于 1.

定理 5.4 设 $\boldsymbol{A}$ 为非奇异矩阵，则存在排列阵 $\boldsymbol{P}$，单位下三角矩阵 $\boldsymbol{L}$ 和上三角阵 $\boldsymbol{U}$，使 $\boldsymbol{PA}=\boldsymbol{LU}$.

证　从(5.9)式可得

$$\boldsymbol{A}=\boldsymbol{I}_{1,i_1}\boldsymbol{L}_1^{-1}\boldsymbol{I}_{2,i_2}\boldsymbol{L}_2^{-1}\cdots\boldsymbol{I}_{n-1,i_{n-1}}\boldsymbol{L}_{n-1}^{-1}\boldsymbol{U},$$

其中 $\boldsymbol{U}$ 为上三角阵. 令排列阵

$$\boldsymbol{P}=\boldsymbol{I}_{n-1,i_{n-1}}\cdots\boldsymbol{I}_{2,i_2}\boldsymbol{I}_{1,i_1},$$

则利用 $\boldsymbol{I}_{ij}^{-1}=\boldsymbol{I}_{ij}$ 有

$$\begin{aligned}\boldsymbol{PA}&=(\boldsymbol{I}_{n-1,i_{n-1}}\cdots\boldsymbol{I}_{2,i_2}\boldsymbol{L}_1^{-1}\boldsymbol{I}_{2,i_2}\cdots\boldsymbol{I}_{n-1,i_{n-1}})\\&\quad\cdot\boldsymbol{I}_{n-1,i_{n-1}}\cdots\boldsymbol{I}_{3,i_3}\boldsymbol{L}_2^{-1}\boldsymbol{I}_{3,i_3}\cdots\boldsymbol{I}_{n-1,i_{n-1}}\boldsymbol{L}_{n-1}^{-1}\boldsymbol{U}\\&=(\boldsymbol{I}_{n-1,i_{n-1}}\cdots\boldsymbol{I}_{2,i_2}\boldsymbol{L}_1^{-1}\boldsymbol{I}_{2,i_2}\cdots\boldsymbol{I}_{n-1,i_{n-1}})\\&\quad\cdot(\boldsymbol{I}_{n-1,i_{n-1}}\cdots\boldsymbol{I}_{3,i_3}\boldsymbol{L}_2^{-1}\boldsymbol{I}_{3,i_3}\cdots\boldsymbol{I}_{n-1,i_{n-1}})\cdots(\boldsymbol{I}_{n-1,i_{n-1}}\boldsymbol{L}_{n-2}^{-1}\boldsymbol{I}_{n-1,i_{n-1}})\boldsymbol{L}_{n-1}^{-1}\boldsymbol{U}.\end{aligned}$$

由此，若记

$$\tilde{\boldsymbol{L}}_k=\boldsymbol{I}_{n-1,i_{n-1}}\cdots\boldsymbol{I}_{k+1,i_{k+1}}\boldsymbol{L}_k^{-1}I_{k+1,i_{k+1}}\cdots\boldsymbol{I}_{n-1,i_{n-1}},\quad k=1,2,\cdots,n-2,$$

$$\tilde{\boldsymbol{L}}_{n-1}=\boldsymbol{L}_{n-1}^{-1},$$

$$\boldsymbol{L}=\tilde{\boldsymbol{L}}_1\tilde{\boldsymbol{L}}_2\cdots\tilde{\boldsymbol{L}}_{n-1},$$

则得 $\boldsymbol{PA}=\boldsymbol{LU}$. 由初等排列阵的性质 $\tilde{\boldsymbol{L}}_k$ 是一个单位下三角阵，$\boldsymbol{L}$ 也是一个单位下三角阵. 定理得证.

例 **5.5**　给定矩阵

$$\boldsymbol{A}=\begin{pmatrix}0&0&1&2\\0&0&3&0\\1&-1&0&1\\2&0&-1&3\end{pmatrix},$$

求排列阵 $\boldsymbol{P}$，使 $\boldsymbol{PA}=\boldsymbol{LU}$.

解　取 $\boldsymbol{I}_{14}$ 和 $\boldsymbol{I}_{23}$，使

$$\boldsymbol{I}_{23}\boldsymbol{I}_{14}\boldsymbol{A}=\begin{pmatrix}2&0&-1&3\\1&-1&0&1\\0&0&3&0\\0&0&1&2\end{pmatrix}.$$

上式右端各阶主子行列式都不为 0，作 $\boldsymbol{LU}$ 分解，

$$\boldsymbol{I}_{23}\boldsymbol{I}_{14}\boldsymbol{A}=\begin{pmatrix}1&&&\\\frac{1}{2}&1&&\\0&0&1&\\0&0&\frac{1}{3}&1\end{pmatrix}\begin{pmatrix}2&0&-1&3\\&-1&\frac{1}{2}&-\frac{1}{2}\\&&3&0\\&&&2\end{pmatrix}=\boldsymbol{LU}.$$

于是有 $\boldsymbol{PA}=\boldsymbol{LU}$，其中

$$\boldsymbol{P}=\boldsymbol{I}_{23}\boldsymbol{I}_{14}=\begin{pmatrix}1&0&0&0\\0&0&1&0\\0&1&0&0\\0&0&0&1\end{pmatrix}\begin{pmatrix}0&0&0&1\\0&1&0&0\\0&0&1&0\\1&0&0&0\end{pmatrix}=\begin{pmatrix}0&0&0&1\\0&0&1&0\\0&1&0&0\\1&0&0&0\end{pmatrix}.$$

当然，在实际计算时，排列阵 $\boldsymbol{P}$ 不是像例 5.5 一样立刻就能找到的，而是像定理 5.4 的证明一样，要通过求若干个初等排列矩阵与求三角分解穿插进行.

5.1.4　Gauss-Jordan 消去法

考虑 Gauss 消去法的一种修正：消去对角线下方和上方的元素. 称这种方法为

Gauss-Jordan 消去法. 设用 Gauss-Jordan 消去法已完成 $k-1$ 步，得到与方程 $\boldsymbol{Ax}=\boldsymbol{b}$ 等价的方程组 $\boldsymbol{A}^{(k)}\boldsymbol{x}=\boldsymbol{b}^{(k)}$，此时对应的增广矩阵是

$$(\boldsymbol{A}^{(k)},\boldsymbol{b}^{(k)})=\begin{pmatrix} 1 & & & a_{1k} & \cdots & a_{1n} & a_{1,n+1} \\ & \ddots & & \vdots & & \vdots & \vdots \\ & & 1 & a_{k-1,k} & \cdots & a_{k-1,n} & a_{k-1,n+1} \\ & & & a_{kk} & \cdots & a_{kn} & a_{k,n+1} \\ & & & \vdots & & \vdots & \vdots \\ & & & a_{nk} & \cdots & a_{nn} & a_{n,n+1} \end{pmatrix},$$

这里，略去了矩阵元素的上标. 在第 k 步计算时，考虑对上述矩阵第 k 列中的第 k 行上、下都进行消元计算. 若用列主元素消去法，仍然是在第 k 列元素 a_{kk} 之下的所有元素中选一个绝对值最大的元素作为主元素，即

$$|a_{i_k,k}|=\max_{k\leqslant i\leqslant n}|a_{ik}|.$$

但是，将第 k 行与第 i_k 行交换后，要通过主行将第 k 列的第 $i(i=1,\cdots,k-1,k+1,\cdots,n)$ 个元素化为零，再将主行的对角线上元素化为 1. 最后得 $(\boldsymbol{A}^{(n+1)},\boldsymbol{b}^{(n+1)})$，这里，$\boldsymbol{A}^{(n+1)}$ 是单位矩阵，$\boldsymbol{b}^{(n+1)}$ 就是计算解.

可见，Gauss-Jordan 消去法用不着回代求解，其计算量大约需要 $\frac{n^3}{2}$ 次乘除法运算，要比 Gauss 消去法的计算量大，但用 Gauss-Jordan 消去法求一个矩阵的逆矩阵是很合适的.

定理 5.5 设 $\boldsymbol{A}$ 为 n 阶非奇异矩阵，方程组 $\boldsymbol{AX}=\boldsymbol{I}$ 的增广矩阵为 $\boldsymbol{C}=(\boldsymbol{A},\boldsymbol{I})$. 如果对 $\boldsymbol{C}$ 用 Gauss-Jordan 消去法化为 $(\boldsymbol{I},\boldsymbol{T})$，则 $\boldsymbol{T}=\boldsymbol{A}^{-1}$.

证 设 $\boldsymbol{A}^{-1}=\boldsymbol{X}=(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_n)$，则 $\boldsymbol{AX}=\boldsymbol{I},\boldsymbol{A\alpha}_j=\boldsymbol{e}_j,j=1,2,\cdots,n$. 这里，$\boldsymbol{e}_j$ 为单位矩阵 $\boldsymbol{I}$ 的第 j 列. 用 Gauss-Jordan 消去法解方程组 $\boldsymbol{Ax}_j=\boldsymbol{e}_j$，其解在 $\boldsymbol{C}$ 中 $\boldsymbol{I}$ 的第 j 列，即为 $\boldsymbol{T}$ 的第 j 列，即 $\boldsymbol{\alpha}_j=\boldsymbol{Te}_j$. 因此，$\boldsymbol{T}=(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_n)=X=A^{-1}$. 定理得证.

例 5.6 用 Gauss-Jordan 消去法求矩阵

$$\boldsymbol{A}=\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{pmatrix}$$

的逆矩阵.

解 用列主元素消去法有

$$\boldsymbol{C}=\boldsymbol{C}^{(1)}=(\boldsymbol{A},\boldsymbol{I})=\begin{pmatrix} 1 & 2 & 3 & 1 & 0 & 0 \\ 2 & 4 & 5 & 0 & 1 & 0 \\ 3 & 5 & 6 & 0 & 0 & 1 \end{pmatrix},$$

$$\boldsymbol{C}^{(2)}=\begin{pmatrix}1 & \frac{5}{3} & 0 & 0 & 0 & \frac{1}{3}\\ 0 & \frac{2}{3} & 1 & 0 & 1 & -\frac{2}{3}\\ 0 & \frac{1}{3} & 1 & 1 & 0 & -\frac{1}{3}\end{pmatrix},$$

$$\boldsymbol{C}^{(3)}=\begin{pmatrix}1 & 0 & -\frac{1}{2} & 0 & -\frac{5}{2} & 2\\ 0 & 1 & \frac{3}{2} & 0 & \frac{3}{2} & -1\\ 0 & 0 & \frac{1}{2} & 1 & -\frac{1}{2} & 0\end{pmatrix},$$

$$\boldsymbol{C}^{(4)}=\begin{pmatrix}1 & 0 & 0 & 1 & -3 & 2\\ 0 & 1 & 0 & -3 & 3 & -1\\ 0 & 0 & 1 & 2 & -1 & 0\end{pmatrix}=(\boldsymbol{I},\boldsymbol{A}^{-1}).$$

在实际计算中，为了节省内存单元，单位矩阵不必存放. 在上例中，可将 $\boldsymbol{C}^{(2)}$ 的最后一列存放在 $\boldsymbol{A}$ 的第 1 列，将 $\boldsymbol{C}^{(3)}$ 的第 5 列存放在 $\boldsymbol{A}$ 的第 2 列，将 $\boldsymbol{C}^{(4)}$ 的第 4 列存放在 $\boldsymbol{A}$ 的第 3 列. 一般地，第 k 步消元时，可将 $\boldsymbol{A}$ 的第 k 列

$$\boldsymbol{a}_k=(a_{1k},\cdots,a_{kk},\cdots,a_{nk})^{\mathrm{T}}$$

用向量

$$\boldsymbol{l}_k=\left(-\frac{a_{1k}}{a_{kk}},\cdots,-\frac{a_{k-1,k}}{a_{kk}},\frac{1}{a_{kk}},-\frac{a_{k+1,k}}{a_{kk}},\cdots,-\frac{a_{nk}}{a_{kk}}\right)^{\mathrm{T}}$$

取代. 最后再调整一下列的次序就可以在 $\boldsymbol{A}$ 的位置得到 $\boldsymbol{A}^{-1}$. 事实上，在 $\boldsymbol{A}$ 的位置最后得到的矩阵是 $\boldsymbol{PA}=\widetilde{\boldsymbol{A}}$ 的逆矩阵 $\widetilde{\boldsymbol{A}}^{-1}$，其中 $\boldsymbol{P}$ 为行交换形成的排列阵，于是 $\boldsymbol{A}^{-1}=\widetilde{\boldsymbol{A}}^{-1}\boldsymbol{P}$.

5.2　直接三角分解方法

5.2.1　一般矩阵的直接三角分解法

本节讨论矩阵 $\boldsymbol{A}$ 的三角分解的直接计算以及直接利用 $\boldsymbol{A}$ 的三角分解式来求解方程组.

(1) 不选主元的三角分解法.

设 $\boldsymbol{A}=\boldsymbol{LU}$，记 $\boldsymbol{A}=(a_{ij})$，$\boldsymbol{L}=(l_{ij})$，$\boldsymbol{U}=(u_{ij})$，其中 $\boldsymbol{L}$ 为单位下三角阵，$\boldsymbol{U}$ 为上三角阵. 我们可直接给出 $\boldsymbol{L}$ 和 $\boldsymbol{U}$ 的元素的计算公式.

由 $\boldsymbol{A}$ 的第 1 行和第 1 列可计算出 $\boldsymbol{U}$ 的第 1 行和 $\boldsymbol{L}$ 的第 1 列，即

$$u_{1j}=a_{1j},\quad j=1,2,\cdots,n, \tag{5.10}$$

$$l_{k1}=\frac{a_{k1}}{u_{11}},\quad k=2,3,\cdots,n. \tag{5.11}$$

如果 $\boldsymbol{U}$ 的第 1 行至第 $k-1$ 行和 $\boldsymbol{L}$ 的第 1 列至第 $k-1$ 列已经算出，则由

$$a_{kj}=\sum_{r=1}^{k}l_{kr}u_{rj},\quad j=k,k+1,\cdots,n,$$

可得 $\boldsymbol{U}$ 的第 k 行元素

$$u_{kj}=a_{kj}-\sum_{r=1}^{k-1}l_{kr}u_{rj},\quad j=k,k+1,\cdots,n. \tag{5.12}$$

同理，由

$$a_{ik}=\sum_{r=1}^{k}l_{ir}u_{rk},\quad i=k+1,k+2,\cdots,n,$$

可得 $\boldsymbol{L}$ 的第 k 列元素

$$l_{ik}=\frac{1}{u_{kk}}\left(a_{ik}-\sum_{r=1}^{k-1}l_{ir}u_{rk}\right),\quad i=k+1,k+2,\cdots,n \tag{5.13}$$

交替使用(5.12)式和(5.13)式，就能逐次计算出 $\boldsymbol{U}$(按行)和 $\boldsymbol{L}$(按列)的全部元素，而且可以把它们存放在矩阵 $\boldsymbol{A}$ 对应的位置上($\boldsymbol{L}$ 的对角线元素不必存放). 这就完成了 $\boldsymbol{A}$ 的 LU 分解.

由(5.10)－(5.13)式求得 $\boldsymbol{L}$ 和 $\boldsymbol{U}$ 后，解方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 就化为求解 $\boldsymbol{LUx}=\boldsymbol{b}$. 若记 $\boldsymbol{Ux}=\boldsymbol{y}$，则有 $\boldsymbol{Ly}=\boldsymbol{b}$. 于是可分两步求解方程组 $\boldsymbol{LUx}=\boldsymbol{b}$. 第一步求解 $\boldsymbol{Ly}=\boldsymbol{b}$，只要逐次用向前代入的方法即可求得 $\boldsymbol{y}$. 第二步求解 $\boldsymbol{Ux}=\boldsymbol{y}$，只要逐次用向后回代的方法即可得 $\boldsymbol{x}$. 设 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$，$\boldsymbol{y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$，$\boldsymbol{b}=(b_1,b_2,\cdots,b_n)^{\mathrm{T}}$，则有计算公式

$$\begin{cases} y_1=b_1, \\ y_i=b_i-\sum_{r=1}^{i-1}l_{ir}y_r,\quad i=2,3,\cdots,n. \end{cases} \tag{5.14}$$

$$\begin{cases} x_n=\dfrac{y_n}{u_{nn}}, \\ x_i=\dfrac{1}{u_{ii}}\left(y_i-\sum_{r=i+1}^{n}u_{ir}x_r\right),\quad i=n-1,n-2,\cdots,1. \end{cases} \tag{5.15}$$

以上解方程组的计算量与顺序 Gauss 消去法相当. 如果有一系列方程组，其系数矩阵都是相同的，右端向量 $\boldsymbol{b}$ 不同，则只需进行一次 LU 分解计算，上述解方程组的方法称为 LU 分解法，也称 Doolittle 方法.

例 **5.7**　用 LU 分解法求解

$$\begin{pmatrix} 6 & 2 & 1 & -1 \\ 2 & 4 & 1 & 0 \\ 1 & 1 & 4 & -1 \\ -1 & 0 & -1 & 3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}=\begin{pmatrix} 6 \\ -1 \\ 5 \\ -5 \end{pmatrix}.$$

解　由(5.10)—(5.13)式计算可得

$$\boldsymbol{L}=\begin{pmatrix} 1 & & & \\ \frac{1}{3} & 1 & & \\ \frac{1}{6} & \frac{1}{5} & 1 & \\ -\frac{1}{6} & \frac{1}{10} & -\frac{9}{37} & 1 \end{pmatrix},\quad \boldsymbol{U}=\begin{pmatrix} 6 & 2 & 1 & -1 \\ & \frac{10}{3} & \frac{2}{3} & \frac{1}{3} \\ & & \frac{37}{10} & -\frac{9}{10} \\ & & & \frac{191}{74} \end{pmatrix}.$$

由(5.14)式计算得

$$\boldsymbol{y}=\left(6,3,\frac{23}{5},-\frac{191}{74}\right)^{\mathrm{T}}.$$

由(5.15)式计算得

$$\boldsymbol{x}=(1,-1,1,-1)^{\mathrm{T}}.$$

(2) 列选主元的三角分解法.

设从 $\boldsymbol{A}=\boldsymbol{A}^{(1)}$ 开始已完成 $k-1$ 步分解计算，$\boldsymbol{U}$ 的元素(按行)和 $\boldsymbol{L}$ 的元素(按列)存放在 $\boldsymbol{A}$ 对应的位置，得到

$$\widetilde{\boldsymbol{A}}^{(k)}=\begin{pmatrix} u_{11} & u_{12} & \cdots & \cdots & \cdots & \cdots & u_{1n} \\ l_{21} & u_{22} & \cdots & \cdots & \cdots & \cdots & u_{2n} \\ \vdots & l_{32} & \ddots & & & & \vdots \\ \vdots & \vdots & \ddots & u_{k-1,k-1} & \cdots & \cdots & u_{k-1,n} \\ \vdots & \vdots & & l_{k,k-1} & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{n,k-1} & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}.$$

该矩阵与顺序 Gauss 消去法中得到的 $\boldsymbol{A}^{(k)}$ 是不同的，这种存贮方式的形式称为紧凑形式.

现作第 k 步计算，令

$$s_i=a_{ik}^{(k)}-\sum_{r=1}^{k-1}l_{ir}u_{rk},\quad i=k,k+1,\cdots,n.$$

当 $i=k$ 时，s_i 对应于(5.12)式中的 u_{kk}，它可能不宜在(5.13)式中作除法. 当 $i=k+1,\cdots,n$ 时，s_i 对应于(5.13)式中的分子. 记

$$|s_{i_k}|=\max_{k\leqslant i\leqslant n}|s_i|,$$

交换$(\widetilde{\boldsymbol{A}}^{(k)},\boldsymbol{b}^{(k)})$的第 i 行与第 i_k 行的位置，但每个位置上仍用原记号. 然后仍按(5.12)式计算 u_{kj}，$j=k+1,k+2,\cdots,n$，算出 $\boldsymbol{U}$ 的第 k 行. l_{ik} 的计算可用

$$l_{ik}=\frac{s_i}{s_{i_k}},\quad i=k+1,k+2,\cdots,n,$$

这就算出了 $\boldsymbol{L}$ 的第 k 列.

以上分解过程经过 $n-1$ 步，可得到 $\boldsymbol{PA}=\boldsymbol{LU}$. 因为 $\boldsymbol{b}$ 也参加换行运算，所以在其位置上已得到 $\boldsymbol{Pb}$. 最后再分两步求解方程组 $\boldsymbol{LUx}=\boldsymbol{Pb}$，即求解 $\boldsymbol{Ly}=\boldsymbol{Pb}$ 和 $\boldsymbol{Ux}=\boldsymbol{y}$.

例 **5.8** 用列选主元的三角分解法解

$$\begin{pmatrix}1&2&3\\3&1&5\\2&5&2\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}14\\20\\18\end{pmatrix}.$$

解 第一步列选主元后的分解计算结果为

$$(\widetilde{\boldsymbol{A}}^{(2)},\boldsymbol{b}^{(2)})=\begin{pmatrix}3&1&5&20\\\frac{1}{3}&2&3&14\\\frac{2}{3}&5&2&18\end{pmatrix}.$$

由于 $s_2=\frac{5}{3}<s_3=\frac{13}{3}$，所以第二步分解计算前要进行行交换，分解计算结果为

$$(\widetilde{\boldsymbol{A}}^{(3)},\boldsymbol{b}^{(3)})=\begin{pmatrix}3&1&5&20\\\frac{2}{3}&\frac{13}{3}&-\frac{4}{3}&\frac{14}{3}\\\frac{1}{3}&\frac{5}{13}&\frac{72}{39}&\frac{216}{39}\end{pmatrix}.$$

由此知

$$\boldsymbol{L}=\begin{pmatrix}1&&\\\frac{2}{3}&1&\\\frac{1}{3}&\frac{5}{13}&1\end{pmatrix},\quad \boldsymbol{U}=\begin{pmatrix}3&1&5\\&\frac{13}{3}&-\frac{4}{3}\\&&\frac{72}{39}\end{pmatrix},$$

$$\boldsymbol{P}=\boldsymbol{I}_{23}\boldsymbol{I}_{12}=\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix}.$$

由于方程组的右端向量已参与了消元计算，所以 $\boldsymbol{Ly}=\boldsymbol{Pb}$ 的解为 $\boldsymbol{y}=\boldsymbol{b}^{(3)}=$

$\left(20,\frac{14}{3},\frac{216}{39}\right)^{\mathrm{T}}$. 解 $\boldsymbol{U}\boldsymbol{x}=\boldsymbol{y}$，得 $\boldsymbol{x}=(1,2,3)^{\mathrm{T}}$.

5.2.2　三对角方程组的追赶法

设有方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{d}$，其中 $\boldsymbol{d}=(d_1,d_2,\cdots,d_n)^{\mathrm{T}}$，系数矩阵 $\boldsymbol{A}$ 是三对角形矩阵

$$\boldsymbol{A}=\begin{pmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & a_n & b_n \end{pmatrix} \tag{5.16}$$

如果 $\boldsymbol{A}$ 满足 Gauss 消去法可行的条件，当然可以用 LU 分解法求解. 并且，$\boldsymbol{L}$ 和 $\boldsymbol{U}$ 有如下形式

$$\boldsymbol{L}=\begin{pmatrix} 1 & & & & \\ l_2 & 1 & & & \\ & l_3 & 1 & & \\ & & \ddots & \ddots & \\ & & & l_n & 1 \end{pmatrix},\quad \boldsymbol{U}=\begin{pmatrix} u_1 & c_1 & & & \\ & u_2 & c_2 & & \\ & & \ddots & \ddots & \\ & & & \ddots & c_{n-1} \\ & & & & u_n \end{pmatrix} \tag{5.17}$$

利用(5.16)、(5.17)式和 $\boldsymbol{A}=\boldsymbol{L}\boldsymbol{U}$ 可得

$$\begin{cases} u_1=b_1 \\ l_i=\dfrac{a_i}{u_{i-1}}, \quad i=2,3,\cdots,n, \\ u_i=b_i-l_ic_{i-1}, \quad i=2,3,\cdots,n. \end{cases} \tag{5.18}$$

由此可求得 $\boldsymbol{L}$ 和 $\boldsymbol{U}$ 的所有元素. 解原方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{d}$ 可分为两步，即求解 $\boldsymbol{L}\boldsymbol{y}=\boldsymbol{d}$ 和 $\boldsymbol{U}\boldsymbol{x}=\boldsymbol{y}$，计算公式为

$$\begin{cases} y_1=d_1, \\ y_i=d_i-l_iy_{i-1}, \quad i=2,3,\cdots,n. \end{cases} \tag{5.19}$$

$$\begin{cases} x_n=\dfrac{y_n}{u_n}, \\ x_i=\dfrac{1}{u_i}(y_i-c_ix_{i+1}), \quad i=n-1,n-2,\cdots,1. \end{cases} \tag{5.20}$$

称(5.18)、(5.19)和(5.20)式为求解三对角方程组的追赶法，又称为 Thomas 算法.

追赶法能实现的条件是 $u_i\neq 0, i=1,2,\cdots,n$. 下面给出一个使追赶法可行的充分条件.

定理 5.6 设三对角矩阵 $\boldsymbol{A}$ 有(5.16)式的表达式,且满足

$$|b_1|>|c_1|>0,|b_n|>|a_n|>0,$$
$$|b_i|\geqslant|a_i|+|c_i|,\quad a_ic_i\neq0,\quad i=2,3,\cdots,n-1.$$

则 $\boldsymbol{A}$ 非奇异,且有

$$0<\left|\frac{c_i}{u_i}\right|<1,\quad i=1,2,\cdots,n,$$
$$|b_i|-|a_i|<|u_i|<|b_i|+|a_i|,\quad i=2,3,\cdots,n.$$

证 用归纳法.对 $i=1$,有 $|u_1|=|b_1|>|c_1|$,所以 $|u_1|\neq0,0<\frac{|c_1|}{|u_1|}<1$.

现设 $u_{i-1}\neq0,\frac{|c_{i-1}|}{|u_{i-1}|}<1$,我们有

$$|u_i|=|b_i-l_ic_{i-1}|\geqslant|b_i|-\frac{|a_ic_{i-1}|}{|u_{i-1}|}>|b_i|-|a_i|.$$

利用所给条件得到 $|u_i|>|c_i|$,故 $u_i\neq0,\frac{|c_i|}{|u_i|}<1$.另一方面,有

$$|u_i|\leqslant|b_i|+|l_ic_{i-1}|=|b_i|+\frac{|a_ic_{i-1}|}{|u_{i-1}|}\leqslant|b_i|+|a_i|.$$

因为 $\det\boldsymbol{A}=u_1u_2\cdots u_n$,所以 $\det\boldsymbol{A}\neq0$.定理得证.

在定理 5.6 的条件下,追赶法可以进行计算,并且计算过程的中间变量有界,不会产生大的变化,可以有效地算出结果.

在定理 5.6 的条件中,要求 a_i 和 c_i 非零.若有某个 a_i(或 c_i)为零,则三对角方程组可以化为两个低阶的非耦合的方程组.

追赶法公式简单,计算量和存储量都小.整个求解过程仅需 $5n-4$ 次乘除与 $3(n-1)$ 次加减法运算,仅需 4 个一维数组存储系数矩阵的元素和右端向量,l_i,u_i 和 x_i 可分别存放在表示系数矩阵元素的数组和右端向量的位置.

例 5.9 用追赶法求解三对角方程组 $\boldsymbol{Ax}=\boldsymbol{d}$,其中

$$\boldsymbol{A}=\begin{pmatrix}4&-1&\\-1&4&-1\\&-1&4\end{pmatrix},\quad \boldsymbol{d}=\begin{pmatrix}1\\3\\2\end{pmatrix}.$$

解 由(5.18)式得

$$\boldsymbol{L}=\begin{pmatrix}1&&\\-0.25&1&\\&-0.2667&1\end{pmatrix},\quad \boldsymbol{U}=\begin{pmatrix}4&-1&\\&3.75&-1\\&&3.7333\end{pmatrix}.$$

由(5.19)式和(5.20)式得

$$\boldsymbol{y}=(1,3.25,2.8668)^{\mathrm{T}},\quad \boldsymbol{x}=(0.5179,1.0714,0.7679)^{\mathrm{T}}.$$

对于另一类方程组，即在周期样条插值等问题会遇到的循环三对角方程组 $\boldsymbol{Ax}=\boldsymbol{d}$，其中

$$\boldsymbol{A}=\begin{pmatrix} b_1 & c_1 & & & a_1 \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1} & b_{n-1} & c_{n-1} \\ c_n & & & a_n & b_n \end{pmatrix},$$

我们也可以用三角分解的方法. 从矩阵零元素的位置不难验证 $\boldsymbol{L}$ 和 $\boldsymbol{U}$ 可写成下面的形式

$$\boldsymbol{L}=\begin{pmatrix} 1 & & & & \\ l_2 & 1 & & & \\ & l_3 & 1 & & \\ & & \ddots & \ddots & \\ \sigma_1 & \sigma_2 & \cdots & \sigma_{n-1}+l_n & 1 \end{pmatrix},\quad \boldsymbol{U}=\begin{pmatrix} u_1 & c_1 & & & \rho_1 \\ & u_2 & c_2 & & \rho_2 \\ & & \ddots & \ddots & \vdots \\ & & & u_{n-1} & c_{n-1}+\rho_{n-1} \\ & & & & u_n \end{pmatrix}.$$

由此，不难得到 $\boldsymbol{L}$ 和 $\boldsymbol{U}$ 的元素的计算公式，这里不再介绍.

5.2.3　平方根法

当 $\boldsymbol{A}$ 为对称正定矩阵时，对 $\boldsymbol{A}$ 可直接作 LU 分解. 进一步地，由(5.8)式可得下面的定理.

定理 5.7　设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$，$\boldsymbol{A}^{\mathrm{T}}=\boldsymbol{A}$，且 $\boldsymbol{A}$ 的顺序主子式 $D_i\neq 0(i=1,2,\cdots,n)$，则存在唯一的单位下三角阵 $\boldsymbol{L}$ 和对角阵 $\boldsymbol{D}$，使

$$\boldsymbol{A}=\boldsymbol{LDL}^{\mathrm{T}}. \tag{5.21}$$

定理 5.8　设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$，$\boldsymbol{A}$ 为对称正定矩阵，则存在唯一的对角元素为正的下三角阵 $\boldsymbol{L}$，使

$$\boldsymbol{A}=\boldsymbol{LL}^{\mathrm{T}}. \tag{5.22}$$

证　由定理 5.7 可知 $\boldsymbol{A}=\boldsymbol{L}_1\boldsymbol{DL}_1^{\mathrm{T}}$，其中 $\boldsymbol{L}_1$ 为单位下三角阵，$\boldsymbol{D}=\mathrm{diag}(d_1,d_2,\cdots,d_n)$. 若令 $\boldsymbol{U}=\boldsymbol{DL}_1^{\mathrm{T}}$，则 $\boldsymbol{A}=\boldsymbol{L}_1\boldsymbol{U}$ 为 $\boldsymbol{A}$ 的 Dolittle 分解，$\boldsymbol{U}$ 的对角元即为 $\boldsymbol{D}$ 的对角元. 因此，$\boldsymbol{A}$ 的顺序主子式 $D_m=d_1d_2\cdots d_m$，$m=1,2,\cdots,n$. 因为 $\boldsymbol{A}$ 正定，所以 $D_i>0$，$i=1,2,\cdots,n$. 由此推出 $d_i>0$，$i=1,2,\cdots,n$. 记

$$\boldsymbol{D}^{\frac{1}{2}}=\mathrm{diag}(\sqrt{d_1},\sqrt{d_2},\cdots,\sqrt{d_n}),$$

令 $\boldsymbol{L}=\boldsymbol{L}_1\boldsymbol{D}^{\frac{1}{2}}$，则有

$$\boldsymbol{A}=\boldsymbol{L}_1\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{L}_1^{\mathrm{T}}=(\boldsymbol{L}_1\boldsymbol{D}^{\frac{1}{2}})(\boldsymbol{L}_1\boldsymbol{D}^{\frac{1}{2}})^{\mathrm{T}}=\boldsymbol{LL}^{\mathrm{T}}.$$

由分解式 $\boldsymbol{L}_1\boldsymbol{DL}_1^{\mathrm{T}}$ 的唯一性可得分解式(5.22)的唯一性. 定理得证.

称(5.22)式为矩阵 $\boldsymbol{A}$ 的 Cholesky 分解. 利用 $\boldsymbol{A}$ 的 Cholesky 分解式来求解方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的方法称为 Cholesky 方法或平方根法,这是因为计算过程含开方运算.

设 $\boldsymbol{A}=(a_{ij})$,

$$\boldsymbol{L}=\begin{pmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{m} \end{pmatrix}.$$

由(5.22)式可得

$$a_{ij}=\sum_{k=1}^{j-1} l_{ik}l_{jk}+l_{ij}l_{jj},\quad i=j,j+1,\cdots,n.$$

按逐列计算 $\boldsymbol{L}$ 的元素的计算步骤,设第 1 列至第 $j-1$ 列已经计算得到,则有

$$l_{jj}=\Big(a_{jj}-\sum_{k=1}^{j-1} l_{jk}^2\Big)^{\frac{1}{2}}, \tag{5.23}$$

$$l_{ij}=\frac{1}{l_{jj}}\Big(a_{ij}-\sum_{k=1}^{j-1} l_{ik}l_{jk}\Big),\quad i=j+1,j+2,\cdots,n. \tag{5.24}$$

这样,可以从 $j=1$ 直到 $j=n$ 逐列算出 $\boldsymbol{L}$ 的元素,再求解下三角方程组 $\boldsymbol{Ly}=\boldsymbol{b}$ 和上三角方程组 $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{x}=\boldsymbol{y}$,计算公式为

$$y_1=\frac{b_1}{l_{11}},\quad y_i=\frac{1}{l_{ii}}\Big(b_i-\sum_{k=1}^{i-1} l_{ik}y_k\Big),\quad i=2,3,\cdots,n,$$

$$x_n=\frac{y_n}{l_{m}},\quad x_i=\frac{1}{l_{ii}}\Big(y_i-\sum_{k=i+1}^{n} l_{ki}x_k\Big),\quad i=n-1,n-2,\cdots,1.$$

平方根法的原理基于矩阵的 LU 分解,所以它也是 Gauss 消去法的变形. 但由于利用了矩阵对称正定的性质,减少了计算量. 平方根法的乘除法运算次数为 $\frac{n^3+9n^2+2n}{6}$,加减法运算次数 $\frac{n^3+6n^2-7n}{6}$. 另外还有 n 次开平方运算,其所含乘除法和加减法次数可分别看成 n 的常数倍. 因此,平方根法约需 $\frac{n^3}{6}$ 次乘除法,与 Gauss 消去法相比减少了一半.

由(5.23)式可得 $a_{jj}=\sum_{k=1}^{j} l_{jk}^2$,由此推出 $|l_{jk}|\leqslant\sqrt{a_{jj}}$,$k=1,2,\cdots,j$. 所以,平方根法的中间量 l_{jk} 得以控制,不必选主元.

例 5.10 用平方根法求解

$$\begin{cases} 4x_1-x_2+x_3=6, \\ -x_1+4.25x_2+2.75x_3=-0.5, \\ x_1+2.75x_2+3.5x_3=1.25. \end{cases}$$

解　不难验证系数矩阵是对称正定的，按(5.23)和(5.24)式依次计算得

$$\boldsymbol{L}=\begin{pmatrix} 2 & & \\ -0.5 & 2 & \\ 0.5 & 1.5 & 1 \end{pmatrix}.$$

解 $\boldsymbol{Ly}=(6,-0.5,1.25)^{\mathrm{T}}$，得 $\boldsymbol{y}=(3,0.5,-1)^{\mathrm{T}}$，再解 $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{x}=\boldsymbol{y}$ 可以得到 $\boldsymbol{x}=(2,1,-1)^{\mathrm{T}}$.

如果对矩阵 $\boldsymbol{A}$ 采用分解式(5.21)，即

$$\boldsymbol{A}=\begin{pmatrix} 1 & & & \\ l_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \cdots & 1 \end{pmatrix}\begin{pmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_n \end{pmatrix}\begin{pmatrix} 1 & l_{21} & \cdots & l_{n1} \\ & 1 & \cdots & l_{n2} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix}$$

则可避免开平方根运算，称为改进的平方根法. 它既适合于求解对称正定方程组，也适合于求解 $\boldsymbol{A}$ 对称且其顺序主子式全不为零的方程组. 分解式的计算公式为($j=1,2,\cdots,n$)

$$\begin{cases} d_j = a_{jj} - \sum\limits_{k=1}^{j-1} l_{jk}^2 d_k, \\ l_{ij} = \dfrac{1}{d_j}\left(a_{ij} - \sum\limits_{k=1}^{j-1} d_k l_{ik} l_{jk}\right), \quad i = j+1, j+2, \cdots, n, \end{cases}$$

其中 $j=1$ 时，求和部分为零. 这样，求解方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 化为求解 $\boldsymbol{Ly}=\boldsymbol{b}$ 和 $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{x}=\boldsymbol{D}^{-1}\boldsymbol{y}$.

对于例 5.10 给定的方程组，用改进的平方根法有

$$\boldsymbol{L}=\begin{pmatrix} 1 & & \\ -0.25 & 1 & \\ 0.25 & 0.75 & 1 \end{pmatrix},\quad \boldsymbol{D}=\begin{pmatrix} 4 & & \\ & 4 & \\ & & 1 \end{pmatrix}.$$

解 $\boldsymbol{Ly}=\boldsymbol{b}$ 得 $\boldsymbol{y}=(6,1,-1)^{\mathrm{T}}$. 解 $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{x}=\boldsymbol{D}^{-1}\boldsymbol{y}$ 得 $\boldsymbol{x}=(2,1,-1)^{\mathrm{T}}$.

5.3　方程组的性态与误差估计

5.3.1　矩阵的条件数

先看一个例子，说明方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的解对 $\boldsymbol{A}$ 或 $\boldsymbol{b}$ 的扰动的敏感性问题.

例 5.11　方程组

$$\begin{pmatrix} 3 & 1 \\ 3.000\,1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}=\begin{pmatrix} 4 \\ 4.000\,1 \end{pmatrix}$$

的准确解是$(1,1)^{\mathrm{T}}$. 若 $\boldsymbol{A}$ 及 $\boldsymbol{b}$ 作微小的变化,考虑扰动后的方程组

$$\begin{bmatrix} 3 & 1 \\ 2.999\,9 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} 4 \\ 4.000\,2 \end{bmatrix},$$

其准确解是$(-2,10)^{\mathrm{T}}$.

在上例中,$\boldsymbol{A}$ 和 $\boldsymbol{b}$ 的微小变化引起 $\boldsymbol{x}$ 很大的变化,$\boldsymbol{x}$ 对 $\boldsymbol{A}$ 和 $\boldsymbol{b}$ 的扰动是敏感的. 这种现象的出现是完全由方程组的性态决定的.

定义 5.1 如果方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 中,矩阵 $\boldsymbol{A}$ 和右端项 $\boldsymbol{b}$ 的微小变化,引起解向量 $\boldsymbol{x}$ 的很大变化,则称 $\boldsymbol{A}$ 为关于解方程组和矩阵求逆的病态矩阵,称相应的方程组为病态方程组. 否则,称 $\boldsymbol{A}$ 为良态矩阵,称相应的方程组为良态方程组.

我们需要一种能刻画矩阵和方程组"病态"程度的标准. 暂不考虑矩阵 $\boldsymbol{A}$ 的扰动,仅考虑 $\boldsymbol{b}$ 的扰动对方程组解的影响,设方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的扰动方程组为 $\boldsymbol{A}(\boldsymbol{x}+\delta\boldsymbol{x})=\boldsymbol{b}+\delta\boldsymbol{b}$,则

$$\delta\boldsymbol{x}=\boldsymbol{A}^{-1}\delta\boldsymbol{b},\quad \|\delta\boldsymbol{x}\|\leqslant\|\boldsymbol{A}^{-1}\|\,\|\delta\boldsymbol{b}\|.$$

又由于 $\|\boldsymbol{b}\|\leqslant\|\boldsymbol{A}\|\,\|\boldsymbol{x}\|$,即得

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{A}\|\,\frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}.$$

可见,量 $\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{A}\|$ 是相对误差 $\dfrac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}$ 的倍增因子,该量越大,方程组右端变化所引起的解向量的相对误差就可能越大.

定义 5.2 设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 为可逆矩阵,按算子范数,称

$$\mathrm{cond}(\boldsymbol{A})=\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{A}\| \tag{5.25}$$

为矩阵 $\boldsymbol{A}$ 的条件数.

如果矩阵范数取 2 范数,则记 $\mathrm{cond}_2(\boldsymbol{A})=\|\boldsymbol{A}^{-1}\|_2\|\boldsymbol{A}\|_2$. 按(5.25)式,同样可以定义 $\mathrm{cond}_\infty(\boldsymbol{A})$和 $\mathrm{cond}_1(\boldsymbol{A})$.

设 $\boldsymbol{A}^{-1}$存在,条件数有如下一些性质:

(1) $\mathrm{cond}(\boldsymbol{A})\geqslant 1$, $\mathrm{cond}(\boldsymbol{A})=\mathrm{cond}(\boldsymbol{A}^{-1})$, $\mathrm{cond}(\alpha\boldsymbol{A})=\mathrm{cond}(\boldsymbol{A})$,其中 $\alpha\in\mathbf{R}$, $\alpha\neq 0$;

(2) 若 $\boldsymbol{U}$ 为正交阵,即 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}$,则

$$\mathrm{cond}_2(\boldsymbol{U})=1,\ \mathrm{cond}_2(\boldsymbol{A})=\mathrm{cond}_2(\boldsymbol{AU})=\mathrm{cond}_2(\boldsymbol{UA}).$$

(3) 设 λ_1 与 λ_n 为 $\boldsymbol{A}$ 按绝对值最大和最小的特征值,则

$$\mathrm{cond}(\boldsymbol{A})\geqslant\frac{|\lambda_1|}{|\lambda_n|}.$$

若 $\boldsymbol{A}$ 对称,则 $\mathrm{cond}_2(\boldsymbol{A})=\dfrac{|\lambda_1|}{|\lambda_n|}$.

例 5.12 下列 Hilbert 矩阵是一族著名的病态矩阵:

$$\boldsymbol{H}_n=\begin{pmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n+1} \\ \vdots & \vdots & & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \cdots & \frac{1}{2n-1} \end{pmatrix}.$$

它是一个 $n\times n$ 的对称矩阵，可以证明它是正定的. 计算条件数有 $\mathrm{cond}_2(\boldsymbol{H}_4)=1.551\,4\times10^4$，$\mathrm{cond}_2(\boldsymbol{H}_6)=1.495\,1\times10^7$，$\mathrm{cond}_2(\boldsymbol{H}_8)=1.525\times10^{10}$. 由此可见，随着 n 的增加，$\boldsymbol{H}_n$ 的病态可能越严重. $\boldsymbol{H}_n$ 常常在数据拟合和函数逼近中出现.

对于实际问题，条件数一般是很难计算的. 下列现象可能表示方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 是病态的.

(1) 如果矩阵 $\boldsymbol{A}$ 的按绝对值最大特征值与最小特征值之比很大，则 $\boldsymbol{A}$ 是病态的；

(2) 如果系数矩阵 $\boldsymbol{A}$ 的元素间数量级相差很大，并且无一定规则，则 $\boldsymbol{A}$ 可能病态；

(3) 如果系数矩阵 $\boldsymbol{A}$ 的某些行或列是近似线性相关的，或系数矩阵 $\boldsymbol{A}$ 的行列式值相对来说很小，则 $\boldsymbol{A}$ 可能病态；

(4) 如果在 $\boldsymbol{A}$ 的三角化过程中出现小主元，或采用选主元技术时，主元素数量级相差悬殊，则 $\boldsymbol{A}$ 可能病态.

对于病态方程组，数值求解必须小心进行，否则达不到所要求的准确度. 有时可以用高精度(如双精度或扩充精度)的运算，以改善或减轻方程组的病态程度. 有时也可对原方程作某些预处理，以降低系数矩阵的条件数，即选择非奇异矩阵 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$，一般选择 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$ 为对角阵或三角矩阵，使

$$\mathrm{cond}(\boldsymbol{PAQ})<\mathrm{cond}(\boldsymbol{A}).$$

然后，求解等价方程组 $\boldsymbol{PAQy}=\boldsymbol{Py}$，$\boldsymbol{y}=\boldsymbol{Q}^{-1}\boldsymbol{x}$.

例如，对矩阵

$$\boldsymbol{A}=\begin{bmatrix} 1 & 10^5 \\ 1 & 1 \end{bmatrix},\quad \boldsymbol{A}^{-1}=\frac{1}{1-10^5}\begin{bmatrix} 1 & -10^5 \\ -1 & 1 \end{bmatrix},$$

有 $\mathrm{cond}_\infty\approx10^5$. 若进行预处理

$$\boldsymbol{B}=\boldsymbol{PA}=\begin{bmatrix} 10^{-5} & 0 \\ 0 & 1 \end{bmatrix}\boldsymbol{A}=\begin{bmatrix} 10^{-5} & 1 \\ 1 & 1 \end{bmatrix}.$$

则 $\mathrm{cond}_\infty(\boldsymbol{B})=4$，条件数得到改善.

5.3.2 方程组解的误差估计

由于舍入误差,我们解方程组往往得到的是近似解.下面利用条件数给出近似解的事前误差估计和事后误差估计,即计算之前和计算之后的误差估计.

定理 5.9 设 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$,$\boldsymbol{A}$ 为非奇异阵,$\boldsymbol{b}$ 为非零向量,$\boldsymbol{A}$ 和 $\boldsymbol{b}$ 分别有扰动 $\delta\boldsymbol{A}$ 和 $\delta\boldsymbol{b}$,$(\boldsymbol{A}+\delta\boldsymbol{A})(\boldsymbol{x}+\delta\boldsymbol{x})=\boldsymbol{b}+\delta\boldsymbol{b}$.若 $\|\boldsymbol{A}^{-1}\|\ \|\delta\boldsymbol{A}\|<1$,则有误差估计式

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\frac{\operatorname{cond}(\boldsymbol{A})}{1-\|\boldsymbol{A}^{-1}\|\ \|\delta\boldsymbol{A}\|}\left(\frac{\|\delta\boldsymbol{A}\|}{\|\boldsymbol{A}\|}+\frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}\right). \tag{5.26}$$

证 将 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 代入扰动方程组$(\boldsymbol{A}+\delta\boldsymbol{A})(\boldsymbol{x}+\delta\boldsymbol{x})=\boldsymbol{b}+\delta\boldsymbol{b}$,整理后有

$$\delta\boldsymbol{x}=\boldsymbol{A}^{-1}[\delta\boldsymbol{b}-(\delta\boldsymbol{A})\boldsymbol{x}-(\delta\boldsymbol{A})(\delta\boldsymbol{x})].$$

将上式两端取范数,则有

$$\|\delta\boldsymbol{x}\|\leqslant\|\boldsymbol{A}^{-1}\|(\|\delta\boldsymbol{b}\|+\|\delta\boldsymbol{A}\|\ \|\boldsymbol{x}\|+\|\delta\boldsymbol{A}\|\ \|\delta\boldsymbol{x}\|),$$

经整理后得

$$(1-\|\boldsymbol{A}^{-1}\|\ \|\delta\boldsymbol{A}\|)\|\delta\boldsymbol{x}\|\leqslant\|\boldsymbol{A}^{-1}\|(\|\delta\boldsymbol{b}\|+\|\boldsymbol{A}\|\ \|\boldsymbol{x}\|).$$

由于 $\|\boldsymbol{A}^{-1}\|\ \|\delta\boldsymbol{A}\|<1$,则有

$$\|\delta\boldsymbol{x}\|\leqslant\frac{\|\boldsymbol{A}^{-1}\|}{1-\|\boldsymbol{A}^{-1}\|\ \|\delta\boldsymbol{A}\|}(\|\delta\boldsymbol{b}\|+\|\delta\boldsymbol{A}\|\ \|\boldsymbol{x}\|).$$

再利用 $\|\boldsymbol{b}\|\leqslant\|\boldsymbol{A}\|\ \|\boldsymbol{x}\|$ 即得所证.

若 $\|\delta\boldsymbol{A}\|=0$,$\|\delta\boldsymbol{b}\|\neq0$ 时,则由(5.26)式有

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\operatorname{cond}(\boldsymbol{A})\frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}.$$

若 $\|\delta\boldsymbol{A}\|\neq0$,$\|\delta\boldsymbol{b}\|=0$,则由(5.26)式有

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\frac{\operatorname{cond}(\boldsymbol{A})}{1-\|\boldsymbol{A}^{-1}\|\ \|\delta\boldsymbol{A}\|}\frac{\|\delta\boldsymbol{A}\|}{\|\boldsymbol{A}\|}=\operatorname{cond}(\boldsymbol{A})\frac{\|\delta\boldsymbol{A}\|}{\|\boldsymbol{A}\|}(1+O(\|\delta\boldsymbol{A}\|)).$$

例 5.13 设矩阵 $\boldsymbol{A}$ 可逆,$\delta\boldsymbol{A}$ 为扰动矩阵.试证当 $\|\boldsymbol{A}^{-1}\delta\boldsymbol{A}\|<1$ 时,$\boldsymbol{A}+\delta\boldsymbol{A}$ 也可逆.

证 考虑行列式

$$\det(\boldsymbol{A}^{-1})\det(\boldsymbol{A}+\delta\boldsymbol{A})=\det(\boldsymbol{A}^{-1}(\boldsymbol{A}+\delta\boldsymbol{A}))=\det(\boldsymbol{I}+\boldsymbol{A}^{-1}\delta\boldsymbol{A}).$$

因为 $\|\boldsymbol{A}^{-1}\delta\boldsymbol{A}\|<1$,所以 $I+\boldsymbol{A}^{-1}\delta\boldsymbol{A}$ 可逆.于是 $\det(\boldsymbol{A}+\delta\boldsymbol{A})\neq0$,即矩阵 $\boldsymbol{A}+\delta\boldsymbol{A}$ 可逆.

例 5.14 设有方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$,其中

$$\boldsymbol{A}=\begin{pmatrix}1&0&-1\\2&2&1\\0&2&2\end{pmatrix},\quad \boldsymbol{b}=\begin{pmatrix}\frac{1}{2}\\[2pt]\frac{1}{3}\\[2pt]-\frac{2}{3}\end{pmatrix}.$$

已知它有解 $x=\left(\frac{1}{2},-\frac{1}{3},0\right)^{\mathrm{T}}$，如果右端有小扰动 $\|\delta \boldsymbol{b}\|_{\infty}=\frac{1}{2}\times 10^{-6}$，试估计由此引起的解的相对误差.

解　由于

$$\boldsymbol{A}^{-1}=\begin{pmatrix}-1 & 1 & -1\\ 2 & -1 & 1.5\\ -2 & 1 & -1\end{pmatrix},$$

从而 $\mathrm{cond}_{\infty}(\boldsymbol{A})=22.5$，于是

$$\frac{\|\delta \boldsymbol{x}\|_{\infty}}{\|\boldsymbol{x}\|_{\infty}}\leqslant \mathrm{cond}_{\infty}(\boldsymbol{A})\frac{\|\delta \boldsymbol{b}\|_{\infty}}{\|\boldsymbol{b}\|_{\infty}}$$

$$=22.5\times\frac{0.5\times 10^{-6}}{\frac{2}{3}}=1.687\,5\times 10^{-5}.$$

定理 5.10　设 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b},\boldsymbol{b}\neq 0$ 则对方程组的近似解 $\tilde{\boldsymbol{x}}$ 有误差估计式

$$\frac{1}{\mathrm{cond}(\boldsymbol{A})}\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|}\leqslant\frac{\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\mathrm{cond}(\boldsymbol{A})\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|},$$

其中 $\boldsymbol{r}=\boldsymbol{b}-\boldsymbol{A}\tilde{\boldsymbol{x}}$ 为剩余向量.

证　由 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 有

$$\boldsymbol{r}=\boldsymbol{A}\boldsymbol{x}-\boldsymbol{A}\tilde{\boldsymbol{x}}=\boldsymbol{A}(\boldsymbol{x}-\tilde{\boldsymbol{x}}),$$

$$\frac{\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\|\boldsymbol{A}^{-1}\boldsymbol{r}\|\cdot\frac{\|\boldsymbol{A}\|}{\|\boldsymbol{b}\|}\leqslant\mathrm{cond}(\boldsymbol{A})\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|}.$$

又由 $\boldsymbol{x}=\boldsymbol{A}^{-1}\boldsymbol{b}$，有

$$\frac{\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\geqslant\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{A}\|}\frac{1}{\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{b}\|}=\frac{1}{\mathrm{cond}(\boldsymbol{A})}\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|}.$$

定理得证.

该定理说明，当 $\mathrm{cond}(\boldsymbol{A})$ 很大时，即使方程组余量 $\boldsymbol{r}$ 的相对误差已经很小，近似解的相对误差仍然可能很大.

如果用直接解法得到的近似解 $\tilde{\boldsymbol{x}}$ 误差较大，我们可以用迭代改善的办法对近似解进行修正. 设 $\boldsymbol{r}=\boldsymbol{b}-\boldsymbol{A}\tilde{\boldsymbol{x}}$，$\Delta\boldsymbol{x}$ 为修正量，$\bar{\boldsymbol{x}}=\tilde{\boldsymbol{x}}+\Delta\boldsymbol{x}$ 为新的近似解. 这样，我们可以通过求解

$$\boldsymbol{A}\Delta\boldsymbol{x}=\boldsymbol{r}\tag{5.27}$$

得到 $\bar{\boldsymbol{x}}$. 显然，在准确运算下有

$$\boldsymbol{A}\bar{\boldsymbol{x}}=\boldsymbol{A}(\tilde{\boldsymbol{x}}+\Delta\boldsymbol{x})=\boldsymbol{b}-\boldsymbol{r}+\boldsymbol{A}\Delta\boldsymbol{x}=\boldsymbol{b}.$$

然而，在实际计算时，方程组(5.27)不大可能准确求解，所以解(5.27)式只能提供有限的修正. 因此，需要反复求解形为(5.27)式的方程组，不断对所得的近似解进行改进. 这种使近似值逐渐接近真解的过程称为选代改善. 为了节省计算

量，可事先对矩阵 $\boldsymbol{A}$ 进行 LU 分解，把反复解形为(5.27)的方程组改为反复解形为 $\boldsymbol{Ly}=\boldsymbol{r},\boldsymbol{U}\Delta\boldsymbol{x}=\boldsymbol{y}$ 的方程组. 为了保证计算精度，计算剩余向量 $\boldsymbol{r}$ 可采用高精度计算.

方程组直接解法的稳定性是应当注重的问题. 如果通过直接计算每一步舍入误差对解的影响来获得近似解的误差界，那将是非常困难的. J. H. Wilkinson 等人提出了“向后误差分析法”，其基本思想是把计算过程中舍入误差对解的影响归结为原始数据扰动对解的影响. 下面给出一个定理来说明这方面的结果.

定理 5.11 设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$，$\boldsymbol{A}$ 为非奇异阵，用列主元法或全主元法解方程组 $\boldsymbol{Ax}=\boldsymbol{b}$，其计算解 $\tilde{\boldsymbol{x}}$ 满足 $(\boldsymbol{A}+\delta\boldsymbol{A})\tilde{\boldsymbol{x}}=\boldsymbol{b}$. 记计算机尾数字长为 t，且 $n2^{-t}\leqslant 0.01$. 记 $\rho=\dfrac{\max\limits_{1\leqslant i,j\leqslant n}|a_{ij}^{(k)}|}{\|\boldsymbol{A}\|_\infty}$，$a_{ij}^{(k)}$ 是消去过程中 $\boldsymbol{A}^{(k)}$ 中的元素，则有

(1) 若 $\boldsymbol{A}$ 的 LU 分解计算结果为 $\tilde{\boldsymbol{L}},\tilde{\boldsymbol{U}}$，则

$$\tilde{\boldsymbol{L}}\tilde{\boldsymbol{U}}=\boldsymbol{A}+\boldsymbol{E},$$

$$\|\boldsymbol{E}\|_\infty\leqslant\rho n^2\|\boldsymbol{A}\|_\infty 2^{-t}.$$

(2) $\|\delta\boldsymbol{A}\|_\infty\leqslant 1.01\rho(n^3+3n^2)\|\boldsymbol{A}\|_\infty 2^{-t}$.

(3) 计算解有精度估计：

$$\frac{\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|_\infty}{\|\boldsymbol{x}\|_\infty}\leqslant\frac{\mathrm{cond}_\infty(\boldsymbol{A})}{1-\|\boldsymbol{A}^{-1}\|_\infty\|\delta\boldsymbol{A}\|_\infty}[1.01\rho(n^3+3n^2)2^{-t}].$$

该定理说明，矩阵 $\boldsymbol{A}$ 的阶数越高、条件数越大、矩阵元素的增长因子 ρ 越大和计算机字长越短，则舍入误差对解的影响越严重. 因此，计算精度取决于矩阵的规模、方程组的性态、所选取的算法和所用计算机字长.

评　注

本章介绍了求解线性代数方程组的两种直接方法：Gauss 消去法和三角分解法. 简单论述了方程组的性态和误差估计. 直接方法是古典的方法，我国古代数学名著《九章算术》中就有消去法低阶情形的叙述，直到今天人们用高速计算机解方程组，特别是阶数不太大的或系数矩阵稀疏的方程组，消去法仍然是一种有力的工具，一般情形下它的计算量为 $O(n^3)$.

在 Gauss 消去法中引进选主元的技巧，就得了解方程组的完全主元素消去法和列主元素消去法. 完全选主元素和列主元素方法都是稳定的算法. 用完全主元素消去法解非病态方程组具有较高的精确度，但它需要花费较多的机器时间. 列主元素消去法是比完全主元素消去法更实用的算法，一般使用较多. 用 Gauss-Jordan 消去法求逆矩阵是比较方便的. 当系数矩阵呈三对角形时，特别是

对角线元素的绝对值大于它所在行的其他元素的绝对值之和的对角占优矩阵，追赶法通常是一种既快速又数值稳定的方法. 当方程组的系数矩阵是对称正定或对角占优时，则不必选主元而直接用 Gauss 顺序消去法或 Doolittle 分解方法. 系数矩阵为对称正定的情形，在非病态的情况下，Cholesky 方法是一种有效的方法.

关于矩阵的条件数、病态方程组、算法的稳定性、误差估计，这些概念都是计算数学中比较重要的概念，本章只作了简单的介绍. 对于病态问题，最好扩大运算字长，如采用双精度或扩充精度.

习　题　5

5.1　用 Gauss 消去法和 Doolittle 分解法求解

$$\begin{pmatrix} 7 & 1 & -1 \\ 2 & 4 & 2 \\ -1 & 1 & 3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}=\begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}.$$

5.2　用 Gauss 消去法和 Gauss 列主元消去法求解

$$\begin{pmatrix} 0.729 & 0.81 & 0.9 \\ 1 & 1 & 1 \\ 1.331 & 1.21 & 1.1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}=\begin{pmatrix} 0.686\,7 \\ 0.833\,8 \\ 1.000\,0 \end{pmatrix}.$$

计算过程取 4 位有效数字，并同精确解$(0.224\,5, 0.281\,4, 0.327\,9)^{\mathrm{T}}$比较.

5.3　设 $\boldsymbol{A}=(a_{ij})_{n\times n}$，$a_{11}\neq 0$，经过一步 Gauss 消去法得到

$$\boldsymbol{A}^{(2)}=\begin{pmatrix} a_{11} & a_1^{\mathrm{T}} \\ 0 & \boldsymbol{A}_2 \end{pmatrix}.$$

试证明：

(1) 若 $\boldsymbol{A}$ 对称，则 $\boldsymbol{A}_2$ 对称；

(2) 若 $\boldsymbol{A}$ 对称正定，则 $\boldsymbol{A}_2$ 对称正定；

(3) 若 $\boldsymbol{A}$ 严格对角占优，即 $|a_{ii}|>\sum\limits_{j\neq 1}|a_{ij}|$，$i=1,2,\cdots,n$，则 $\boldsymbol{A}_2$ 也严格对角占优.

5.4　设 $\boldsymbol{A}=(a_{ij})_{n\times n}$ 是对称正定矩阵，经 $k-1$ 步 Gauss 消元后约化为 $\boldsymbol{A}^{(k)}=(a_{ij}^{(k)})$. 试证明：

(1) $a_{ii}>0$，$i=1,2,\cdots,n$；

(2) $\boldsymbol{A}$ 的绝对值最大的元素必在对角线上；

(3) $a_{ii}^{(2)}\leqslant a_{ii}$，$i=1,2,\cdots,n$；

(4) $\max\limits_{1\leqslant i,j\leqslant n}|a_{ij}^{(k)}|\leqslant\max\limits_{1\leqslant i,j\leqslant n}|a_{ij}|$, $k=2,3,\cdots,n$.

5.5 设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$,其第 k 列为 $(a_{1k},a_{2k},\cdots,a_{nk})^{\mathrm{T}}$, $a_{kk}\neq0$,其他各列依次为单位向量 $\boldsymbol{e}_1,\cdots,\boldsymbol{e}_{k-1},\boldsymbol{e}_{k+1},\cdots,\boldsymbol{e}_n$. 试证 $\boldsymbol{A}^{-1}$ 的第 k 列为

$$-\frac{1}{a_{kk}}(a_{1k},\cdots,a_{k-1,k},-1,a_{k+1,k},\cdots,a_{nk})^{\mathrm{T}},$$

其他各列与 $\boldsymbol{A}$ 的各列相同.

5.6 用 Gauss-Jordan 消去法解方程组

$$\begin{pmatrix}2&3&4\\1&1&9\\1&2&-6\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}0\\2\\1\end{pmatrix}.$$

5.7 下列矩阵能否作 Doolittle 分解? 若能分解,分解式是否唯一?

$$\boldsymbol{A}=\begin{pmatrix}1&2&3\\2&4&1\\4&6&7\end{pmatrix},\quad \boldsymbol{B}=\begin{pmatrix}1&1&1\\2&2&1\\3&3&1\end{pmatrix},\quad \boldsymbol{C}=\begin{pmatrix}1&2&6\\2&5&15\\6&15&46\end{pmatrix}.$$

5.8 用追赶法求解 $\boldsymbol{Ax}=\boldsymbol{b}$,其中

$$\boldsymbol{A}=\begin{pmatrix}2&-1&&&\\-1&2&-1&&\\&-1&2&-1&\\&&-1&2&-1\\&&&-1&2\end{pmatrix},\quad \boldsymbol{b}=\begin{pmatrix}1\\0\\0\\0\\0\end{pmatrix}.$$

5.9 用平方根法求解

$$\begin{pmatrix}15&-4&-2\\-4&10&3\\-2&3&20\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}-2\\3\\5\end{pmatrix}.$$

5.10 用改进的平方根法求解

$$\begin{pmatrix}2&-1&1\\-1&-2&3\\1&3&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}4\\5\\5\end{pmatrix}.$$

5.11 计算 $\mathrm{cond}_2(\boldsymbol{A})$ 和 $\mathrm{cond}_\infty(\boldsymbol{A})$,其中

$$\boldsymbol{A}=\begin{bmatrix}100&99\\99&98\end{bmatrix}.$$

5.12 证明:如果 $\boldsymbol{A}$ 是正交阵,则 $\mathrm{cond}_2(\boldsymbol{A})=1$.

5.13 设 $\boldsymbol{A},\boldsymbol{B}\in\mathbf{R}^{n\times n}$,对矩阵的算子范数,证明

$$\mathrm{cond}(\boldsymbol{AB})\leqslant\mathrm{cond}(\boldsymbol{A})\mathrm{cond}(\boldsymbol{B}).$$

5.14　设方程组 $\boldsymbol{Ax}=\boldsymbol{b}$，其中

$$\boldsymbol{A}=\begin{bmatrix}2 & -1\\ -2 & 1.000\,1\end{bmatrix},\quad \boldsymbol{b}=\begin{bmatrix}-1\\ 1.000\,1\end{bmatrix}.$$

当右端向量 $\boldsymbol{b}$ 有误差 $\delta\boldsymbol{b}=(0,0.000\,1)^{\mathrm{T}}$ 时，引起解向量 $\boldsymbol{x}$ 的误差为 $\delta\boldsymbol{x}$. 试求出 $\dfrac{\|\delta\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty}$ 的上界，并分析这个结果.

5.15　设方程组 $\boldsymbol{Ax}=\boldsymbol{b}$，其中

$$\boldsymbol{A}=\begin{bmatrix}1 & 2\\ 1.000\,1 & 2\end{bmatrix},\boldsymbol{b}=\begin{bmatrix}3\\ 3.000\,1\end{bmatrix},$$

其精确解为 $(1,1)^{\mathrm{T}}$，给 $\boldsymbol{A}$ 一个扰动

$$\delta\boldsymbol{A}=\begin{bmatrix}0 & 0\\ -0.000\,02 & 0\end{bmatrix},$$

引起解的变化为 $\delta\boldsymbol{x}$. 试求出 $\dfrac{\|\delta\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty}$ 的上界.

5.16　设 $\boldsymbol{A}$ 为非奇异矩阵，并且 $\|\boldsymbol{A}^{-1}\|\ \|\delta\boldsymbol{A}\|<1$. 试证 $(\boldsymbol{A}+\delta\boldsymbol{A})^{-1}$ 存在，且有

$$\frac{\|\boldsymbol{A}^{-1}-(\boldsymbol{A}+\delta\boldsymbol{A})^{-1}\|}{\|\boldsymbol{A}^{-1}\|}\leqslant\frac{\|\boldsymbol{A}^{-1}\|\ \|\delta\boldsymbol{A}\|}{1-\|\boldsymbol{A}^{-1}\|\ \|\delta\boldsymbol{A}\|}.$$

5.17　设 $\boldsymbol{A}$ 是 n 阶非奇异矩阵，$\boldsymbol{B}$ 是奇异矩阵，求证：对于算子范数有

$$\|\boldsymbol{A}\|\leqslant\|\boldsymbol{A}-\boldsymbol{B}\|\,\mathrm{cond}(\boldsymbol{A}).$$

数值试验题 5

5.1　设方程组 $\boldsymbol{Ax}=\boldsymbol{b}$，其中

$$\boldsymbol{A}=\begin{pmatrix}0.3\times10^{-15} & 59.14 & 3 & 1\\ 5.291 & -6.13 & -1 & 2\\ 11.2 & 9 & 5 & 2\\ 1 & 2 & 1 & 1\end{pmatrix},\quad \boldsymbol{b}=\begin{pmatrix}59.17\\ 46.78\\ 1\\ 2\end{pmatrix}.$$

分别用不选主元素的三角分解法和列选主元素的三角分解法解方程组并比较计算结果.

5.2　设主方程组 $\boldsymbol{Ax}=\boldsymbol{b}$，其中

$$\boldsymbol{A}=\begin{pmatrix} 1 & x_0 & {x_0}^2 & \cdots & {x_0}^5 \\ 1 & x_1 & {x_1}^2 & \cdots & {x_1}^5 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_5 & {x_5}^2 & \cdots & {x_5}^5 \end{pmatrix},$$

$x_k=1+0.1k, k=0,1,\cdots,5$, $\boldsymbol{b}$ 由相应的矩阵元素计算，使解向量 $\boldsymbol{x}=(1,1,\cdots,1)^{\mathrm{T}}$.

(1) $\boldsymbol{A}$ 不变，对 $\boldsymbol{b}$ 的元素 b_6 加一个扰动 10^{-4}，应用数学软件求解方程组；

(2) $\boldsymbol{b}$ 不变，对 $\boldsymbol{A}$ 的元素 a_{22} 和 a_{66} 分别加一个扰动 10^{-6}，应用数学软件求解方程组；

(3) 对上述两种扰动方程组的解作误差分析.

5.3 给定两个不同的方程组，第 1 个方程组的系数矩阵为著名的 10 阶 Hilbert 矩阵 $\boldsymbol{H}_{10}$，右端项 $\boldsymbol{b}=(1,0,\cdots,0)^{\mathrm{T}}$，且有 $\|\boldsymbol{H}_{10}\|_1=2.93$，$\|\boldsymbol{H}_{10}^{-1}\|_1=1.21\times10^{13}$. 第 2 个方程组是一个下三角方程组，其系数矩阵为 4 阶 Wilkinson 矩阵：

$$A=\begin{pmatrix} 0.914\,3\times10^{-4} & & & \\ 0.876\,2 & 0.715\,6\times10^{-4} & & \\ 0.794\,3 & 0.814\,3 & 0.950\,4\times10^{-4} & \\ 0.801\,7 & 0.612\,3 & 0.716\,5 & 0.712\,3\times10^{-4} \end{pmatrix},$$

右端项 $\boldsymbol{b}=(0.000\,091\,43, 0.876\,271\,56, 1.608\,695\,04, 2.130\,571\,23)^{\mathrm{T}}$，准确解为 $\boldsymbol{x}=(1,1,1,1)^{\mathrm{T}}$，且有 $\|\boldsymbol{A}\|_1=2.13$，$\|\boldsymbol{A}^{-1}\|_1=1.15\times10^{16}$.

(1) 对上面提供的两个方程组，用你掌握的解法求出计算解 $\tilde{\boldsymbol{x}}$，并计算剩余向量；

(2) 对上面两个方程组右端项产生 10^{-7} 的扰动后，分别解方程组；再对系数矩阵和右端项都产生 10^{-7} 的扰动，再分别解方程组. 观察解产生的误差变化情况.

5.4 给定 n 阶方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$，其中

$$\boldsymbol{A}=\begin{pmatrix} 6 & 1 & & & \\ 8 & 6 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 8 & 6 & 1 \\ & & & 8 & 6 \end{pmatrix}, \quad \boldsymbol{b}=\begin{pmatrix} 7 \\ 15 \\ \vdots \\ 15 \\ 14 \end{pmatrix},$$

则方程组有解 $\boldsymbol{x}=(1,1,\cdots,1)^{\mathrm{T}}$.

(1) 对 $n=10$ 和 $n=84$，分别用 Gauss 消去法和列主元素消去法解方程组，并比较计算结果；

(2) 试估计矩阵 $\boldsymbol{A}$ 的条件数.

第 6 章　线性方程组的迭代解法

对于大型线性代数方程组，常用迭代解法. 它是从某些初始向量出发，用设计好的步骤逐次算出近似解向量 $\boldsymbol{x}^{(k)}$，从而得到向量序列$\{\boldsymbol{x}^{(k)}\}$. 一般 $\boldsymbol{x}^{(k+1)}$ 的计算公式是

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{F}_k(\boldsymbol{x}^{(k)},\boldsymbol{x}^{(k-1)},\cdots,\boldsymbol{x}^{(k-m)}),\quad k=0,1,\cdots.$$

称之为多步迭代法. 若 $\boldsymbol{x}^{(k+1)}$ 只与 $\boldsymbol{x}^{(k)}$ 有关，且 $\boldsymbol{F}_k$ 是线性的，即

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{B}_k\boldsymbol{x}^{(k)}+\boldsymbol{f}_k,\quad k=0,1,\cdots,$$

其中 $\boldsymbol{B}_k\in\mathbf{R}^{n\times n}$，称为单步线性迭代法，$\boldsymbol{B}_k$ 称为迭代矩阵. 若 $\boldsymbol{B}_k$ 和 $\boldsymbol{f}_k$ 都与 k 无关，即

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{B}\boldsymbol{x}^{(k)}+\boldsymbol{f},\quad k=0,1,\cdots,$$

称为单步定常线性迭代法. 本章主要讨论具有这种形式的各种迭代方法.

6.1　基本迭代方法

6.1.1　迭代公式的构造

设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$，$\boldsymbol{b}\in\mathbf{R}^n$，$\boldsymbol{A}$ 非奇异，$\boldsymbol{x}\in\mathbf{R}^n$ 满足方程组

$$\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}. \tag{6.1}$$

如果能找到矩阵 $\boldsymbol{B}\in\mathbf{R}^{n\times n}$，向量 $\boldsymbol{f}\in\mathbf{R}^n$，使 $\boldsymbol{I}-\boldsymbol{B}$ 可逆，而且方程组

$$\boldsymbol{x}=\boldsymbol{B}\boldsymbol{x}+\boldsymbol{f} \tag{6.2}$$

的唯一解就是方程组(6.1)的解，则可从(6.2)式构造一个定常的线性迭代公式

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{B}\boldsymbol{x}^{(k)}+\boldsymbol{f}. \tag{6.3}$$

给定初始向量 $\boldsymbol{x}^{(0)}\in\mathbf{R}^n$，由(6.3)式可以产生序列$\{\boldsymbol{x}^{(k)}\}$，若它有极限 $\boldsymbol{x}^*$，显然 $\boldsymbol{x}^*$ 就是(6.1)式和(6.2)式的解.

定义 6.1　若对任意初始向量 $\boldsymbol{x}^{(0)}\in\mathbf{R}^n$，迭代公式(6.3)产生的序列$\{\boldsymbol{x}^{(k)}\}$都有

$$\lim_{k\to\infty}\boldsymbol{x}^{(k)}=\boldsymbol{x}^*,$$

则称迭代法(6.3)是收敛的.

从(6.1)式出发，可以由不同的途径得到各种不同的等价方程组(6.2)，从而得到不同的迭代法(6.3). 例如，设 $\boldsymbol{A}$ 可以分解为 $\boldsymbol{A}=\boldsymbol{M}-\boldsymbol{N}$，其中 $\boldsymbol{M}$ 非奇异，则

由(6.1)式可得

$$\boldsymbol{x}=\boldsymbol{M}^{-1}\boldsymbol{N}\boldsymbol{x}+\boldsymbol{M}^{-1}\boldsymbol{b}.$$

令 $\boldsymbol{B}=\boldsymbol{M}^{-1}\boldsymbol{N}$，$\boldsymbol{f}=\boldsymbol{M}^{-1}\boldsymbol{b}$，就可以得到(6.2)式的形式. 不同的分解方式 $\boldsymbol{A}=\boldsymbol{M}-\boldsymbol{N}$，可得不同的 $\boldsymbol{B}$ 和 $\boldsymbol{f}$，下面给出对应不同分解方式的常用迭代计算公式.

6.1.2 Jacobi 迭代法和 Gauss-Seidel 迭代法

(1) Jacobi 迭代法.

记 $\boldsymbol{A}=(a_{ij})$，可以把 $\boldsymbol{A}$ 分解为

$$\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{L}-\boldsymbol{U}, \tag{6.4}$$

其中 $\boldsymbol{D}=\mathrm{diag}(a_{11},a_{22},\cdots,a_{nn})$，

$$\boldsymbol{L}=-\begin{pmatrix} 0 & & & \\ a_{21} & 0 & & \\ \vdots & \ddots & \ddots & \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{pmatrix},\quad \boldsymbol{U}=-\begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ & 0 & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ & & & 0 \end{pmatrix}.$$

现设 $\boldsymbol{D}$ 非奇异，即 $a_{ii}\neq 0$，$i=1,2,\cdots,n$. 方程组(6.1)等价于

$$\boldsymbol{x}=\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})\boldsymbol{x}+\boldsymbol{D}^{-1}\boldsymbol{b}.$$

由此构造迭代公式

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{B}_{\mathrm{J}}\boldsymbol{x}^{(k)}+\boldsymbol{f}_{\mathrm{J}},\quad k=0,1,2,\cdots. \tag{6.5}$$

其中迭代矩阵 $\boldsymbol{B}_{\mathrm{J}}$ 和向量 $\boldsymbol{f}_{\mathrm{J}}$ 分别为

$$\boldsymbol{B}_{\mathrm{J}}=\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})=\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{A}, \tag{6.6}$$

$$\boldsymbol{f}_{\mathrm{J}}=\boldsymbol{D}^{-1}\boldsymbol{b}. \tag{6.7}$$

称(6.5)为解(6.1)的 Jacobi 迭代法，简称 J 法.

用 J 法计算向量序列 $\{\boldsymbol{x}^{(k)}\}$，要用两组单元存放向量 $\boldsymbol{x}^{(k)}$ 和 $\boldsymbol{x}^{(k+1)}$. 迭代法可以写成分量形式

$$x_i^{(k+1)}=\frac{1}{a_{ii}}\Big(b_i-\sum_{j=1}^{i-1}a_{ij}x_j^{(k)}-\sum_{j=i+1}^{n}a_{ij}x_j^{(k)}\Big),\quad i=1,2,\cdots,n. \tag{6.8}$$

(2) Gauss-Seidel 迭代法.

在 J 法中，计算 $x_i^{(k+1)}$ 时，分量 $x_1^{(k+1)},\cdots,x_{i-1}^{(k+1)}$ 已经算出，所以可考虑对 J 法进行修改. 在每个分量计算出来之后，下一个分量的计算就利用最新的计算结果. 这样，在整个迭代过程中只要使用一组单元存放迭代向量，其分量形式的计算公式为

$$x_i^{(k+1)}=\frac{1}{a_{ii}}\Big(b_i-\sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)}-\sum_{j=i+1}^{n}a_{ij}x_j^{(k)}\Big),\quad i=1,2,\cdots,n. \tag{6.9}$$

这就是 Gauss-Seidel 迭代法，简称 GS 法.

将(6.9)式写成矩阵形式

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{D}^{-1}(\boldsymbol{L}\boldsymbol{x}^{(k+1)}+\boldsymbol{U}\boldsymbol{x}^{(k)}+\boldsymbol{b}),$$

经整理有

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{B}_{\mathrm{GS}}\boldsymbol{x}^{(k)}+\boldsymbol{f}_{\mathrm{GS}},\quad k=0,1,\cdots, \tag{6.10}$$

其中迭代矩阵 $\boldsymbol{B}_{\mathrm{GS}}$和向量 $\boldsymbol{f}_{\mathrm{GS}}$为

$$\boldsymbol{B}_{\mathrm{GS}}=(\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{U}, \tag{6.11}$$

$$\boldsymbol{f}_{\mathrm{GS}}=(\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{b}. \tag{6.12}$$

Jacobi 迭代法和 Gauss-Seidel 迭代法的分量形式适合于计算编程用，它们的矩阵形式适合于研究迭代序列是否收敛等理论分析.

例 6.1　用 J 法和 GS 法分别求解方程组

$$\begin{pmatrix}10 & 3 & 1\\ 2 & -10 & 3\\ 1 & 3 & 10\end{pmatrix}\begin{pmatrix}x_1\\ x_2\\ x_3\end{pmatrix}=\begin{pmatrix}14\\ -5\\ 14\end{pmatrix},$$

其准确解为 $\boldsymbol{x}^{*}=(1,1,1)^{\mathrm{T}}$.

解　用 J 法计算，按(6.8)式有

$$\begin{cases}x_1^{(k+1)}=\dfrac{1}{10}(\quad -3x_2^{(k)}-x_3^{(k)}+14),\\ x_2^{(k+1)}=\dfrac{1}{10}(2x_1^{(k)}\quad +3x_3^{(k)}+5),\\ x_3^{(k+1)}=\dfrac{1}{10}(-x_1^{(k)}-3x_2^{(k)}\quad +14).\end{cases}$$

用 GS 法计算，按(6.9)式有

$$\begin{cases}x_1^{(k+1)}=\dfrac{1}{10}(\quad -3x_2^{(k)}-x_3^{(k)}+14),\\ x_2^{(k+1)}=\dfrac{1}{10}(2x_1^{(k+1)}\quad +3x_3^{(k)}+5),\\ x_3^{(k+1)}=\dfrac{1}{10}(-x_1^{(k+1)}-3x_2^{(k+1)}\quad +14).\end{cases}$$

取 $\boldsymbol{x}^{(0)}=(0,0,0)^{\mathrm{T}}$，J 法迭代 4 次的计算结果是

$$\boldsymbol{x}^{(4)}=(0.990\,6,0.964\,5,0.990\,6)^{\mathrm{T}},$$

$$\|\boldsymbol{x}^{(4)}-\boldsymbol{x}^{*}\|_{\infty}=0.035\,6.$$

GS 法迭代 4 次的计算结果是

$$\boldsymbol{x}^{(4)}=(0.991\,54,0.995\,78,1.002\,1)^{\mathrm{T}},$$

$$\|\boldsymbol{x}^{(4)}-\boldsymbol{x}^{*}\|_{\infty}=0.008\,5.$$

从计算结果看，本例用 GS 法显然比用 J 法收敛快.

6.2 迭代法的收敛性

6.2.1 一般迭代法的收敛性

设 $\boldsymbol{x}^*$ 是方程组(6.2)的解,即 $\boldsymbol{x}^*=\boldsymbol{B}\boldsymbol{x}^*+\boldsymbol{f}$. 该式与(6.3)式相减,并记误差向量 $\boldsymbol{e}^{(k)}=\boldsymbol{x}^{(k)}-\boldsymbol{x}^*$,则有

$$\boldsymbol{e}^{(k+1)}=\boldsymbol{B}\boldsymbol{e}^{(k)},\quad k=0,1,\cdots.$$

由此可推得

$$\boldsymbol{e}^{(k)}=\boldsymbol{B}^k\boldsymbol{e}^{(0)},\tag{6.13}$$

其中 $\boldsymbol{e}^{(0)}=\boldsymbol{x}^{(0)}-\boldsymbol{x}^*$ 与 k 无关. 所以,迭代法(6.3)式收敛就意味着对任意初始向量 $\boldsymbol{x}^{(0)}\in\mathbf{R}^n$,都有

$$\lim_{k\to\infty}\boldsymbol{e}^{(k)}=\lim_{k\to\infty}\boldsymbol{B}^k\boldsymbol{e}^{(0)}=0.$$

下面给出迭代法收敛的充分必要条件.

定理 6.1 设矩阵 $\boldsymbol{B}\in\mathbf{R}^{n\times n}$,则 $\lim\limits_{k\to\infty}\boldsymbol{B}^k=0$ 的充分必要条件是 $\boldsymbol{B}$ 的谱半径 $\rho(\boldsymbol{B})<1$.

证 根据矩阵 Jordan 标准型的结论,对矩阵 $\boldsymbol{B}$,存在非奇异矩阵 $\boldsymbol{P}$,使得

$$\boldsymbol{P}^{-1}\boldsymbol{B}\boldsymbol{P}=\boldsymbol{J}=\operatorname{diag}(\boldsymbol{J}_1,\boldsymbol{J}_2,\cdots,\boldsymbol{J}_r),$$

其中

$$\boldsymbol{J}_i=\begin{pmatrix}\lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1\\ & & & \lambda_i\end{pmatrix}_{n_i\times n_i}.$$

显然,$\boldsymbol{B}=\boldsymbol{P}\boldsymbol{J}\boldsymbol{P}^{-1}$,$\boldsymbol{B}^k=\boldsymbol{P}\boldsymbol{J}^k\boldsymbol{P}^{-1}$,

$$\boldsymbol{J}^k=\operatorname{diag}(\boldsymbol{J}_1^k,\boldsymbol{J}_2^k,\cdots,\boldsymbol{J}_r^k).$$

因此,$\lim\limits_{k\to\infty}\boldsymbol{B}^k=0$ 的充分必要条件是 $\lim\limits_{k\to\infty}\boldsymbol{J}_i^k=0$,$i=1,2,\cdots,r$.

记 $\boldsymbol{J}_i=\lambda_i\boldsymbol{I}+\boldsymbol{E}_i$,则有

$$\boldsymbol{E}_i^{\,k}=\begin{pmatrix}0 & \cdots & 0 & 1 & 0 & \cdots & 0\\ & \ddots & \ddots & 0 & 1 & \ddots & \vdots\\ & & \ddots & \ddots & \ddots & \ddots & 0\\ & & & \ddots & \ddots & \ddots & 1\\ & & & & \ddots & \ddots & 0\\ & & & & & \ddots & \vdots\\ & & & & & & 0\end{pmatrix}_{n_i\times n_i},$$

其中第 1 行的第 $k+1$ 个元素为 1. 于是有

$$\boldsymbol{J}_i^k=(\lambda_i\boldsymbol{I}+\boldsymbol{E}_i)^k=\sum_{j=0}^{k}\mathrm{C}_k^j\lambda_i^{k-j}\boldsymbol{E}_i^j=\sum_{j=0}^{n_i-1}\mathrm{C}_k^j\lambda_i^{k-j}\boldsymbol{E}_i^j$$

$$=\begin{pmatrix}\lambda_i^k & k\lambda_i^{k-1} & \cdots & \mathrm{C}_k^{n_i-1}\lambda_i^{k-n_i+1}\\ & \lambda_i^k & \ddots & \vdots\\ & & \ddots & k\lambda_i^{k-1}\\ & & & \lambda_i^k\end{pmatrix}_{n_i\times n_i},$$

其中，$\boldsymbol{E}_i^0=\boldsymbol{I}$，$\mathrm{C}_k^j=\dfrac{k!}{j!\ (k-j)!}$.

由于 $\lim\limits_{k\to\infty}k^s\lambda^k=0(|\lambda|<1,s\geqslant 0)$，所以 $\lim\limits_{k\to\infty}\boldsymbol{J}_i^k=0$ 的充分必要条件是 $|\lambda_i|<1(i=1,2,\cdots,r)$. 定理得证.

定理 6.2　对于任意的初始向量 $\boldsymbol{x}^{(0)}$ 和右端向量 $\boldsymbol{f}$，解方程组(6.2)的迭代法(6.3)收敛的充分必要条件是 $\rho(\boldsymbol{B})<1$.

证　先证充分性. 设 $\rho(\boldsymbol{B})<1$，则矩阵 $\boldsymbol{I}-\boldsymbol{B}$ 非奇异，方程组(6.2)有唯一解 $\boldsymbol{x}^*$，从而得(6.13)式. 由定理 6.1 知 $\lim\limits_{k\to\infty}\boldsymbol{B}^k=0$，因此，$\lim\limits_{k\to\infty}\boldsymbol{e}^{(k)}=0$，即 $\lim\limits_{k\to\infty}\boldsymbol{x}^{(k)}=\boldsymbol{x}^*$.

再证必要性. 设对任意初始向量 $\boldsymbol{x}^{(0)}$ 和右端向量 $\boldsymbol{f}$，均有 $\lim\limits_{k\to\infty}\boldsymbol{x}^{(k)}=\boldsymbol{x}^*$，则得 $\boldsymbol{x}^*=\boldsymbol{B}\boldsymbol{x}^*+\boldsymbol{f}$，$\boldsymbol{x}^{(k)}-\boldsymbol{x}^*=\boldsymbol{B}^k(\boldsymbol{x}^{(0)}-\boldsymbol{x}^*)$. 因此，$\lim\limits_{k\to\infty}\boldsymbol{B}^k(\boldsymbol{x}^{(0)}-\boldsymbol{x}^*)=0$，由 $\boldsymbol{x}^{(0)}$ 的任意性推出 $\lim\limits_{k\to\infty}\boldsymbol{B}^k=0$，即得 $\rho(\boldsymbol{B})<1$. 定理得证.

例 6.2　判断用 J 法和 GS 法解方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 的收敛性，其中

(1) $\boldsymbol{A}=\begin{pmatrix}1 & -9 & -10\\ -9 & 1 & 5\\ 8 & 7 & 1\end{pmatrix}$；　　(2) $\boldsymbol{A}=\begin{pmatrix}10 & 4 & 5\\ 4 & 10 & 7\\ 5 & 7 & 10\end{pmatrix}$.

解　(1) 按(6.6)式和(6.11)式有

$$\boldsymbol{B}_{\mathrm{J}}=\begin{pmatrix}0 & 9 & 10\\ 9 & 0 & -5\\ -8 & -7 & 0\end{pmatrix},\quad \boldsymbol{B}_{\mathrm{GS}}=\begin{pmatrix}0 & 9 & 10\\ 0 & 81 & 85\\ 0 & -639 & -675\end{pmatrix}.$$

$\boldsymbol{B}_{\mathrm{J}}$ 的特征值为：$\lambda_1=4.1412+3.930\,6\mathrm{i}$，$\lambda_2=4.1412-3.930\,6\mathrm{i}$，$\lambda_3=-8.2825$. $\rho(\boldsymbol{B}_{\mathrm{J}})=8.2825>1$. $\boldsymbol{B}_{\mathrm{GS}}$ 的特征值为：$\lambda_1=0$，$\lambda_2=0.6054$，$\lambda_3=-594.6054$. $\rho(\boldsymbol{B}_{\mathrm{GS}})=594.6054>1$. 因此，两种迭代法均发散.

(2) 按(6.6)式和(6.11)式求得

$$\boldsymbol{B}_{\mathrm{J}}=\begin{pmatrix}0 & -0.4 & -0.5\\ -0.4 & 0 & -0.7\\ -0.5 & -0.7 & 0\end{pmatrix},\quad \boldsymbol{B}_{\mathrm{GS}}=\begin{pmatrix}0 & -0.4 & -0.5\\ 0 & 0.16 & -0.5\\ 0 & 0.088 & 0.6\end{pmatrix}.$$

$\boldsymbol{B}_{\mathrm{J}}$ 的特征值为：$\lambda_1=0.3653$，$\lambda_2=0.7108$，$\lambda_3=-1.0770$．$\rho(\boldsymbol{B}_{\mathrm{J}})=1.0770>1$．$\boldsymbol{B}_{\mathrm{GS}}$的特征值为：$\lambda_1=0$，$\lambda_2=0.3137$，$\lambda_3=0.4463$．$\rho(\boldsymbol{B}_{\mathrm{GS}})=0.4463<1$．因此，J法发散，而 GS 法收敛．

例 6.3　用 J 法和 GS 法解方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 和 $\widetilde{\boldsymbol{D}}\boldsymbol{Ax}=\boldsymbol{b}$ 有相同的敛散性，其中 $\widetilde{\boldsymbol{D}}$ 是非奇异对角阵．

证　设矩阵 $\boldsymbol{A}$ 分解为 $\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{L}-\boldsymbol{U}$，其中 $\boldsymbol{D}$，$\boldsymbol{L}$ 和 $\boldsymbol{U}$ 分别为对角阵、下三角阵和上三角阵．对方程组 $\boldsymbol{Ax}=\boldsymbol{b}$，由(6.6)式和(6.11)式，J 法和 GS 法的迭代矩阵分别为 $\boldsymbol{B}_{\mathrm{J}}=\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})$，$\boldsymbol{B}_{\mathrm{GS}}=(\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{U}$．

由于 $\widetilde{\boldsymbol{D}}\boldsymbol{A}=\widetilde{\boldsymbol{D}}\boldsymbol{D}-\widetilde{\boldsymbol{D}}\boldsymbol{L}-\widetilde{\boldsymbol{D}}\boldsymbol{U}$，因此，方程组 $\widetilde{\boldsymbol{D}}\boldsymbol{Ax}=\boldsymbol{b}$ 对应于 J 法的迭代矩阵为

$$\widetilde{\boldsymbol{B}}_{\mathrm{J}}=(\widetilde{\boldsymbol{D}}\boldsymbol{D})^{-1}(\widetilde{\boldsymbol{D}}\boldsymbol{L}+\widetilde{\boldsymbol{D}}\boldsymbol{U})=\boldsymbol{D}^{-1}\widetilde{\boldsymbol{D}}^{-1}\widetilde{\boldsymbol{D}}(\boldsymbol{L}+\boldsymbol{U})=\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})=\boldsymbol{B}_{\mathrm{J}},$$

对应于 GS 法的迭代矩阵为

$$\widetilde{\boldsymbol{B}}_{\mathrm{GS}}=(\widetilde{\boldsymbol{D}}\boldsymbol{D}-\widetilde{\boldsymbol{D}}\boldsymbol{L})^{-1}\widetilde{\boldsymbol{D}}\boldsymbol{U}=[\widetilde{\boldsymbol{D}}(\boldsymbol{D}-\boldsymbol{L})]^{-1}\widetilde{\boldsymbol{D}}\boldsymbol{U}=(\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{U}=\boldsymbol{B}_{\mathrm{GS}}.$$

即方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 与 $\widetilde{\boldsymbol{D}}\boldsymbol{Ax}=\boldsymbol{b}$ 有相同的 Jacobi 迭代矩阵和相同的 Gauss-Seidel迭代矩阵．因此敛散性相同．

有时实际判别一个迭代法是否收敛，条件 $\rho(\boldsymbol{B})<1$ 是很难检验的．而一些矩阵范数 $\|\boldsymbol{B}\|$ 可以用 $\boldsymbol{B}$ 的元素表示，所以用 $\|\boldsymbol{B}\|<1$ 作为收敛的充分条件较为方便．

定理 6.3　对某种算子范数，若 $\|\boldsymbol{B}\|<1$，则迭代法(6.3)产生的向量序列 $\{\boldsymbol{x}^{(k)}\}$ 收敛于(6.2)式的精确解 $\boldsymbol{x}^*$，且有误差估计式

$$\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|\leqslant\frac{\|\boldsymbol{B}\|}{1-\|\boldsymbol{B}\|}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|,\tag{6.14}$$

$$\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|\leqslant\frac{\|\boldsymbol{B}\|^k}{1-\|\boldsymbol{B}\|}\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(0)}\|.\tag{6.15}$$

证　利用不等式 $\rho(\boldsymbol{B})\leqslant\|\boldsymbol{B}\|$，由 $\|\boldsymbol{B}\|<1$ 知迭代法是收敛的，且 $\lim\limits_{k\to\infty}\boldsymbol{x}^{(k)}=\boldsymbol{x}^*$．由(6.3)式和 $\boldsymbol{x}^*=\boldsymbol{B}\boldsymbol{x}^*+\boldsymbol{f}$ 易得

$$\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*=\boldsymbol{B}(\boldsymbol{x}^{(k)}-\boldsymbol{x}^*),\quad\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}=\boldsymbol{B}(\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}).$$

于是有

$$\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*\|\leqslant\|\boldsymbol{B}\|\,\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|,$$
$$\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\|\leqslant\|\boldsymbol{B}\|\,\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|.$$

由此可得

$$\begin{aligned}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|&=\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k+1)}+\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*\|\\&\leqslant\|\boldsymbol{B}\|\,\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|+\|\boldsymbol{B}\|\,\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|.\end{aligned}$$

因 $1-\|\boldsymbol{B}\|>0$，由上式即得(6.14)式，反复运用

$$\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|=\|\boldsymbol{B}(\boldsymbol{x}^{(k-1)}-\boldsymbol{x}^{(k-2)})\|\leqslant\|\boldsymbol{B}\|\,\|\boldsymbol{x}^{(k-1)}-\boldsymbol{x}^{(k-2)}\|,$$

即可得(6.15)式，定理得证.

(6.14)式说明，若 $\|\boldsymbol{B}\|<1$ 但不接近于 1，则当相邻两次迭代向量 $\boldsymbol{x}^{(k-1)}$ 和 $\boldsymbol{x}^{(k)}$ 很接近时，$\boldsymbol{x}^{(k)}$ 与精确解很靠近. 因此，在实际计算中，用 $\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\|\leqslant\varepsilon$ 作为迭代终止条件是合理的.

对给定的精度要求，由(6.15)式可以得到需要迭代的次数，并且，由(6.15)式可见，$\|\boldsymbol{B}\|$ 越小，序列$\{\boldsymbol{x}^{(k)}\}$收敛越快. 由于 $\|\boldsymbol{B}\|$ 依赖于所选择的范数，而且 $\rho(\boldsymbol{B})\leqslant\|\boldsymbol{B}\|$，我们以 $\rho(\boldsymbol{B})$给出收敛速度的概念.

定义 6.2　称 $R(\boldsymbol{B})=-\ln\rho(\boldsymbol{B})$为迭代法(6.3)的渐近收敛速度.

由此定义可以看出，$\rho(\boldsymbol{B})<1$ 越小，$R(\boldsymbol{B})$就越大.

例 6.4　用 J 法和 GS 法解下列方程组

$$\begin{cases}10x_1-2x_2-2x_3=1,\\-2x_1+10x_2-x_3=0.5,\\-x_1-2x_2+3x_3=1,\end{cases}$$

必收敛，并比较满足 $\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|_\infty\leqslant10^{-5}$的迭代次数.

解　按(6.6)式和(6.11)式有

$$\boldsymbol{B}_{\mathrm{J}}=\begin{pmatrix}0 & 0.2 & 0.2\\0.2 & 0 & 0.1\\\dfrac{1}{3} & \dfrac{2}{3} & 0\end{pmatrix},\quad \boldsymbol{B}_{\mathrm{GS}}=\frac{1}{150}\begin{pmatrix}0 & 30 & 30\\0 & 6 & 21\\0 & 14 & 24\end{pmatrix}.$$

由于 $\|\boldsymbol{B}_{\mathrm{J}}\|_1=\dfrac{13}{15}<1$，$\|\boldsymbol{B}_{\mathrm{GS}}\|_1=\dfrac{1}{2}<1$，所以，J 法和 GS 法必收敛，并且，$\|\boldsymbol{B}_{\mathrm{GS}}\|_1<\|\boldsymbol{B}_{\mathrm{J}}\|_1$，GS 法比 J 法收敛快.

取 $\boldsymbol{x}^{(0)}=(0,0,0)^{\mathrm{T}}$，J 法的计算结果如表 6－1，GS 法的计算结果如表 6－2. 对 J 法有 $\|\boldsymbol{x}^{(15)}-\boldsymbol{x}^{(14)}\|_\infty=10^{-5}$；对 GS 法有 $\|\boldsymbol{x}^{(9)}-\boldsymbol{x}^{(8)}\|=0.4\times10^{-5}$，实际计算结果也表明 GS 法比 J 法收敛快.

表 6－1

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0	0	0
1	0.100 000	0.050 000	0.333 333
2	0.176 667	0.103 333	0.400 000
⋮	⋮	⋮	⋮
13	0.231 069	0.147 041	0.508 362
14	0.231 081	0.147 050	0.508 383
15	0.231 087	0.147 055	0.508 393

表 6－2

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0	0	0
1	0.100 000	0.070 000	0.413 333
2	0.196 667	0.130 667	0.486 000
⋮	⋮	⋮	⋮
7	0.231 071	0.147 048	0.508 389
8	0.231 087	0.147 056	0.508 399
9	0.231 091	0.147 058	0.508 402

6.2.2 Jacobi 迭代法和 Gauss-Seidel 迭代法的收敛性

显然可以利用定理 6.2 和定理 6.3 判定 J 法和 GS 法的收敛性，但其中只有定理 6.3 对 J 法使用比较方便. 对于大型方程组，要求出迭代矩阵 $\boldsymbol{B}_{GS}$ 和谱半径 $\rho(\boldsymbol{B}_J)$ 以及 $\rho(\boldsymbol{B}_{GS})$ 都是不容易的. 下面给出一些容易验证收敛性的充分条件，先讨论对角占优矩阵的性质.

定义 6.3 若 $\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{n\times n}$ 满足

$$|a_{ii}|>\sum_{j=1,j\neq i}^{n}|a_{ij}|,\quad i=1,2,\cdots,n,$$

则称 $\boldsymbol{A}$ 为严格对角占优矩阵. 若满足

$$|a_{ii}|\geqslant\sum_{j=1,j\neq i}^{n}|a_{ij}|,\quad i=1,2,\cdots,n,$$

且其中至少有一个严格不等式成立，则称 $\boldsymbol{A}$ 为弱对角占优矩阵.

定义 6.4 设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$，若存在一个排列阵 $\boldsymbol{P}$，使得

$$\boldsymbol{P}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{P}=\begin{pmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{O} & \boldsymbol{A}_{22}\end{pmatrix},\tag{6.16}$$

其中 $\boldsymbol{A}_{11}$ 和 $\boldsymbol{A}_{22}$ 均为方阵，则称 $\boldsymbol{A}$ 为可约的. 如果不存在排列阵 $\boldsymbol{P}$ 使(6.16)式成立，则称 $\boldsymbol{A}$ 为不可约的.

如下矩阵 $\boldsymbol{A}$ 是可约的，$\boldsymbol{B}$ 是不可约的.

$$\boldsymbol{A}=\begin{pmatrix}5&3&1&2\\0&1&0&3\\3&2&1&4\\0&2&0&3\end{pmatrix},\quad \boldsymbol{B}=\begin{pmatrix}4&-1&&\\-1&4&-1&\\&-1&4&-1\\&&-1&4\end{pmatrix}.$$

因为，对于矩阵 $\boldsymbol{A}$ 有

$$\boldsymbol{P}=\begin{pmatrix}1 & & & \\ & 0 & 1 & \\ & 1 & 0 & \\ & & & 1\end{pmatrix},\quad \boldsymbol{P}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{P}=\begin{pmatrix}5 & 1 & 3 & 2\\ 3 & 1 & 2 & 4\\ 0 & 0 & 1 & 3\\ 0 & 0 & 2 & 3\end{pmatrix}.$$

而对于矩阵 $\boldsymbol{B}$，不存在一个排列阵使(6.16)式成立.

定理 6.4　若 $\boldsymbol{A}=(a_{ij})$ 严格对角占优，则 $a_{ii}\neq 0, i=1,2,\cdots,n$，且 $\boldsymbol{A}$ 非奇异.

证　由严格对角占优矩阵的定义可知 $a_{ii}\neq 0, i=1,2,\cdots,n$，若 $\boldsymbol{A}$ 奇异，则有 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}\neq 0$，使 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{0}$。设 $|x_k|=\|\boldsymbol{x}\|_\infty>0$，则 $\boldsymbol{A}\boldsymbol{x}=0$ 的第 k 个方程为

$$a_{kk}x_k=-\sum_{j=1,j\neq k}^{n}a_{kj}x_j,$$

由此得到

$$|a_{kk}|\leqslant\sum_{j=1,j\neq k}^{n}|a_{kj}|\frac{|x_j|}{|x_k|}\leqslant\sum_{j=1,j\neq k}^{n}|a_{kj}|,$$

这与严格对角占优矛盾，定理得证.

定理 6.5　若 $\boldsymbol{A}=(a_{ij})$ 为不可约弱对角占优阵，则 $a_{ii}\neq 0, i=0,1,\cdots,n$，且 $\boldsymbol{A}$ 非奇异.

证　若有某个 $a_{kk}=0$，由 $\boldsymbol{A}$ 的弱对角占优性质可知 $\boldsymbol{A}$ 的第 k 行元素均为零. 交换 $\boldsymbol{A}$ 的第 k 行和第 n 行，并交换 $\boldsymbol{A}$ 的第 k 列和第 n 列，就得到(6.16)式的形式，这与 $\boldsymbol{A}$ 的不可约性质矛盾，故 $a_{ii}\neq 0, i=1,2,\cdots,n$.

如果 $\boldsymbol{A}$ 是奇异的，则存在 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}\neq 0$，使 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{0}$，下面分两种情况考虑.

若 $|x_1|=|x_2|=\cdots=|x_n|\neq 0$，由 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{0}$ 的第 k 个方程有

$$|a_{kk}|\leqslant\sum_{j=1,j\neq k}^{n}|a_{kj}|,k=1,2,\cdots,n,$$

这与 $\boldsymbol{A}$ 的弱对角占优性相矛盾.

若 $|x_i|(i=1,2,\cdots,n)$ 不全相等，记 $J=\{k:|x_k|\geqslant|x_i|,i=1,2,\cdots,n\}$，显然 J 非空，J 的补集也非空. 若有 $k\in J$ 和 $m\notin J$，使得 $a_{km}\neq 0$，则由 $|\frac{x_m}{x_k}|<1$ 得知

$$|a_{kk}|\leqslant\sum_{j=1,j\neq k}^{n}|a_{kj}|\frac{|x_j|}{|x_k|}<\sum_{j=1,j\neq k}^{n}|a_{kj}|,$$

这与 $\boldsymbol{A}$ 的弱对角占优性相矛盾，因此

$$a_{km}=0,\quad \forall k\in J,\quad \forall m\notin J,$$

这又导致与 $\boldsymbol{A}$ 的不可约性相矛盾.

故在以上两种情况下，齐次方程组 $\boldsymbol{Ax}=\boldsymbol{0}$ 只有零解，所以 $\boldsymbol{A}$ 非奇异，定理得证.

以上两个定理说明，若 $\boldsymbol{A}$ 为严格对角占优或不可约弱对角占优阵，则 J 法和 GS 法都可以计算. 在这种情况下迭代法的收敛性有如下定理.

定理 6.6 若 $\boldsymbol{A}$ 为严格对角占优矩阵，或为不可约的弱对角占优矩阵，则解方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的 J 法和 GS 法均收敛.

证 设 $\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{L}-\boldsymbol{U}$，这里只给出 $\boldsymbol{A}$ 为严格对角占优阵时的证明.

对 J 法，迭代矩阵 $\boldsymbol{B}_{\mathrm{J}}=\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})$，易得

$$\|\boldsymbol{B}_{\mathrm{J}}\|_{\infty}=\max_{1\leqslant i\leqslant n}\sum_{j=1,j\neq i}\frac{|a_{ij}|}{|a_{ii}|}.$$

由 $\boldsymbol{A}$ 的严格对角占优性，得到 $\|\boldsymbol{B}_{\mathrm{J}}\|_{\infty}<1$，所以 J 法收敛.

对 GS 法，迭代矩阵 $\boldsymbol{B}_{\mathrm{GS}}=(\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{U}$，这里 $\det(\boldsymbol{D}-\boldsymbol{L})^{-1}=\prod_{i=1}^{n}a_{ii}^{-1}\neq 0$. 由于

$$\begin{aligned}\det(\lambda\boldsymbol{I}-\boldsymbol{B}_{\mathrm{GS}})&=\det(\lambda\boldsymbol{I}-(\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{U})\\&=\det(\boldsymbol{D}-\boldsymbol{L})^{-1}\det(\lambda(\boldsymbol{D}-\boldsymbol{L})-\boldsymbol{U}),\end{aligned}$$

我们只要证明方程 $\det(\lambda(\boldsymbol{D}-\boldsymbol{L})-\boldsymbol{U})=0$ 的根 λ，满足 $|\lambda|<1$. 用反证法，假设 $|\lambda|\geqslant 1$，则由 $\boldsymbol{A}$ 的严格对角占优性有

$$\begin{aligned}|\lambda||a_{ii}|&>\sum_{j=1,j\neq i}^{n}|\lambda||a_{ij}|\\&\geqslant\sum_{j=1}^{i-1}|\lambda||a_{ij}|+\sum_{j=i+1}^{n}|a_{ij}|,\quad i=1,2,\cdots,n.\end{aligned}$$

这说明矩阵

$$\lambda(\boldsymbol{D}-\boldsymbol{L})-\boldsymbol{U}=\begin{pmatrix}\lambda a_{11}&a_{12}&\cdots&a_{1n}\\\lambda a_{21}&\lambda a_{22}&\cdots&a_{2n}\\\vdots&\vdots&&\vdots\\\lambda a_{n1}&\lambda a_{n2}&\cdots&\lambda a_{nn}\end{pmatrix}$$

是严格对角占优阵，因此 $\det(\lambda(\boldsymbol{D}-\boldsymbol{L})-\boldsymbol{U})\neq 0$. 这说明只有当 $|\lambda|<1$ 时，才能使 $\det(\lambda(\boldsymbol{D}-\boldsymbol{L})-\boldsymbol{U})=0$. 从而有 $\rho(\boldsymbol{B}_{\mathrm{GS}})<1$，GS 法收敛，定理得证.

由定理 6.6 的证明可见，矩阵 $\boldsymbol{A}$ 严格对角占优等价于 $\|\boldsymbol{B}_{\mathrm{J}}\|_{\infty}<1$. 因此，由定理 6.6 又可知，若 $\|\boldsymbol{B}_{\mathrm{J}}\|_{\infty}<1$，则相应的 GS 法也收敛.

由例 6.1 所给的系数矩阵是严格对角占优的，由例 6.4 所给的系数矩阵是不可约弱对角占优的，所以用 J 法和 GS 法解对应的方程组都收敛.

6.3　超松弛迭代法

在很多情况下，J 法和 GS 法收敛较慢，所以考虑 GS 法的改进. 设计算第 $k+1$ 个近似解 $\boldsymbol{x}^{(k+1)}$ 时，分量 $x_1^{(k+1)}, x_2^{(k+1)}, \cdots, x_{i-1}^{(k+1)}$ 已经算好，按 GS 法给出辅助量

$$\bar{x}_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right),$$

再用参数 ω 将 $x_i^{(k)}$ 与 $\bar{x}_i^{(k+1)}$ 做加权平均，即

$$x_i^{(k+1)} = \omega\bar{x}_i^{(k+1)} + (1-\omega)x_i^{(k)} = x_i^{(k)} + \omega(\bar{x}_i^{(k+1)} - x_i^{(k)}),$$

经整理得

$$x_i^{(k+1)} = x_i^{(k)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij}x_j^{(k)}\right), \quad i=1,2,\cdots,n. \tag{6.17}$$

称此式为逐次超松弛迭代法，简记为 SOR(Successive Over-Relaxation)法，其中 ω 称为松弛因子. 当 $\omega=1$ 时，(6.17)式就是 GS 法.

记 $\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{L}-\boldsymbol{U}$，(6.17)式可写成矩阵形式

$$\boldsymbol{x}^{(k+1)} = (1-\omega)\boldsymbol{x}^{(k)} + \omega\boldsymbol{D}^{-1}(\boldsymbol{b}+\boldsymbol{L}\boldsymbol{x}^{(k+1)}+\boldsymbol{U}\boldsymbol{x}^{(k)}),$$

再整理得

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{L}_\omega\boldsymbol{x}^{(k)} + \omega(\boldsymbol{D}-\omega\boldsymbol{L})^{-1}\boldsymbol{b}, \tag{6.18}$$

其中，迭代矩阵为

$$\boldsymbol{L}_\omega = (\boldsymbol{D}-\omega\boldsymbol{L})^{-1}((1-\omega)\boldsymbol{D}+\omega\boldsymbol{U}). \tag{6.19}$$

例 6.5　方程组

$$\begin{pmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 24 \\ 30 \\ -24 \end{pmatrix}.$$

的准确解为 $(3,4,-5)^{\mathrm{T}}$，如果用 $\omega=1$ 的 SOR 迭代法(即 GS 法)，计算公式是

$$\begin{cases} x_1^{(k+1)} = -0.75x_2^{(k)} + 6, \\ x_2^{(k+1)} = -0.75x_1^{(k+1)} + 0.25x_3^{(k)} + 7.5, \\ x_3^{(k+1)} = 0.25x_2^{(k+1)} - 6. \end{cases}$$

如果用 $\omega=1.25$ 的 SOR 迭代法，计算公式是

$$\begin{cases} x_1^{(k+1)} = -0.25x_1^{(k)} - 0.937\,5x_2^{(k)} + 7.5, \\ x_2^{(k+1)} = -0.937\,5x_1^{(k+1)} - 0.25x_2^{(k)} + 0.312\,5x_3^{(k)} + 9.375, \\ x_3^{(k+1)} = 0.312\,5x_2^{(k+1)} - 0.25x_3^{(k)} - 7.5. \end{cases}$$

取 $\boldsymbol{x}^{(0)}=(1,1,1)^{\mathrm{T}}$,迭代 7 次,则 $\omega=1$ 时得

$$\boldsymbol{x}^{(7)}=(3.013\,411\,0,3.988\,824\,1,-5.002\,794\,0)^{\mathrm{T}},$$

$\omega=1.25$ 时得

$$\boldsymbol{x}^{(7)}=(3.000\,049\,8,4.000\,258\,6,-5.000\,348\,6)^{\mathrm{T}}.$$

若继续算下去,要达到 7 位数字精确度,$\omega=1$ 时,要迭代 34 次,而 $\omega=1.25$ 时,只需要迭代 14 次,显然选 $\omega=1.25$ 收敛要快些.

按一般迭代法收敛的理论,SOR 迭代法收敛的充分必要条件是 $\rho(\boldsymbol{L}_\omega)<1$,而 $\rho(\boldsymbol{L}_\omega)$与松弛因子 ω 有关. 下面讨论松弛因子 ω 在什么范围内取值,SOR 迭代法才可能收敛.

定理 6.7 如果解方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的 SOR 法收敛,则有 $0<\omega<2$.

证 设 $\boldsymbol{L}_\omega$ 的特征值为 $\lambda_1,\lambda_2,\cdots,\lambda_n$,则

$$\begin{aligned}\det(\boldsymbol{L}_\omega)&=\det(\boldsymbol{D}-\omega\boldsymbol{L})^{-1}\det((1-\omega)\boldsymbol{D}+\omega\boldsymbol{U})\\&=\det\boldsymbol{D}^{-1}\det((1-\omega)\boldsymbol{D})=(1-\omega)^n.\end{aligned}$$

由于 SOR 法收敛,所以有

$$|1-\omega|=|\det(\boldsymbol{L}_\omega)|^{\frac{1}{n}}=|\lambda_1\lambda_2\cdots\lambda_n|^{\frac{1}{n}}\leqslant\rho(\boldsymbol{L}_\omega)<1,$$

由此得 $0<\omega<2$,定理得证.

该定理说明,只有当松弛因子 ω 在区间$(0,2)$内取值时,SOR 法才可能收敛. 下面给出 SOR 法收敛的充分条件.

定理 6.8 如果 $\boldsymbol{A}$ 为对称正定矩阵,且 $0<\omega<2$,则解 $\boldsymbol{Ax}=\boldsymbol{b}$ 的 SOR 法收敛.

证 设 λ 是 $\boldsymbol{L}_\omega$ 的一个特征值,对应特征向量 $\boldsymbol{x}$. 由(6.19)式可得

$$((1-\omega)\boldsymbol{D}+\omega\boldsymbol{U})\boldsymbol{x}=\lambda(\boldsymbol{D}-\omega\boldsymbol{L})\boldsymbol{x},$$

这里,$\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{L}-\boldsymbol{U}$ 是实对称矩阵,所以有 $\boldsymbol{L}^{\mathrm{T}}=\boldsymbol{U}$. 上式两边与 $\boldsymbol{x}$ 作内积得

$$(1-\omega)(\boldsymbol{Dx},\boldsymbol{x})+\omega(\boldsymbol{Ux},\boldsymbol{x})=\lambda((\boldsymbol{Dx},\boldsymbol{x})-\omega(\boldsymbol{Lx},\boldsymbol{x})).\tag{6.20}$$

因为 $\boldsymbol{A}$ 正定,$\boldsymbol{D}$ 亦正定,记 $p=(\boldsymbol{Dx},\boldsymbol{x})$,有 $p>0$. 又记$(\boldsymbol{Lx},\boldsymbol{x})=\alpha+\mathrm{i}\beta$,则有

$$(\boldsymbol{Ux},\boldsymbol{x})=(\boldsymbol{x},\boldsymbol{Lx})=\overline{(\boldsymbol{Lx},\boldsymbol{x})}=\alpha-\mathrm{i}\beta.$$

由(6.20)式有

$$\lambda=\frac{(1-\omega)p+\omega\alpha-\mathrm{i}\omega\beta}{p-\omega\alpha-\mathrm{i}\omega\beta},$$

$$|\lambda|^2=\frac{(p-\omega(p-\alpha))^2+\omega^2\beta^2}{(p-\omega\alpha)^2+\omega^2\beta^2}.$$

因 $\boldsymbol{A}$ 正定,$(\boldsymbol{Ax},\boldsymbol{x})=p-2\alpha>0$,$0<\omega<2$,所以$(p-\omega(p-\alpha))^2-(p-\omega\alpha)^2=p\omega(2-\omega)(2\alpha-p)<0$,即$|\lambda|^2<1$,从而 $\rho(\boldsymbol{L}_\omega)<1$,SOR 方法收敛. 定理得证.

当 $\omega=1$ 时,SOR 法就是 GS 法,所以上面的定理说明,当系数矩阵是对称正

定矩阵时，GS 法收敛.

对于例 6.5 所给的方程组，其系数矩阵是对称正定的，因此对 $\omega=1$ 和 $\omega=1.25$的 SOR 迭代法都收敛.

例 6.6　设矩阵 $\boldsymbol{A}$ 非奇异，求证用 GS 法求解方程组 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 时是收敛的.

证　对 $\boldsymbol{x}\neq 0$，由 $\boldsymbol{A}$ 非奇异知 $\boldsymbol{A}\boldsymbol{x}\neq 0$，从而

$$(\boldsymbol{A}\boldsymbol{x},\boldsymbol{A}\boldsymbol{x})=(\boldsymbol{A}\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x})=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}>0,$$

即 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ 是对称正定的，因此，用 GS 法求解方程组 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 时收敛.

对于超松弛迭代法，自然希望能找到最优松弛因子 ω_{opt}，使对应 ω_{opt}的 SOR 方法收敛最快. 对于一类有特殊性质的矩阵(即所谓 2 -循环的和相容次序的矩阵，它们常在偏微分方程的数值解法中出现)，有关 ω_{opt}的理论在 20 世纪 50 年代已得到. 因为证明较复杂，这里只叙述其结果，即

$$\omega_{\mathrm{opt}}=\frac{2}{1+\sqrt{1-\mu^2}},$$

其中 $\mu=\rho(\boldsymbol{B}_{\mathrm{J}})$是 J 法迭代矩阵 $\boldsymbol{B}_{\mathrm{J}}$ 的谱半径.

可以证明，对称正定的三对角矩阵满足最优松弛因子 ω_{opt}的条件. 在实际应用中，一般地说计算 $\rho(\boldsymbol{B}_{\mathrm{J}})$较困难. 对某些微分方程数值解问题，可考虑用求特征值的近似值的方法，也可由计算实践摸索出近似最佳松弛因子.

例 6.7　用 SOR 迭代法解方程组

$$\begin{pmatrix} 10 & -1 & -2 \\ -1 & 10 & -2 \\ -1 & -1 & 5 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}=\begin{pmatrix} 7.2 \\ 8.3 \\ 4.2 \end{pmatrix}.$$

解　取 $\boldsymbol{x}^{(0)}=(0,0,0)^{\mathrm{T}}$，迭代公式为

$$\begin{cases} x_1^{(k+1)}=x_1^{(k)}+\omega(0.72-x_1^{(k)}+0.1x_2^{(k)}+0.2x_3^{(k)}), \\ x_2^{(k+1)}=x_2^{(k)}+\omega(0.83+0.1x_1^{(k+1)}-x_2^{(k)}+0.2x_3^{(k)}), \\ x_3^{(k+1)}=x_3^{(k)}+\omega(0.84+0.2x_1^{(k+1)}+0.2x_2^{(k+1)}-x_3^{(k)}). \end{cases}$$

对 ω 取不同值，计算结果满足

$$\| x^{(k)}-x^{*} \|_{\infty}\leqslant 10^{-5}$$

的迭代次数如表 6 - 3，这里，准确解为 $\boldsymbol{x}^{*}=(1.1,1.2,1.3)^{\mathrm{T}}$.

表 6 - 3

ω	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
k	163	77	49	34	26	20	15	12	9	6
ω	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	
k	6	8	10	13	17	22	31	51	105	

从表 6－3 可见，本例的最佳松弛因子应在 1 与 1.1 之间，当 $\omega=1.055$ 时，计算结果为表 6－4.

表 6－4

k	0	1	2	3	4	5
$x_1^{(k)}$	0	0.759 6	1.082 02	1.100 88	1.099 98	1.1
$x_2^{(k)}$	0	0.955 788	1.200 59	1.199 89	1.200 05	1.2
$x_3^{(k)}$	0	1.248 15	1.299 18	1.300 21	1.3	1.3

6.4 分块迭代法

前面所讨论的迭代法，一次只计算一个分量. 要完成一次迭代，需要逐个地计算迭代解向量中的每一个分量，直到算出全部分量的值. 然后再进行下一次迭代，使得解向量达到计算精确度为止. 通常，称这种迭代法为点迭代法.

下面介绍更一般的迭代法，其基本思想是将方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 中的 $\boldsymbol{A}$ 分块，将 $\boldsymbol{x}$ 和 $\boldsymbol{b}$ 也进行相应地分块，然后将每个子块视为一个元素，并按照点迭代法类似地进行迭代，称这种迭代法为块迭代法. 下面给出具体描述.

设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 可写成分块形式

$$\boldsymbol{A}=\begin{pmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1r}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{A}_{2r}\\ \vdots & \vdots & & \vdots\\ \boldsymbol{A}_{r1} & \boldsymbol{A}_{r2} & \cdots & \boldsymbol{A}_{rr}\end{pmatrix},$$

其中 $\boldsymbol{A}_{ii}$ 为 $n_i\times n_i$ 的方阵，$n_1+n_2+\cdots+n_r=n$. 向量 $\boldsymbol{x}$ 和 $\boldsymbol{b}$ 也相应地进行分块.

$$\boldsymbol{x}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_r)^{\mathrm{T}},\quad \boldsymbol{b}=(\boldsymbol{b}_1,\boldsymbol{b}_2,\cdots,\boldsymbol{b}_r)^{\mathrm{T}},$$

其中 $\boldsymbol{x}_i$ 和 $\boldsymbol{b}_i$ 都是 n_i 维的向量. 令 $\boldsymbol{A}=\boldsymbol{D}_B-\boldsymbol{L}_B-\boldsymbol{U}_B$，其中 $\boldsymbol{D}_B=\mathrm{diag}(\boldsymbol{A}_{11},\boldsymbol{A}_{22},\cdots,\boldsymbol{A}_{rr})$，

$$\boldsymbol{L}_B=-\begin{pmatrix}\boldsymbol{O} & & & \\ \boldsymbol{A}_{21} & \boldsymbol{O} & & \\ \vdots & \ddots & \ddots & \\ \boldsymbol{A}_{r1} & \cdots & \boldsymbol{A}_{r,r-1} & \boldsymbol{O}\end{pmatrix},\quad \boldsymbol{U}_B=\begin{pmatrix}\boldsymbol{O} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1r}\\ & \boldsymbol{O} & \ddots & \vdots\\ & & \ddots & \boldsymbol{A}_{r-1,r}\\ & & & \boldsymbol{O}\end{pmatrix}.$$

类似于点迭代法，可分别得到求解方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 的块 Jacobi 迭代法

$$\boldsymbol{A}_{ii}\boldsymbol{x}_i^{(k+1)}=\boldsymbol{b}_i-\sum_{j=1,j\neq i}^{r}\boldsymbol{A}_{ij}\boldsymbol{x}_j^{(k)},\quad i=1,2,\cdots,r. \tag{6.21}$$

块 Gauss-Seidel 迭代法

$$\boldsymbol{A}_{ii}\boldsymbol{x}_i^{(k+1)} = \boldsymbol{b}_i - \sum_{j=1}^{i-1}\boldsymbol{A}_{ij}\boldsymbol{x}_j^{(k+1)} - \sum_{j=i+1}^{r}\boldsymbol{A}_{ij}\boldsymbol{x}_j^{(k)}, \quad i = 1,2,\cdots,r. \tag{6.22}$$

块超松弛法

$$\boldsymbol{A}_{ii}\boldsymbol{x}_i^{(k+1)} = \boldsymbol{A}_{ii}\boldsymbol{x}_i^{(k)} + \omega\left(\boldsymbol{b}_i - \sum_{j=1}^{i-1}\boldsymbol{A}_{ij}\boldsymbol{x}_j^{(k+1)} - \sum_{j=i}^{r}\boldsymbol{A}_{ij}\boldsymbol{x}_j^{(k)}\right), \quad i = 1,2,\cdots,r. \tag{6.23}$$

在实际计算中，对每个 i，(6.21)—(6.23)式都分别是 $n_i \times n_i$ 的方程组，一般用直接方法求解. 对大型方程组的情形，n 是大数，而 n_i 相对是较小的. 当 $n_1 = n_2 = \cdots = n_r = 1$ 时，就是点迭代法.

对于块迭代法，也有相应于点迭代法的收敛性判定定理. 在偏微分方程数值解中，常常会遇到特殊形状的分块矩阵.

评　注

本章介绍了解线性代数方程组的迭代法的一些基本理论及 Jacobi 迭代法、Gauss-Seidel 迭代法和 SOR 迭代法，这 3 种迭代法都是一阶定常迭代法. 它们的理论在 20 世纪 50 年代已经形成. 其他如加速收敛的方法等没有叙述.

在计算机大规模集成电路设计、结构分析、网络理论、电力分布系统、图论，特别是数值求解多维偏微分方程组中，常常会遇到大规模的稀疏的线性代数方程组(其系数矩阵是非零元素占很小百分比的稀疏矩阵)，这时，常常用点迭代法或块迭代法. 迭代法有存贮空间小、程序简单等特点，在使用时，能保持系数矩阵的稀疏性不变.

迭代法的收敛性和收敛速度是使用的关键问题，实际使用的应该是收敛快的方法. 通常，Gauss-Seidel 迭代法要比 Jacobi 迭代法收敛快. SOR 方法的松弛因子如果选择适当，则收敛更快. 一些特殊类型的方程组，松弛因子的选择已有成熟的方法或经验，此时，SOR 方法就用得更多. 迭代法的收敛性与系数矩阵的性质有密切的关系，一些具有特殊性质的矩阵的应用在实际工作中也是很重要的.

习　题　6

6.1　给定方程组：

(1) $\begin{pmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \\ 1 & 2 & -3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}$；　(2) $\begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 10 \\ 1 \end{pmatrix}$.

试证明:对方程组(1),J 法收敛而 GS 法不收敛;对方程组(2),J 法不收敛而 GS 法收敛.

6.2 设方程组

$$\begin{cases} a_{11}x_1+a_{12}x_2=b_1, \\ a_{21}x_1+a_{22}x_2=b_2, \end{cases}$$

其中 $a_{11}a_{22}\neq 0$. 求 J 法收敛的充要条件.

6.3 给定方程组

$$\begin{pmatrix} 8 & -1 & 1 \\ 2 & 10 & -1 \\ 1 & 1 & -5 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}=\begin{pmatrix} 1 \\ 4 \\ 3 \end{pmatrix},$$

试判别 J 法和 GS 法的收敛性. 若收敛,取初迭代向量 $\boldsymbol{x}^{(0)}=(0,0,0)^{\mathrm{T}}$,求满足 $\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\|_{\infty}<10^{-3}$ 的解.

6.4 证明矩阵

$$\boldsymbol{A}=\begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix}$$

对于 $-0.5<a<1$ 是正定的,而 J 法只对 $-0.5<a<0.5$ 是收敛的.

6.5 给定迭代过程 $\boldsymbol{x}^{(k+1)}=\boldsymbol{C}\boldsymbol{x}^{(k)}+\boldsymbol{g}$,其中 $\boldsymbol{C}\in\mathbf{R}^{n\times n}$,$k=0,1,\cdots$. 试证明:如果矩阵 $\boldsymbol{C}$ 的特征值 $\lambda_i(\boldsymbol{C})=0(i=1,2,\cdots,n)$,则此迭代过程最多迭代 n 次就收敛于方程组的解.

6.6 用 SOR 法解方程组(取 $\omega=0.9$)

$$\begin{cases} 5x_1+2x_2+x_3=-12, \\ -x_1+4x_2+2x_3=20, \\ 2x_1-3x_2+10x_3=3. \end{cases}$$

要求当 $\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\|_{\infty}<10^{-4}$ 时迭代终止.

6.7 用 SOR 法解方程组(分别取松弛因子 $\omega=1.03$,$\omega=1$,$\omega=1.1$)

$$\begin{cases} 4x_1-x_2=1, \\ -x_1+4x_2-x_3=4, \\ -x_2+4x_3=-3. \end{cases}$$

要求 $\|\boldsymbol{x}^{*}-\boldsymbol{x}^{(k)}\|_{\infty}<0.5\times 10^{-5}$,对每一个 ω 确定迭代次数.

6.8 设有方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$,$\boldsymbol{A}$ 为对称正定阵,其特征值 $\lambda(\boldsymbol{A})\leqslant\beta$. 证明迭代公式

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}+\omega(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}^{(k)}),\quad k=0,1,\cdots,$$

当 $0<\omega<\dfrac{2}{\beta}$ 时收敛.

6.9　设 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 为 n 阶矩阵，$\boldsymbol{A}$ 非奇异. 考虑方程组

$$\boldsymbol{A}\boldsymbol{z}_1+\boldsymbol{B}\boldsymbol{z}_2=\boldsymbol{b}_1,\quad \boldsymbol{B}\boldsymbol{z}_1+\boldsymbol{A}\boldsymbol{z}_2=\boldsymbol{b}_2,$$

其中 $\boldsymbol{z}_1,\boldsymbol{z}_2,\boldsymbol{b}_1,\boldsymbol{b}_2\in\mathbf{R}^n$.

(1) 找出下述迭代方法收敛的充要条件：

$$\boldsymbol{A}\boldsymbol{z}_1^{(m+1)}=\boldsymbol{b}_1-\boldsymbol{B}\boldsymbol{z}_2^{(m)},\quad \boldsymbol{A}\boldsymbol{z}_2^{(m+1)}=\boldsymbol{b}_2-\boldsymbol{B}\boldsymbol{z}_1^{(m)},\quad m=0,1,\cdots;$$

(2) 找出下述迭代法收敛的充要条件：

$$\boldsymbol{A}\boldsymbol{z}_1^{(m+1)}=\boldsymbol{b}_1-\boldsymbol{B}\boldsymbol{z}_2^{(m)},\quad \boldsymbol{A}\boldsymbol{z}_2^{(m+1)}=\boldsymbol{b}_2-\boldsymbol{B}\boldsymbol{z}_1^{(m+1)},\quad m=0,1,\cdots.$$

6.10　用块 Gauss-Seidel 迭代法求解方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$，其中

$$\boldsymbol{A}=\begin{pmatrix} 1 & 0 & -\frac{1}{4} & -\frac{1}{4} \\ 0 & 1 & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & 1 & 0 \\ -\frac{1}{4} & -\frac{1}{4} & 0 & 1 \end{pmatrix},\quad \boldsymbol{b}=\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

取初始迭代向量 $\boldsymbol{x}^{(0)}=\boldsymbol{b}$，直到 $\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\|_\infty<10^{-3}$.

数值试验题 6

6.1　讨论用 Jacobi 迭代法求解方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 的收敛性，其中，$\boldsymbol{A}$ 为如下的 100 阶的方阵，$\boldsymbol{b}$ 为如下的 100 维的列向量，并要求写出简单的理由：

$$\boldsymbol{A}=\begin{pmatrix} \frac{1}{1}+\frac{1}{2} & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3}+\frac{1}{2} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ & \vdots & \vdots & & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n-1}+\frac{1}{2} \end{pmatrix},$$

$$\boldsymbol{b}=\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

6.2　设有方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$，其中 $\boldsymbol{A}\in\mathbf{R}^{20\times 20}$，

$$\boldsymbol{A}=\begin{pmatrix} 3 & -0.5 & -0.25 & & & & \\ -0.5 & 3 & -0.5 & \ddots & & & \\ -0.25 & -0.5 & \ddots & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots & -0.5 & -0.25 \\ & & & \ddots & -0.5 & 3 & -0.5 \\ & & & & -0.25 & -0.5 & 3 \end{pmatrix}.$$

(1) 选取不同的初始向量 $\boldsymbol{x}^{(0)}$ 和不同的右端向量 $\boldsymbol{b}$，给定迭代误差要求，用 J 法和 GS 法计算，观测得出的迭代向量序列是否均收敛. 若收敛，记录迭代次数，分析计算结果并得出你的结论.

(2) 取定初始向量 $\boldsymbol{x}^{(0)}$ 和右端向量 $\boldsymbol{b}$，如取 $\boldsymbol{x}^{(0)}=\boldsymbol{0}$，$\boldsymbol{b}=\boldsymbol{A}\boldsymbol{e}$，$\boldsymbol{e}=(1,1,\cdots,1)^{\mathrm{T}}$. 将 $\boldsymbol{A}$ 的主对角线元素成倍增长若干次，非主对角线元素不变，每次用 J 法计算，要求迭代误差满足 $\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\|_{\infty}<10^{-6}$，比较收敛速度，分析现象并得出你的结论.

6.3 用 SOR 法解方程组

$$\begin{pmatrix} 5 & -1 & -1 & -1 & -1 \\ -1 & 5 & -1 & -1 & -1 \\ -1 & -1 & 5 & -1 & -1 \\ -1 & -1 & -1 & 5 & -1 \\ -1 & -1 & -1 & -1 & 5 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}=\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

其精确解是 $\boldsymbol{x}^{*}=(1,1,1,1,1)^{\mathrm{T}}$. 取不同的松弛因子 ω，要求每次迭代计算的误差满足 $\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\|_{\infty}<10^{-6}$，记录迭代次数，得出最佳松弛因子.

第 7 章　非线性方程和方程组的数值解法

一般的非线性方程组可写成 $\boldsymbol{F}(\boldsymbol{x})=0$，其中 $\boldsymbol{F}$ 和 $\boldsymbol{x}$ 都是 n 维向量，或写成

$$f_i(x_1,x_2,\cdots,x_n)=0,\quad i=1,2,\cdots,n.$$

其中，$f_1,f_2,\cdots,f_n$ 中至少有一个是 $x_1,x_2,\cdots,x_n$ 的非线性函数. 当 $n=1$ 时，就是单个的方程 $f(x)=0$. 非线性方程和方程组的求解是工程和科学领域中最常见的问题.

与线性方程组不同，除特殊情况外，求解非线性方程不能用直接法求数值解，而是要用迭代法. 迭代法的基本问题是收敛性、收敛速度和计算效率.

对于线性方程组，如前所述，若某迭代法收敛，则取任何初值都收敛. 但是，对于非线性方程，不同的初值可能有不同的收敛性态，有的初值使迭代收敛，有的则不收敛. 一般来说，为使迭代法收敛，初值应取在解的附近.

我们先详细讨论单个方程的情形，其中有一类是形如

$$a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0=0$$

的代数方程. 当 $n\geqslant 5$ 时，其根是不能用加、减、乘、除和开方的有限次运算公式表示的，所以代数方程的解法主要是迭代法.

方程的数值解法的收敛性，也与方程根的重数有关. 对于一般的函数 $f(x)\in C[a,b]$，若有

$$f(x)=(x-x^*)^mg(x),\quad g(x^*)\neq 0,$$

其中 m 为正整数，我们称 x^* 是 $f(x)$ 的 m 重零点，或称 x^* 是方程 $f(x)=0$ 的 m 重根. 显然，若 x^* 是 $f(x)$ 的 m 重零点，且 $g(x)$ 充分光滑，则有

$$f(x^*)=f'(x^*)=\cdots=f^{(m-1)}(x^*)=0,\quad f^{(m)}(x^*)\neq 0.$$

当 m 为奇数时，$f(x)$ 在点 x^* 处变号；当 m 为偶数时，$f(x)$ 在点 x^* 处不变号.

7.1　方程求根的二分法

设 $f(x)\in C[a,b]$，若 $f(a)f(b)<0$，则方程 $f(x)=0$ 在 $[a,b]$ 内至少有一个根，称 $[a,b]$ 为方程的有根区间. 通过缩小有根区间，我们可以形成方程求根的数值解法. 下面给出方程求根的二分法.

设 $[a_0,b_0]=[a,b]$ 为有根区间，$[a_n,b_n]$ 的中点是 x_n，若有 $f(a_n)f(x_n)<0$，则令新的有根区间 $[a_{n+1},b_{n+1}]=[a_n,x_n]$；若 $f(a_n)f(x_n)>0$，则令 $[a_{n+1},b_{n+1}]$

$=[x_n,b_n]$. 这样，若反复二分下去，即可得出一系列有根区间

$$[a,b]\supset[a_1,b_1]\supset[a_2,b_2]\supset\cdots\supset[a_n,b_n]\supset\cdots,$$

其中每个区间都是前一个区间的一半. 因此，$[a_k,b_k]$的长度

$$b_k-a_k=\frac{b-a}{2^k}\to 0 \quad (k\to\infty).$$

由此可见，如果二分过程能无限地继续下去，这些区间最终必收敛于一点 x^*，该点显然就是所求的根.

每次二分后，设取有根区间$[a_k,b_k]$的中点 $x_k=\frac{a_k+b_k}{2}$作为根的近似值，则在二分过程中可以获得一个近似根的序列$\{x_k\}$，该序列必以根 x^* 为极限.

在实际计算时，我们不可能完成这个无限过程，其实也没有这种必要，因为数值分析的结果允许带有一定的误差. 由于

$$|x^*-x_k|\leqslant\frac{b_k-a_k}{2}=\frac{b-a}{2^{k+1}},$$

只要二分足够多次(即 k 充分大)，则有$|x^*-x_k|<\varepsilon$，这里 ε 为预定的精度.

例 7.1 求方程 $f(x)=x^3-x-1=0$ 在区间$[1,1.5]$内的一个实根，要求准确到小数点后的第 2 位.

解 这里 $a=1,b=1.5,f(a)<0,f(b)>0$. 取$[a,b]$的中点 $x_0=1.25$，将区间$[a,b]$二等分，由于 $f(x_0)<0$，即 $f(a)$与 $f(x_0)$同号，故在 x_0 的右侧有方程的一个实根，这时，令 $a_1=x_0=1.25,b_1=b=1.5$，而新的有根区间为$[a_1,b_1]$. 二分过程可如此反复下去，计算结果如表 7－1.

表 7－1

k	0	1	2	3	4	5	6
a_k	1	1.25	1.25	1.312 5	1.312 5	1.312 5	1.320 3
b_k	1.5	1.5	1.375	1.375	1.343 8	1.328 1	1.328 1
x_k	1.25	1.375	1.312 5	1.343 8	1.328 1	1.320 3	1.324 2
$f(x_k)$	−	+	−	+	+	−	−

为了预估达到要求的二分的次数，令$\frac{b-a}{2^{k+1}}\leqslant 0.005$ 可得 $k\geqslant 6$，即二分 6 次就能达到预定的精度$|x^*-x_6|\leqslant 0.005$，与实际计算结果相符.

上述二分法的优点是算法简单，而且在有根区间内，收敛性总能得到保证. 值得注意的是，为了求出足够精确的近似解，往往需要计算很多次函数值，是一种收敛较慢的方法，通常用来求根的粗略近似值，把它作为后面要讨论的迭代法

的初始值. 另一方面，二分法只适用于求一元方程的奇数重实根.

在二分法中，是逐次将有根区间折半. 更一般地是，从有根区间的左端点出发，按预定的步长 h 一步一步地向右跨，每跨一步进行一次根的“搜索”，即检查所在节点上的函数值的符号，一旦发现其与左端的函数值异号，则可确定一个缩小了的有根区间，其宽度等于预定的步长 h. 然后，再对新的有根区间，取新的更小的预定步长，继续“搜索”，直到有根区间的宽度足够小. 称这种方法为逐步搜索法.

7.2　一元方程的不动点迭代法

7.2.1　不动点迭代法及其收敛性

设一元函数 $f(x)$ 是连续的，为了求一元非线性方程

$$f(x)=0 \tag{7.1}$$

的实根，先将它转换成等价形式

$$x=\varphi(x), \tag{7.2}$$

其中 $\varphi(x)$ 是一个连续函数. 然后构造迭代公式

$$x_{k+1}=\varphi(x_k),\quad k=0,1,\cdots. \tag{7.3}$$

对于给定的初始值 x_0，若由此迭代公式生成的序列 $\{x_k\}$ 的极限存在，$\lim\limits_{k\to\infty}x_k=x^*$，则有 $x^*=\varphi(x^*)$，即 x^* 满足方程(7.2)，从而按等价性，x^* 也是方程(7.1)的根.

迭代式(7.3)称为基本迭代法，$\varphi(x)$ 称为迭代函数，x^* 称为 $\varphi(x)$ 的不动点，(7.3)式也称为不动点迭代法. 迭代过程中，x_{k+1} 仅由 x_k 决定，因此，这是一种单步法.

把(7.1)式转换成等价形式(7.2)的方法很多，迭代函数的不同选择对应不同的迭代法，它们的收敛性可能有很大的差异. 当方程有多个解时，同一迭代法的不同初值，也可能收敛到不同的根. 举例说明如下.

例 **7.2**　求 $f(x)=x^3-x-1=0$ 的一个实根.

解　把 $f(x)=0$ 转换成两种等价形式

$$x=\varphi_1(x)=\sqrt[3]{x+1},\quad x=\varphi_2(x)=x^3-1,$$

对应的迭代法分别为

$$x_{k+1}=\sqrt[3]{x_k+1},\quad x_{k+1}={x_k}^3-1,\quad k=0,1,\cdots.$$

由于 $f(1)=-1$，$f(2)=5$，即连续函数 $f(x)$ 在区间[1,2]内变号，从而[1,2]为

有限区间.取它的中点为初值,即令 $x_0=1.5$,迭代结果列于表 7-2.此方程有唯一实根 $x^*=1.324\ 717\ 957\ 244\ 75$.显然,第一个迭代法收敛,第二个迭代法发散.

表 7-2

k	0	1	2	…	11
$\varphi_1(x_k)$	1.5	1.357 208 81	1.330 860 96	…	1.324 717 96
$\varphi_2(x_k)$	1.5	2.375 000 00	12.396 484 4	…	$\to+\infty$

例 7.3 求 $f(x)=x^2-2=0$ 的根.

解 把 $f(x)=0$ 转换成等价形式

$$x=\varphi(x)=\frac{1}{2}\left(x+\frac{2}{x}\right),$$

对应的迭代法为

$$x_{k+1}=\frac{1}{2}\left(x_k+\frac{2}{x_k}\right),\quad k=0,1,\cdots.$$

取初值 $x_0=\pm1$,迭代结果分别收敛到 $x^*=\pm\sqrt{2}$,计算结果如表 7-3 所示.

表 7-3

k	0	1	2	3	4	5
x_k	1	1.5	1.416 666 67	1.414 215 69	1.414 213 56	1.414 213 56
x_k	−1	−1.5	−1.416 666 67	−1.414 215 69	−1.414 213 56	−1.414 213 56

由此可见,基本迭代法的收敛性质取决于迭代函数 $\varphi(x)$和初值 x_0 的选取.下面给出迭代法(7.3)的收敛性基本定理.

定理 7.1 设函数 $\varphi(x)$在闭区间$[a,b]$上连续,并且满足

(1) 对任意 $x\in[a,b]$,有 $\varphi(x)\in[a,b]$;

(2) 存在正数 $L<1$,使对任意 $x,y\in[a,b]$,有

$$|\varphi(x)-\varphi(y)|\leqslant L|x-y|. \tag{7.4}$$

则对方程(7.2)有

(1) 函数 $\varphi(x)$在闭区间$[a,b]$上存在唯一的不动点 x^*;

(2) 对于任何初值 $x_0\in[a,b]$,由迭代法(7.3)生成的序列$\{x_k\}$收敛到不动点 x^*;

(3) 有误差估计式

$$|x_k-x^*|\leqslant\frac{L}{1-L}|x_k-x_{k-1}|. \tag{7.5}$$

证　令 $\psi(x)=x-\varphi(x)$，则由 $\varphi(x)\in[a,b]$ 知，$\varphi(a)\leqslant 0$，$\varphi(b)\geqslant 0$. 因为 $\varphi(x)$ 是连续函数，故它在 $[a,b]$ 上有零点，即 $\varphi(x)$ 在 $[a,b]$ 上有不动点 x^*. 若 $\varphi(x)$ 在 $[a,b]$ 上有两个相异的不动点 x_1^*，x_2^*，则有

$$|x_1^*-x_2^*|=|\varphi(x_1^*)-\varphi(x_2^*)|\leqslant L|x_1^*-x_2^*|<|x_1^*-x_2^*|.$$

这是个矛盾式子，因此 $\varphi(x)$ 在 $[a,b]$ 上只有一个不动点.

显然有 $x_k\in[a,b]$，$k=0,1,\cdots$，进而

$$|x_k-x^*|=|\varphi(x_{k-1})-\varphi(x^*)|\leqslant L|x_{k-1}-x^*|\leqslant\cdots\leqslant L^k|x_0-x^*|.$$

从而 $\lim\limits_{k\to\infty}|x_k-x^*|=0$，即 $\lim\limits_{k\to\infty}x_k=x^*$.

显然有

$$|x_{k+1}-x_k|=|\varphi(x_k)-\varphi(x_{k-1})|\leqslant L|x_k-x_{k-1}|,$$

进而，对任何正整数 p，同理可得

$$\begin{aligned}|x_{k+p}-x_k|&\leqslant|x_{k+p}-x_{k+p-1}|+\cdots+|x_{k+2}-x_{k+1}|+|x_{k+1}-x_k|\\&\leqslant(L^{p-1}+\cdots+L+1)|x_{k+1}-x_k|.\end{aligned}$$

因为 $0<L<1$，从而 $(1-L)^{-1}=\sum\limits_{k=0}^{\infty}L^k>1+L+\cdots+L^{p-1}$，

$$|x_{k+p}-x_k|\leqslant\frac{1}{1-L}|x_{k+1}-x_k|\leqslant\frac{L}{1-L}|x_k-x_{k-1}|.$$

令 $p\to+\infty$，由收敛性即得(7.5)式，定理得证.

如果函数 $\varphi(x)$ 在区间 (a,b) 内可导，那么定理 7.1 中的条件(2)可用更强的条件

$$|\varphi'(x)|\leqslant L<1,\quad\forall x\in(a,b)\tag{7.6}$$

代替. 事实上，若上式成立，则由微分中值定理，对任何 $x,y\in[a,b]$ 都有

$$|\varphi(x)-\varphi(y)|=|\varphi'(\xi)(x-y)|\leqslant L|x-y|,$$

其中 ξ 在 x 与 y 之间，从而条件(7.4)式成立.

由估计式(7.5)可知，只要相邻两次计算结果的偏差 $|x_k-x_{k-1}|$ 足够小，且 L 不很接近 1，即可保证近似值 x_k 具有足够的精度. 因此，可以通过检查 $|x_k-x_{k-1}|$ 的大小来判断迭代过程是否终止. 并且，由(7.5)式有

$$|x_k-x^*|\leqslant\frac{L^k}{1-L}|x_1-x_0|.\tag{7.7}$$

如果能恰当地估计出 L 的值，则由(7.7)式，我们可对给定的精度确定出需要迭代的次数.

函数 $\varphi(x)$ 的不动点 x^*，在几何上是直线 $y=x$ 与曲线 $y=\varphi(x)$ 的交点的横坐标. 因此，迭代过程(7.3)式的几何解释如图 7-1 所示，其中图(a)是收敛的情形，图(b)是发散的情形.

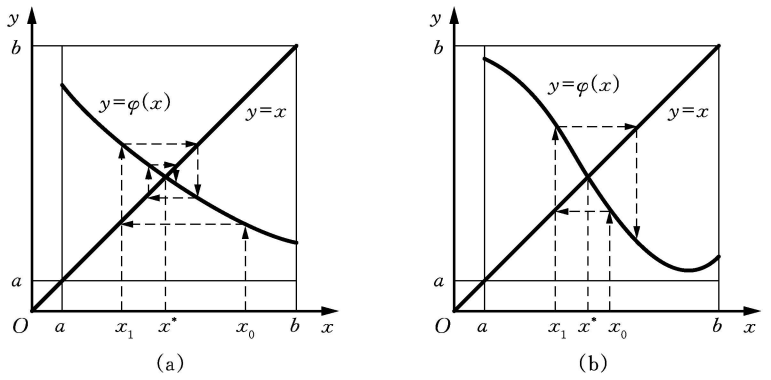

图 7－1

例 7.4 对于例 7.2 中的两种迭代法，讨论它们的收敛性.

解 对于迭代函数 $\varphi_1(x)=\sqrt[3]{x+1}$，其导数 $\varphi_1'(x)=\dfrac{1}{3}(x+1)^{-\frac{2}{3}}$. 容易验证，对任意 $x\in[1,2]$有

$$\varphi_1(x)\in[1.26,1.45]\subset[1,2],\quad \varphi_1'(x)\leqslant 0.21<1.$$

因此，对于任何初值 $x_0\in[1,2]$，由 $\varphi_1(x)$给出的迭代法都收敛到区间$[1,2]$上的唯一不动点 x^*.

对于迭代函数 $\varphi_2(x)=x^3-1$，其导数 $\varphi_2'(x)=3x^2$. 显然，对 $x\in[1,2]$有 $\varphi_2(x)\in[0,7]$，$\varphi_2'(x)>1$，不满足定理 7.1 的条件. 从几何上可以说明，只要初值 $x_0\neq x^*$，该迭代法发散.

有时，对于一些不满足定理 7.1 的条件的问题，可以通过转化，化为适合于迭代的形式. 这要针对具体情况进行讨论.

例 7.5 已知 $x=\varphi(x)$的 $\varphi'(x)$满足$|\varphi'(x)-3|<1$，试问如何利用 $\varphi(x)$构造一个收敛的简单迭代函数？

解 由 $x=\varphi(x)$，可得

$$x-3x=\varphi(x)-3x,$$

即可得等价的方程

$$x=\frac{1}{2}(3x-\varphi(x)).$$

因此，令

$$\psi(x)=\frac{1}{2}(3x-\varphi(x)),$$

则有

$$|\psi'(x)|=\frac{1}{2}|3-\varphi'(x)|<\frac{1}{2}.$$

因此，迭代式 $x_{k+1}=\psi(x_k)(k=0,1,\cdots)$ 收敛.

7.2.2　局部收敛性和加速收敛法

由于定理 7.1 讨论的是迭代法在区间 $[a,b]$ 上的收敛性，因而，可以称之为全局收敛性定理. 全局收敛性也包括在无穷区间上收敛的情形. 但一般来说，全局收敛性的情形不易检验. 所以常常讨论在根 x^* 附近的收敛性问题. 为此，给出如下定义.

定义 7.1　设 x^* 是 $\varphi(x)$ 的不动点，若存在 x^* 的一个邻域 $S(x^*,\delta)=[x^*-\delta,x^*+\delta]$，$\delta>0$，使得对任何初值 $x_0\in S(x^*,\delta)$，由迭代法(7.3)生成的序列满足 $\{x_k\}\subset S(x^*,\delta)$，且收敛到 x^*，则称迭代法(7.3)是局部收敛的.

定理 7.2　设 x^* 是 $\varphi(x)$ 的一个不动点，$\varphi'(x)$ 在 x^* 的某个邻域上连续，并且有 $|\varphi'(x^*)|<1$，则迭代法(7.3)局部收敛.

证　因为 $\varphi'(x)$ 在 x^* 连续，且 $|\varphi'(x^*)|<1$，所以存在 x^* 的一个闭邻域 $[x^*-\delta,x^*+\delta]$，在其上 $|\varphi'(x)|\leqslant L<1$，并且有

$$|\varphi(x)-x^*|=|\varphi(x)-\varphi(x^*)|\leqslant L|x-x^*|<\delta,$$

即对一切 $x\in[x^*-\delta,x^*+\delta]$，有 $\varphi(x)\in[x^*-\delta,x^*+\delta]$. 根据定理 7.1，对任意 $x_0\in[x^*-\delta,x^*+\delta]$，迭代法(7.3)收敛. 定理得证.

上述定理称为局部收敛性定理，它给出了局部收敛的一个充分条件. 当迭代收敛时，收敛的快慢用下述收敛阶来衡量.

定义 7.2　设序列 $\{x_k\}$ 收敛到 x^*，记误差 $e_k=x_k-x^*$. 若存在实数 $p\geqslant 1$ 和 $c\neq 0$，使得

$$\lim_{k\to\infty}\frac{e_{k+1}}{e_k^{\,p}}=c,\tag{7.8}$$

则称序列 $\{x_k\}$ 是 p 阶收敛的. 当 $p=1$ 时，称为线性收敛；当 $p>1$ 时，称为超线性收敛；当 $p=2$ 时，称为平方收敛.

(7.8)式表明，当 $k\to\infty$ 时，e_{k+1} 是 e_k 的 p 阶无穷小量，因此，阶数 p 越大，收敛越快. 如果是线性收敛的，(7.8)式中的常数满足 $0<|c|\leqslant 1$.

如果在定理 7.2 中，还有 $\varphi'(x^*)\neq 0$，即 $\varphi'(x^*)$ 满足 $0<|\varphi'(x^*)|<1$，则对 $x_0\neq x^*$，必有 $x_k\neq x^*$，$k=1,2,\cdots$，而且

$$e_{k+1}=x_{k+1}-x^*=\varphi(x_k)-\varphi(x^*)=\varphi'(\xi_k)e_k,$$

其中 ξ_k 在 x_k 与 x^* 之间. 于是

$$\lim_{k\to\infty}\frac{e_{k+1}}{e_k}=\lim_{k\to\infty}\varphi'(\xi_k)=\varphi'(x^*)\neq 0.$$

从而，在这种情况下，$\{x_k\}$是线性收敛的. 可见，提高收敛阶的一个途径是选择迭代函数 $\varphi(x)$，使它满足 $\varphi'(x^*)=0$. 下面给出整数阶超线性收敛的一个充分条件.

定理 7.3 设 x^* 是 $\varphi(x)$的一个不动点. 若有正整数 $p\geqslant 2$，使得 $\varphi^{(p)}(x)$在 x^* 的邻域上连续，并且满足

$$\varphi'(x^*)=\varphi''(x^*)=\cdots=\varphi^{(p-1)}(x^*)=0,\quad \varphi^{(p)}(x^*)\neq 0, \tag{7.9}$$

则由迭代法(7.3)生成的序列$\{x_k\}$在 x^* 的邻域是 p 阶收敛的，且有

$$\lim_{k\to\infty}\frac{e_{k+1}}{e_k^{\,p}}=\frac{\varphi^{(p)}(x^*)}{p!}. \tag{7.10}$$

证 因 $\varphi'(x^*)=0$，由定理 7.2 知迭代法(7.3)是局部收敛的. 取充分接近 x^* 的 x_0，设 $x_0\neq x^*$，有 $x_k\neq x^*$，$k=1,2,\cdots$. 由 Taylor 展开式有

$$\begin{aligned}x_{k+1}=\varphi(x_k)=&\varphi(x^*)+\varphi'(x^*)(x_k-x^*)+\cdots\\&+\frac{\varphi^{(p-1)}(x^*)}{(p-1)!}(x_k-x^*)^{p-1}+\frac{\varphi^{(p)}(\xi_k)}{p!}(x_k-x^*)^p,\end{aligned}$$

其中 ξ_k 在 x_k 与 x^* 之间. 由(7.9)式有

$$x_{k+1}-x^*=\frac{\varphi^{(p)}(\xi_k)}{p!}(x_k-x^*)^p.$$

由 $\varphi^{(p)}(x)$的连续性可得(7.10)式. 定理得证.

对于线性收敛的迭代法，收敛很慢，所以要在这些迭代法的基础上考虑加速收敛的方法. 设$\{x_k\}$线性收敛到 x^*，则迭代误差 $e_k=x_k-x^*$ 满足

$$\lim_{k\to\infty}\frac{e_{k+1}}{e_k}=\lim_{k\to\infty}\frac{x_{k+1}-x^*}{x_k-x^*}=c\neq 0.$$

因此，当 k 充分大时有

$$\frac{x_{k+1}-x^*}{x_k-x^*}\approx\frac{x_{k+2}-x^*}{x_{k+1}-x^*},$$

从中解出 x^* 得

$$x^*\approx x_{k+2}-\frac{(x_{k+2}-x_{k+1})^2}{x_{k+2}-2x_{k+1}+x_k}.$$

所以，我们在计算了 x_k，x_{k+1}和 x_{k+2}之后，可以用上式右端作为 x_{k+2}的一个修正值. 这样，我们可将迭代法 $x_{k+1}=\varphi(x_k)$改造成下述过程，称为 Steffensen 迭代法：

$$\begin{cases}y_k=\varphi(x_k),\quad z_k=\varphi(y_k),\\ x_{k+1}=z_k-\dfrac{(z_k-y_k)^2}{z_k-2y_k+x_k},\quad k=0,1,\cdots.\end{cases} \tag{7.11}$$

它的不动点迭代形式是

$$x_{k+1}=\psi(x_k),\quad k=0,1,\cdots, \tag{7.12}$$

其中的迭代函数为

$$\psi(x)=\frac{x\varphi(\varphi(x))-\varphi^2(x)}{\varphi(\varphi(x))-2\varphi(x)+x}=x-\frac{(\varphi(x)-x)^2}{\varphi(\varphi(x))-2\varphi(x)+x}. \tag{7.13}$$

例 7.6　求方程 $f(x)=xe^x-1=0$ 的根.

解　此方程等价于 $x=\varphi(x)=e^{-x}$. 由函数 $y=x$ 和 $y=e^{-x}$ 的图形可以看出,$\varphi(x)$只有一个不动点 $x^*>0$. 因为对任何 $x>0$,都有 $0<|\varphi'(x)|=e^{-x}<1$,所以迭代法 $x_{k+1}=e^{-x_k}$ 线性收敛. 取初始值 $x_0=0.5$,迭代结果列于表 7-4. 准确解是 $x^*=0.567\ 143\ 290\ 409\ 78\cdots$,可见线性收敛的速度是很慢的.

表 7-4

k	0	1	…	28	29
x_k	0.5	0.606 530 660	…	0.567 143 282	0.567 143 295

如果使用 Steffensen 迭代法,仍取初值 $x_0=0.5$,则

$$y_k=e^{-x_k},\quad z_k=e^{-y_k},$$

$$x_{k+1}=z_k-\frac{(z_k-y_k)^2}{z_k-2y_k+x_k},\quad k=0,1,\cdots.$$

计算结果列于表 7-5. 与表 7-4 相比,可见 Steffensen 迭代法比原方法收敛快得多,仅迭代 4 次就达到了原方法迭代 29 次的结果.

表 7-5

k	0	1	2	3	4
x_k	0.5	0.567 623 876	0.567 143 314	0.567 143 290	0.567 143 290

定理 7.4　设函数 $\psi(x)$由 $\varphi(x)$按(7.13)式定义.

(1) 若 x^* 是 $\varphi(x)$的不动点,$\varphi'(x)$在 x^* 处连续,且 $\varphi'(x^*)\neq1$,则 x^* 也是 $\psi(x)$的不动点;反之,若 x^* 是 $\psi(x)$的不动点,则 x^* 也是 $\varphi(x)$的不动点;

(2) 若 x^* 是 $\varphi(x)$的不动点,$\varphi''(x)$在 x^* 处连续,且 $\varphi'(x^*)\neq1$,则 Steffensen迭代法(7.11)至少具有二阶局部收敛性.

证　(1) 若 $x^*=\varphi(x^*)$,则当 $x=x^*$ 时,(7.13)式的分子分母都为零. 对它的极限用 L'Hospitale 法则,由于 $\varphi'(x^*)\neq1$,得知

$$\begin{aligned}\lim_{x\to x^*}\psi(x)&=\lim_{x\to x^*}\frac{\varphi(\varphi(x))+x\varphi'(\varphi(x))\varphi'(x)-2\varphi(x)\varphi'(x)}{\varphi'(\varphi(x))\varphi'(x)-2\varphi'(x)+1}\\&=\frac{x^*[\varphi'(x^*)-1]^2}{[\varphi'(x^*)-1]^2}=x^*,\end{aligned}$$

从而 $x^*=\psi(x^*)$. 反之，若 $x^*=\psi(x^*)$，则由(7.13)式得知 $x^*=\varphi(x^*)$.

(2) 由(1)可知 x^* 是 $\psi(x)$ 的不动点，于是，由定理 7.3，只要证明 $\psi'(x^*)=0$. 对(7.13)式两边求导得

$$1-\psi'(x)=\frac{p(x)}{q(x)}, \tag{7.14}$$

其中

$$\begin{aligned} p(x)=&2(\varphi(x)-x)(\varphi'(x)-1)(\varphi(\varphi(x))-2\varphi(x)+x)\\ &-(\varphi(x)-x)^2(\varphi'(\varphi(x))\varphi'(x)-2\varphi'(x)+1),\\ q(x)=&(\varphi(\varphi(x))-2\varphi(x)+x)^2, \end{aligned}$$

并且容易算出

$$p''(x^*)=q''(x^*)=2(\varphi'(x^*)-1)^4.$$

于是，由 $\varphi'(x^*)\neq 1$，可知 $p''(x^*)=q''(x^*)\neq 0$. 对(7.14)式的两边求极限，因为 x^* 至少是 $p(x)$和 $q(x)$的二重根，所以，使用两次 L'Hospitale 法则得

$$1-\psi'(x^*)=\lim_{x\to x^*}(1-\psi'(x))=\lim_{x\to x^*}\frac{p''(x)}{q''(x)}=1,$$

从而 $\psi'(x^*)=0$. 定理得证.

可见，在定理 7.4 的条件下，不管原迭代法 $x_{k+1}=\varphi(x_k)$收敛还是不收敛，由它构成的 Steffensen 迭代法(7.11)至少平方收敛. 因此，Steffensen 迭代法是对原迭代法的一种改善. 关于原迭代法不收敛的情形，举例如下.

例 7.7 用 Steffensen 迭代法求方程 $f(x)=x^3-x-1=0$ 的实根.

解 由例 7.4 可知，迭代法 $x_{k+1}=x_k{}^3-1$ 发散. 现用 $\varphi_2(x)=x^3-1$ 构造 Steffensen 迭代法.

$$y_k=x_k{}^3-1,\quad z_k=y_k{}^3-1,$$

$$x_{k+1}=z_k-\frac{(z_k-y_k)^2}{z_k-2y_k+x_k}.$$

仍取初值 $x_0=1.5$，计算结果如表 7-6. 可见，Steffensen 迭代法对这种不收敛的情形同样有效.

表 7-6

k	0	1	…	5	6
x_k	1.5	1.416 292 97	…	1.324 717 99	1.324 717 96

7.3　一元方程的常用迭代法

7.3.1　Newton 迭代法

设 x^* 是方程 $f(x)=0$ 的实根，x_k 是一个近似根 $x_k\approx x^*$，由 Taylor 展开式有

$$0=f(x^*)=f(x_k)+f'(x_k)(x^*-x_k)+\frac{f''(\xi)}{2}(x^*-x_k)^2,$$

这里假设 $f''(x)$存在并连续. 若 $f'(x_k)\neq 0$，可得

$$x^*=x_k-\frac{f(x_k)}{f'(x_k)}-\frac{f''(\xi)}{2f'(x_k)}(x^*-x_k)^2, \tag{7.15}$$

其中 ξ 在 x^* 与 x_k 之间. 若(7.15)式的右端最后一项忽略不计，作为 x^* 新的一个近似值，就有

$$x_{k+1}=x_k-\frac{f(x_k)}{f'(x_k)},\quad k=0,1,\cdots, \tag{7.16}$$

这就是 Newton 迭代法.

对(7.16)式可作如下的几何解释：x_{k+1} 为函数 $f(x)$在点 x_k 处的切线与横坐标轴的交点，见图 7－2. 因此，Newton 迭代法也称为切线法.

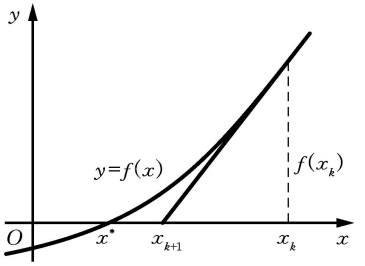

图 7－2

将(7.16)式写成一般的不动点迭代(7.3)式的形式，有

$$\varphi(x)=x-\frac{f(x)}{f'(x)},\quad \varphi'(x)=\frac{f(x)f''(x)}{[f'(x)]^2}.$$

所以有 $\varphi'(x^*)=0(f'(x^*)\neq 0)$，Newton 迭代法是超线性收敛的. 更准确地，从(7.15)式和(7.16)式可得下面的定理.

定理 7.5　设 $f(x^*)=0$，$f'(x^*)\neq 0$，且 $f(x)$在包含 x^* 的一个区间上有 2 阶连续导数，若 Newton 迭代法收敛于 x^*，则至少 2 阶收敛，并且

$$\lim_{k\to\infty}\frac{x_{k+1}-x^*}{(x_k-x^*)^2}=\frac{f''(x^*)}{2f'(x^*)}.$$

以上讨论的是 Newton 法的局部收敛性. 对于某些非线性方程,Newton 法具有全局收敛性.

例 7.8 设 $a>0$,对方程 $x^2-a=0$,试证:取任何初值 $x_0>0$,Newton 迭代法都收敛到算术根 $\sqrt{a}$.

证 对 $f(x)=x^2-a$,Newton 迭代法为

$$x_{k+1}=\frac{1}{2}\left(x_k+\frac{a}{x_k}\right),\quad k=0,1,\cdots. \tag{7.17}$$

由此可知

$$x_{k+1}-\sqrt{a}=\frac{1}{2x_k}(x_k{}^2-2x_k\sqrt{a}+a)=\frac{1}{2x_k}(x_k-\sqrt{a})^2,$$

$$x_k-x_{k+1}=\frac{1}{2x_k}(x_k{}^2-a).$$

可见,对于任何 $x_0>0$,都有 $x_k\geqslant\sqrt{a}(k=1,2,\cdots)$,并且 $\{x_k\}$ 非增. 因此,$\{x_k\}$ 是有下界的非增序列,从而有极限 x^*. 对(7.17)式的两边取极限,得到 $(x^*)^2-a=0$,因为 $x_k>0$,故有 $x^*=\sqrt{a}$.

在定理 7.5 中,要求 $f(x^*)=0$,$f'(x)\neq0$,即 x^* 是方程 $f(x)=0$ 的单根时,Newton 法至少具有 2 阶局部收敛性. 下面讨论重根的情形.

设 x^* 是 $f(x)=0$ 的 m 重根,$m\geqslant2$,即

$$f(x)=(x-x^*)^m g(x),\quad g(x^*)\neq0.$$

由 Newton 迭代函数 $\varphi(x)$ 的导数表达式,容易求出

$$\varphi'(x^*)=1-\frac{1}{m}.$$

从而,$0<\varphi'(x^*)<1$. 因此,只要 $f'(x_k)\neq0$,这时的 Newton 迭代法线性收敛.

为了改善重根时 Newton 法的收敛性,有如下两种方法.

若改为取

$$\varphi(x)=x-\frac{mf(x)}{f'(x)}, \tag{7.18}$$

容易验证 $\varphi'(x^*)=0$,迭代至少 2 阶收敛.

若令 $\mu(x)=\dfrac{f(x)}{f'(x)}$,由 x^* 是 $f(x)$ 的 m 重零点,有

$$\mu(x)=\frac{(x-x^*)g(x)}{mg(x)+(x-x^*)g(x)}.$$

所以,x^* 是 $\mu(x)$ 的单零点. 可将 Newton 法的迭代函数修改为

$$\varphi(x)=x-\frac{\mu(x)}{\mu'(x)}=x-\frac{f(x)f'(x)}{[f'(x)]^2-f(x)f''(x)}. \tag{7.19}$$

这种方法也是至少 2 阶收敛的.

例 7.9　方程 $x^4-4x^2+4=0$ 的根 $x^*=\sqrt{2}$是二重根. 用 3 种方法求解.

解　(1) 用 Newton 法有

$$x_{k+1}=x_k-\frac{x_k^{\,2}-2}{4x_k}.$$

(2) 由(7.18)式,$m=2$,迭代公式为

$$x_{k+1}=x_k-\frac{x_k^{\,2}-2}{2x_k}.$$

(3) 由(7.19)式确定的修改方法,迭代公式化简为

$$x_{k+1}=x_k-\frac{x_k(x_k^{\,2}-2)}{x_k^{\,2}+2}.$$

3 种方法均取 $x_0=1.5$,计算结果列于表 7－7. 方法(2)和方法(3)都是二阶方法,x_3 都达到了误差限为 10^{-9}的精确度,而普通的 Newton 法是一阶的,要近 30 次迭代才有相同精度的结果.

表 7－7

x_k	x_0	x_1	x_2	x_3
方法(1)	1.5	1.458 333 333	1.436 607 143	1.425 497 619
方法(2)	1.5	1.416 666 667	1.414 215 686	1.414 213 562
方法(3)	1.5	1.411 764 706	1.414 211 438	1.414 213 562

Newton 法的每步计算都要求提供函数的导数值,当函数 $f(x)$比较复杂时,提供它的导数值往往是有困难的. 此时,在 Newton 迭代法(7.16)中,可用 $f'(x_0)$或常数 D 取代 $f'(x_k)$,迭代公式变为

$$x_{k+1}=x_k-\frac{f(x_k)}{f'(x_0)} \quad 或 \quad x_{k+1}=x_k-\frac{f(x_k)}{D}.$$

这称为简化 Newton 法,其迭代函数为 $\varphi(x)=x-\dfrac{f(x)}{f'(x_0)}$或 $\varphi(x)=x-\dfrac{f(x)}{D}$. 通常 $\varphi'(x^*)\neq 0$,简化 Newton 法一般为线性收敛.

7.3.2　割线法与抛物线法

为了回避导数值 $f'(x_k)$的计算,除了前面的简化 Newton 法之外,我们也可用 x_k,x_{k-1}点上的差商代替 $f'(x_k)$,得到迭代公式

$$x_{k+1}=x_k-\frac{x_k-x_{k-1}}{f(x_k)-f(x_{k-1})}f(x_k),\quad k=1,2,\cdots, \tag{7.20}$$

这就是割线法的计算公式. 其几何解释为通过$(x_k,f(x_k))$和$(x_{k-1},f(x_{k-1}))$作$y=f(x)$的割线,割线与 x 轴交点的横坐标就是 x_{k+1}.

与 Newton 法不同的是,用割线法计算 x_{k+1}时,需要有两个初始值 x_0 和 x_1. 计算 x_{k+1}时,要保留上一步的 $f(x_{k-1})$和 x_{k-1},再计算一次函数值 $f(x_k)$. 所以割线法是一种两步迭代法,不能直接用单步迭代法收敛性分析的结果. 下面给出割线法收敛性的定理.

定理 7.6 设 $f(x^*)=0$,在区间 $\Delta=[x^*-\delta,x^*+\delta]$上 $f(x)$的 2 阶导数连续,且 $f'(x)\neq 0$. 又设 $M\delta<1$,其中

$$M=\frac{\max\limits_{x\in\Delta}|f''(x)|}{2\min\limits_{x\in\Delta}|f'(x)|}. \tag{7.21}$$

则当 $x_0,x_1\in\Delta$ 时,由(7.20)式产生的序列$\{x_k\}\subset\Delta$,并且按 $p=\dfrac{1+\sqrt{5}}{2}\approx 1.618$ 阶收敛到根 x^*.

证 由(7.20)式两边减去 x^*,利用均差的记号有

$$\begin{aligned}x_{k+1}-x^*&=x_k-x^*-\frac{f(x_k)-f(x^*)}{f(x_{k-1},x_k)}=(x_k-x^*)\left(1-\frac{f(x_k,x^*)}{f(x_{k-1},x_k)}\right)\\&=(x_k-x^*)(x_{k-1}-x^*)\frac{f(x_{k-1},x_k,x^*)}{f(x_{k-1},x_k)}.\end{aligned} \tag{7.22}$$

因 $f(x)$有 2 阶连续导数,所以有

$$f(x_{k-1},x_k)=f'(\eta_k),\quad f(x_{k-1},x_k,x^*)=\frac{1}{2}f''(\xi_k),$$

其中 η_k 在 x_{k-1},x_k 之间,ξ_k 在包含 x_{k-1},x_k,x^* 的最小区间上. 仍记 $e_k=x_k-x^*$,由(7.22)式有

$$e_{k+1}=\frac{f''(\xi_k)}{2f'(\eta_k)}e_ke_{k-1}. \tag{7.23}$$

若$|e_{k-1}|<\delta$,$|e_k|<\delta$,则利用(7.21)式和 $M\delta<1$ 得

$$|e_{k+1}|\leqslant M|e_k||e_{k-1}|\leqslant M\delta^2<\delta.$$

这说明 $x_0,x_1\in\Delta$ 时,序列$\{x_k\}\subset\Delta$. 又由于

$$|e_k|\leqslant M|e_{k-1}||e_{k-2}|\leqslant M\delta|e_{k-1}|\leqslant\cdots\leqslant(M\delta)^k|e_0|,$$

所以,当 $k\to\infty$时,$e_k\to 0$,即$\{x_k\}$收敛到 x^*. 从上式也可知割线法至少是一阶收敛的.

进一步确定收敛的阶,这里我们给出一个不严格的证明. 由(7.23)式有

$$|e_{k+1}|\approx M^*|e_k||e_{k-1}|, \tag{7.24}$$

这里 $M^*=\dfrac{|f''(x^*)|}{|2f'(x^*)|}$. 令 $d^{m_k}=M^*|e_k|$，代入(7.24)式得

$$m_{k+1}\approx m_k+m_{k-1},\quad m_0=M^*|e_0|,\quad m_1=M^*|e_1|.$$

我们知道，差分方程 $z_{k+1}=z_k+z_{k-1}$ 的通解为 $z_k=c_1\lambda_1{}^k+c_2\lambda_1{}^k$，这里 c_1，c_2 为任意常数，

$$\lambda_1=\frac{1+\sqrt{5}}{2}\approx 1.618,\quad \lambda_2=\frac{1-\sqrt{5}}{2}\approx -0.618,$$

λ_1 和 λ_2 是方程 $\lambda^2-\lambda-1=0$ 的两个根. 当 k 充分大时，设 $m_k\approx c\lambda_1{}^k$，$c$ 为常数，则有

$$\frac{|e_{k+1}|}{|e_k{}^{\lambda_1}|}=(M^*)^{\lambda_1-1}d^{m_{k+1}-\lambda_1 m_k}\approx(M^*)^{\lambda_1-1}.$$

这说明割线法的收敛阶为 $\lambda_1\approx 1.618$. 定理证毕.

类似于简单 Newton 法，有如下的单点割线法.

$$x_{k+1}=x_k-\frac{x_k-x_0}{f(x_k)-f(x_0)}f(x_k),\quad k=1,2,\cdots,$$

其迭代函数为

$$\varphi(x)=x-\frac{f(x)(x-x_0)}{f(x)-f(x_0)},$$

于是

$$\varphi'(x^*)=1-\frac{f'(x^*)}{f'(\xi)},$$

其中 ξ 在 x_0 与 x^* 之间. 由此可见，单点割线法一般为线性收敛. 但当 $f'(x)$ 变化不大时，$\varphi'(x^*)\approx 0$，收敛仍可能很快.

例 7.10　分别用单点割线法、割线法和 Newton 法求解 Leonardo 方程

$$f(x)=x^3+2x^2+10x-20=0.$$

解　$f'(x)=3x^2+4x+10$，$f''(x)=6x+4$. 由于 $f'(x)>0$，$f(1)=-7<0$，$f(2)=12>0$，故 $f(x)=0$ 在(1,2)内仅有一根. 对于单点割线法和割线法，都取 $x_0=1$，$x_1=2$，计算结果如表 7-8. 对于 Newton 法，由于在(0,2)内 $f''(x)>0$，$f(2)>0$，故取 $x_0=2$，计算结果如表 7-8.

表 7-8

x_k	单点割线法	割线法	Newton 法
x_2	1.368 421 053	1.368 421 053	1.383 388 704
x_3	1.368 851 263	1.368 850 469	1.368 869 419
x_4	1.368 803 298	1.368 808 104	1.368 808 109
x_5	1.368 808 644	1.368 808 108	1.368 808 108

由计算结果知，对单点割线法有 $|x_5-x_4|\approx 0.5\times 10^{-5}$；对割线法有 $|x_5-x_4|=0.4\times 10^{-8}$；对 Newton 法有 $|x_5-x_4|=0.1\times 10^{-8}$，故取 $x^*\approx 1.368\,808\,108$.

割线法的收敛阶虽然低于 Newton 法，但迭代一次只需计算一次函数值 $f(x_k)$，不需计算导数值 $f'(x_k)$，所以效率高，实际问题中经常使用.

与割线法类似，我们可通过 3 点 $(x_i,f(x_i))(i=k-2,k-1,k)$ 作一条抛物线，适当选取它与 x 轴交点的横坐标作为 x_{k+1}. 这样产生迭代序列的方法称为抛物线法，亦称 Muller 方法.

下面给出抛物线法的计算公式. 过 3 点 $(x_i,f(x_i))(i=k-2,k-1,k)$ 的插值多项式为

$$\begin{aligned}p_2(x)&=f(x_k)+f(x_k,x_{k-1})(x-x_k)+f(x_k,x_{k-1},x_{k-2})(x-x_k)(x-x_{k-1})\\&=f(x_k)+\omega_k(x-x_k)+f(x,x_{k-1},x_{k-2})(x-x_k)^2,\end{aligned}$$

其中

$$\omega_k=f(x_k,x_{k-1})+(x_k-x_{k-1})f(x_k,x_{k-1},x_{k-2}).$$

2 次方程 $p_2(x)=0$ 有两个根，我们选择接近 x_k 的一个作 x_{k+1}，即得迭代公式

$$x_{k+1}=x_k-\frac{2f(x_k)}{\omega_k+\operatorname{sgn}(\omega_k)\sqrt{\omega_k{}^2-4f(x_k)f(x_k,x_{k-1},x_{k-2})}}. \tag{7.25}$$

把根式写到分母是为了避免有效数字的损失.

可以证明(7.25)式产生的序列局部收敛到 $f(x)$ 的零点 x^*，即有类似于定理 7.6 的结论. 这里要假设 $f(x)$ 在 x^* 的邻域内三阶导数连续，$f'(x^*)\neq 0$. 它的收敛阶是 $p\approx 1.839$，这是方程 $\lambda^3-\lambda^2-\lambda-1=0$ 的根. 收敛速度比割线法更接近于 Newton 法.

7.4 非线性方程组的数值解法

7.4.1 非线性方程组的不动点迭代法

设含有 n 个未知数和 n 个方程的非线性方程组为

$$\boldsymbol{F}(\boldsymbol{x})=0, \tag{7.26}$$

其中 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 为 n 维列向量

$$\boldsymbol{F}(\boldsymbol{x})=(f_1(\boldsymbol{x}),f_2(\boldsymbol{x}),\cdots,f_n(\boldsymbol{x}))^{\mathrm{T}},$$

$f_i(\boldsymbol{x})(i=1,2,\cdots,n)$ 中至少有一个是 $\boldsymbol{x}$ 的非线性函数，并假设自变量和函数值都是实数. 多元非线性方程组(7.26)与一元非线性方程 $f(x)=0$ 具有相同的形式，可以与一元非线性方程并行地讨论它的迭代解法. 例如不动点迭代法和 Newton 型迭代法. 但是，这里某些定理的证明较为复杂，我们将略去其证明.

把方程组(7.26)改写成下面便于迭代的等价形式

$$\boldsymbol{x}=\boldsymbol{\Phi}(\boldsymbol{x})=(\varphi_1(\boldsymbol{x}),\varphi_2(\boldsymbol{x}),\cdots,\varphi_n(\boldsymbol{x}))^{\mathrm{T}},\tag{7.27}$$

并构造不动点迭代法

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{\Phi}(\boldsymbol{x}^{(k)}),\quad k=0,1,\cdots.\tag{7.28}$$

对于给定的初始点 $\boldsymbol{x}^{(0)}$，若由此生成的序列$\{\boldsymbol{x}^{(k)}\}$收敛，$\lim\limits_{k\to\infty}\boldsymbol{x}^{(k)}=\boldsymbol{x}^{*}$，且 $\boldsymbol{\Phi}(\boldsymbol{x})$是连续的，即 $\varphi_1(\boldsymbol{x}),\varphi_2(\boldsymbol{x}),\cdots,\varphi_n(\boldsymbol{x})$是关于自变量 $x_1,x_2,\cdots,x_n$ 的连续函数. 则 $\boldsymbol{x}^{*}$ 满足 $\boldsymbol{x}^{*}=\boldsymbol{\Phi}(\boldsymbol{x}^{*})$，即 $\boldsymbol{x}^{*}$ 是迭代函数 $\boldsymbol{\Phi}(\boldsymbol{x})$的不动点，从而 $\boldsymbol{x}^{*}$ 是方程组(7.26)的解.

例 7.11　设有非线性方程组

$$\begin{cases}x_1^{\,2}-10x_1+x_2^{\,2}+8=0,\\ x_1x_2^{\,2}+x_1-10x_2+8=0.\end{cases}\tag{7.29}$$

把它写成等价形式

$$\begin{cases}x_1=\varphi_1(x_1,x_2)=\dfrac{1}{10}(x_1^{\,2}+x_2^{\,2}+8),\\ x_2=\varphi_2(x_1,x_2)=\dfrac{1}{10}(x_1x_2^{\,2}+x_1+8).\end{cases}$$

并由此构造不动点迭代法

$$\begin{cases}x_1^{(k+1)}=\varphi_1(x_1^{(k)},x_2^{(k)})=\dfrac{1}{10}((x_1^{(k)})^2+(x_2^{(k)})^2+8),\\ x_2^{(k+1)}=\varphi_2(x_1^{(k)},x_2^{(k)})=\dfrac{1}{10}(x_1^{(k)}(x_2^{(k)})^2+x_1^{(k)}+8),\quad k=0,1,\cdots.\end{cases}\tag{7.30}$$

取初始点 $\boldsymbol{x}^{(0)}=(0,0)^{\mathrm{T}}$. 计算结果列于表 7－9，可见迭代结果收敛到方程组的解 $\boldsymbol{x}^{*}=(1,1)^{\mathrm{T}}$.

表 7－9

k	0	1	2	…	18	19
$x_1^{(k)}$	0	0.8	0.928 0	…	0.999 999 972	0.999 999 989
$x_2^{(k)}$	0	0.8	0.931 2	…	0.999 999 972	0.999 999 989

函数也称为映射，若函数 $\boldsymbol{\Phi}(\boldsymbol{x})$的定义域为 $D\subset\mathbf{R}^n$，则可用映射符号简便地表示为 $\boldsymbol{\Phi}:D\subset\mathbf{R}^n\to\mathbf{R}^n$. 为了讨论不动点迭代法(7.28)的收敛性，先定义向量值函数的映内性和压缩性.

定义 7.3　设有函数 $\boldsymbol{\Phi}:D\subset\mathbf{R}^n\to\mathbf{R}^n$. 若

$$\boldsymbol{\Phi}(\boldsymbol{x})\in D,\forall\,\boldsymbol{x}\in D,$$

则称 $\boldsymbol{\Phi}(\boldsymbol{x})$ 在 D 上是映内的，记作 $\boldsymbol{\Phi}(D)\subset D$. 又若存在常数 $L\in(0,1)$，使得

$$\|\boldsymbol{\Phi}(\boldsymbol{x})-\boldsymbol{\Phi}(\boldsymbol{y})\|\leqslant L\|\boldsymbol{x}-\boldsymbol{y}\|,\ \forall\,\boldsymbol{x},\boldsymbol{y}\in D,$$

则称 $\boldsymbol{\Phi}(\boldsymbol{x})$ 在 D 上是压缩的，L 称为压缩系数.

压缩性与所用的向量范数有关，函数 $\boldsymbol{\Phi}(\boldsymbol{x})$ 对某种范数是压缩的，对另一种范数可能不是压缩的.

定理 7.7（Brouwer 不动点定理） 若 $\boldsymbol{\Phi}$ 在有界闭凸集 $D_0\subset D$ 上连续并且映内，则 $\boldsymbol{\Phi}$ 在 D_0 内存在不动点.

映内性可保证不动点存在，但不能保证唯一. 为了保证唯一性，还需要附加压缩性条件.

定理 7.8（压缩映射原理） 设函数 $\boldsymbol{\Phi}:D\subset\mathbf{R}^n\to\mathbf{R}^n$ 在闭集 $D_0\subset D$ 上是映内的，并且对某一种范数是压缩的，压缩系数为 L，则

(1) $\boldsymbol{\Phi}(\boldsymbol{x})$ 在 D_0 上存在唯一的不动点 $\boldsymbol{x}^*$；

(2) 对任何初值 $\boldsymbol{x}^{(0)}\in D_0$，迭代法(7.28)生成的序列 $\{\boldsymbol{x}^{(k)}\}\subset D_0$ 且收敛到 $\boldsymbol{x}^*$，并有误差估计式

$$\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|\leqslant\frac{L}{1-L}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|.$$

例 7.12 对于例 7.11，设 $D_0=\{(x_1,x_2)^{\mathrm{T}}:-1.5\leqslant x_1,x_2\leqslant 1.5\}$. 试证：对任何初始点 $\boldsymbol{x}^{(0)}\in D_0$，由迭代法(7.30)生成的序列都收敛到方程(7.29)在 D_0 中的唯一解 $\boldsymbol{x}^*=(1,1)^{\mathrm{T}}$.

证 首先容易算出，对于任何 $\boldsymbol{x}=(x_1,x_2)^{\mathrm{T}}\in D_0$，都有

$$0.8\leqslant\varphi_1(x_1,x_2)\leqslant 1.25,\quad 0.312\,5\leqslant\varphi_2(x_1,x_2)\leqslant 1.287\,5.$$

因此，迭代函数 $\boldsymbol{\Phi}$ 在 D_0 上是映内的. 进而，对于任何

$$\boldsymbol{x}=(x_1,x_2)^{\mathrm{T}}\in D_0,\quad \boldsymbol{y}=(y_1,y_2)^{\mathrm{T}}\in D_0,$$

都有

$$\begin{aligned}
|\varphi_1(\boldsymbol{x})-\varphi_1(\boldsymbol{y})|&=\frac{1}{10}|(x_1+y_1)(x_1-y_1)+(x_2+y_2)(x_2-y_2)|\\
&\leqslant\frac{3}{10}(|x_1-y_1|+|x_2-y_2|)=0.3\|\boldsymbol{x}-\boldsymbol{y}\|_1,\\
|\varphi_2(\boldsymbol{x})-\varphi_2(\boldsymbol{y})|&=\frac{1}{10}|x_1-y_1+x_1x_2{}^2-y_1y_2{}^2|\\
&=\frac{1}{10}|x_1-y_1+x_1x_2{}^2-y_1x_2{}^2+y_1x_2{}^2-y_1y_2{}^2|\\
&=\frac{1}{10}|(1+x_2{}^2)(x_1-y_1)+y_1(x_2+y_2)(x_2-y_2)|\\
&\leqslant\frac{1}{10}(3.25|x_1-y_1|+4.5|x_2-y_2|)\leqslant 0.45\|\boldsymbol{x}-\boldsymbol{y}\|_1,
\end{aligned}$$

从而

$$\begin{aligned}\| \boldsymbol{\Phi}(\boldsymbol{x})-\boldsymbol{\Phi}(\boldsymbol{y}) \|_1 &= |\varphi_1(\boldsymbol{x})-\varphi_1(\boldsymbol{y})|+|\varphi_2(\boldsymbol{x})-\varphi_2(\boldsymbol{y})| \\ &\leqslant 0.75 \| \boldsymbol{x}-\boldsymbol{y} \|_1 .\end{aligned}$$

可见，函数 $\boldsymbol{\Phi}$ 在 D_0 上是压缩的. 因此，由定理 7.8 得知结论成立.

以上讨论了迭代法在 D_0 的收敛性，下面讨论局部收敛性.

定义 7.4　设 $\boldsymbol{x}^*$ 为 $\boldsymbol{\Phi}$ 的不动点，若存在 $\boldsymbol{x}^*$ 的一个邻域 $S\subset D$，对一切 $\boldsymbol{x}^{(0)}\in S$，由(7.28)式产生的序列 $\{\boldsymbol{x}^{(k)}\}\subset S$，且 $\lim\limits_{k\to\infty}x^{(k)}=\boldsymbol{x}^*$，则称 $\{\boldsymbol{x}^{(k)}\}$ 具有局部收敛性.

定义 7.5　设 $\{\boldsymbol{x}^{(k)}\}$ 收敛于 $\boldsymbol{x}^*$，存在常数 $p\geqslant 2$ 及常数 $c>0$，使

$$\lim_{k\to\infty}\frac{\| \boldsymbol{x}^{(k+1)}-\boldsymbol{x}^* \|}{\| \boldsymbol{x}^{(k)}-\boldsymbol{x}^* \|^p}=c,$$

则称 $\{\boldsymbol{x}^{(k)}\}$ 为 p 阶收敛.

定理 7.9　设 $\boldsymbol{\Phi}: D\subset\mathbf{R}^n\to\mathbf{R}^n$，$\boldsymbol{x}^*\in D$ 为 $\boldsymbol{\Phi}$ 的不动点，若存在开球 $S=S(\boldsymbol{x}^*,\delta)=\{\boldsymbol{x}:\| \boldsymbol{x}-\boldsymbol{x}^* \|<\delta\}\subset D$，常数 $L\in(0,1)$，使

$$\| \boldsymbol{\Phi}(\boldsymbol{x})-\boldsymbol{\Phi}(\boldsymbol{x}^*) \|\leqslant L\| \boldsymbol{x}-\boldsymbol{x}^* \|,\quad \forall\,\boldsymbol{x}\in S,$$

则由(7.28)式产生的序列 $\{\boldsymbol{x}^{(k)}\}$ 局部收敛至 $\boldsymbol{x}^*$.

证　任给 $\boldsymbol{x}^{(0)}\in S$，一般地，设 $\boldsymbol{x}^{(k)}\in S$，即 $\| \boldsymbol{x}^{(k)}-\boldsymbol{x}^* \|<\delta$，则

$$\begin{aligned}\| \boldsymbol{x}^{(k+1)}-\boldsymbol{x}^* \| &= \| \boldsymbol{\Phi}(\boldsymbol{x}^{(k)})-\boldsymbol{\Phi}(\boldsymbol{x}^*) \| \\ &\leqslant L\| \boldsymbol{x}^{(k)}-\boldsymbol{x}^* \|<L\delta<\delta,\end{aligned}$$

即 $\boldsymbol{x}^{(k+1)}\in S$. 进而，由

$$\| \boldsymbol{x}^{(k)}-\boldsymbol{x}^* \|\leqslant L\| \boldsymbol{x}^{(k-1)}-\boldsymbol{x}^* \|\leqslant\cdots\leqslant L^k\| \boldsymbol{x}^{(0)}-\boldsymbol{x}^* \|,$$

得知 $\lim\limits_{k\to\infty}\| \boldsymbol{x}^{(k)}-\boldsymbol{x}^* \|=0$，从而有 $\lim\limits_{k\to\infty}\boldsymbol{x}^{(k)}=x^*$. 于是，由定义 7.4 知迭代法(7.28)在点 $\boldsymbol{x}^*$ 处局部收敛. 定理得证.

与单个方程的情形类似，有时可以用关于导数的条件代替压缩条件来判别收敛性.

定理 7.10　设 $\boldsymbol{\Phi}: D\subset\mathbf{R}^n\to\mathbf{R}^n$，$\boldsymbol{\Phi}$ 在 D 内有一不动点 $\boldsymbol{x}^*$，$\boldsymbol{\Phi}$ 在 $\boldsymbol{x}^*$ 处可导，且谱半径 $\rho(\boldsymbol{\Phi}'(\boldsymbol{x}^*))=\sigma<1$，则迭代法(7.28)在点 $\boldsymbol{x}^*$ 处局部收敛，其中，函数 $\boldsymbol{\Phi}(\boldsymbol{x})$ 的导数为 Jacobi 矩阵

$$\boldsymbol{\Phi}'(\boldsymbol{x})=\begin{pmatrix}\nabla\varphi_1(\boldsymbol{x})^{\mathrm{T}}\\ \nabla\varphi_2(\boldsymbol{x})^{\mathrm{T}}\\ \cdots\\ \nabla\varphi_n(\boldsymbol{x})^{\mathrm{T}}\end{pmatrix}=\begin{pmatrix}\dfrac{\partial\varphi_1(\boldsymbol{x})}{\partial x_1} & \dfrac{\partial\varphi_1(\boldsymbol{x})}{\partial x_2} & \cdots & \dfrac{\partial\varphi_1(\boldsymbol{x})}{\partial x_n}\\ \dfrac{\partial\varphi_2(\boldsymbol{x})}{\partial x_1} & \dfrac{\partial\varphi_2(\boldsymbol{x})}{\partial x_2} & \cdots & \dfrac{\partial\varphi_2(\boldsymbol{x})}{\partial x_n}\\ \vdots & \vdots & & \vdots\\ \dfrac{\partial\varphi_n(\boldsymbol{x})}{\partial x_1} & \dfrac{\partial\varphi_n(\boldsymbol{x})}{\partial x_2} & \cdots & \dfrac{\partial\varphi_n(\boldsymbol{x})}{\partial x_n}\end{pmatrix}.$$

利用谱半径与范数的关系 $\rho(\boldsymbol{A})\leqslant\|\boldsymbol{A}\|$，我们可用 $\|\boldsymbol{\Phi}'(\boldsymbol{x}^*)\|<1$ 代替定理 7.10 中的条件 $\rho(\boldsymbol{\Phi}'(\boldsymbol{x}^*))<1$.

例如，对于例 7.11，有

$$\boldsymbol{\Phi}'(\boldsymbol{x})=\frac{1}{10}\begin{pmatrix}2x_1 & 2x_2\\ {x_2}^2+1 & 2x_1x_2\end{pmatrix}.$$

对于例 7.12 所取的区域 D_0，$\boldsymbol{\Phi}$ 的不动点 $\boldsymbol{x}^*$ 在它的内部. 容易检验，在 D_0 上有 $\|\boldsymbol{\Phi}'(\boldsymbol{x}^*)\|\leqslant 0.75$. 因此，迭代法(7.30)在点 $\boldsymbol{x}^*$ 处局部收敛.

7.4.2 非线性方程组的 Newton 法

对于非线性方程组，也可以构造类似于一元方程的 Newton 迭代法. 设 $\boldsymbol{x}^*$ 是方程组(7.26)的解，$\boldsymbol{x}^{(k)}$ 是方程组的一个近似解. 用点 $\boldsymbol{x}^{(k)}$ 处的一阶 Taylor 展开式近似每一个分量函数值 $f_i(\boldsymbol{x}^*)=0$，有

$$f_i(\boldsymbol{x}^*)\approx f_i(\boldsymbol{x}^{(k)})+\sum_{j=1}^{n}\frac{\partial f_i(\boldsymbol{x}^{(k)})}{\partial x_j}({x_j}^*-{x_j}^{(k)}),\quad i=1,2,\cdots,n.$$

写成向量形式有

$$\boldsymbol{F}(\boldsymbol{x}^*)\approx\boldsymbol{F}(\boldsymbol{x}^{(k)})+\boldsymbol{F}'(\boldsymbol{x}^{(k)})(\boldsymbol{x}^*-\boldsymbol{x}^{(k)}).\tag{7.31}$$

其中 $\boldsymbol{F}'(\boldsymbol{x}^{(k)})$ 为 $\boldsymbol{F}(\boldsymbol{x})$ 的 Jacobi 矩阵 $\boldsymbol{F}'(\boldsymbol{x})$ 在 $\boldsymbol{x}^{(k)}$ 的值，而

$$\boldsymbol{F}'(\boldsymbol{x})=\begin{pmatrix}\nabla f_1(\boldsymbol{x})^{\mathrm{T}}\\ \nabla f_2(\boldsymbol{x})^{\mathrm{T}}\\ \cdots\\ \nabla f_n(\boldsymbol{x})^{\mathrm{T}}\end{pmatrix}=\begin{pmatrix}\dfrac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \dfrac{\partial f_1(\boldsymbol{x})}{\partial x_2} & \cdots & \dfrac{\partial f_1(\boldsymbol{x})}{\partial x_n}\\ \dfrac{\partial f_2(\boldsymbol{x})}{\partial x_1} & \dfrac{\partial f_2(\boldsymbol{x})}{\partial x_2} & \cdots & \dfrac{\partial f_2(\boldsymbol{x})}{\partial x_n}\\ \vdots & \vdots & & \vdots\\ \dfrac{\partial f_n(\boldsymbol{x})}{\partial x_1} & \dfrac{\partial f_n(\boldsymbol{x})}{\partial x_2} & \cdots & \dfrac{\partial f_n(\boldsymbol{x})}{\partial x_n}\end{pmatrix}.$$

若矩阵 $\boldsymbol{F}'(\boldsymbol{x}^{(k)})$ 非奇异，则可以用使(7.31)式右端为零的向量作为 $\boldsymbol{x}^*$ 新的一个近似值，记为 $\boldsymbol{x}^{(k+1)}$，于是得到 Newton 迭代法

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}-(\boldsymbol{F}'(\boldsymbol{x}^{(k)}))^{-1}\boldsymbol{F}(\boldsymbol{x}^{(k)}),\quad k=0,2,\cdots,\tag{7.32}$$

其中 $\boldsymbol{x}^{(0)}$ 是给定的初值向量. 如果写成一般不动点迭代 $\boldsymbol{x}^{(k+1)}=\boldsymbol{\Phi}(\boldsymbol{x}^{(k)})$ 的形式，则 Newton 迭代函数为

$$\boldsymbol{\Phi}(\boldsymbol{x})=\boldsymbol{x}-(\boldsymbol{F}'(\boldsymbol{x}))^{-1}\boldsymbol{F}(\boldsymbol{x}).\tag{7.33}$$

在 Newton 法实际计算过程中，第 k 步是先解线性方程组

$$\boldsymbol{F}'(\boldsymbol{x}^{(k)})\Delta\boldsymbol{x}^{(k)}=-\boldsymbol{F}(\boldsymbol{x}^{(k)}),\tag{7.34}$$

解出 $\Delta\boldsymbol{x}^{(k)}$ 后，再令 $\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}+\Delta\boldsymbol{x}^{(k)}$，其中包括了计算向量 $\boldsymbol{F}(\boldsymbol{x}^{(k)})$ 和矩阵 $\boldsymbol{F}'(\boldsymbol{x}^{(k)})$.

例 7.13　用 Newton 法解例 7.11 的方程组(7.29).

解　对该方程组有

$$\boldsymbol{F}(\boldsymbol{x})=\begin{pmatrix} x_1{}^2-10x_1+x_2{}^2+8 \\ x_1x_2{}^2+x_1-10x_2+8 \end{pmatrix},\quad \boldsymbol{F}'(\boldsymbol{x})=\begin{pmatrix} 2x_1-10 & 2x_2 \\ x_2{}^2+1 & 2x_1x_2-10 \end{pmatrix}.$$

取初始向量 $\boldsymbol{x}^{(0)}=(0,0)^{\mathrm{T}}$,解方程组 $\boldsymbol{F}'(\boldsymbol{x}^{(0)})\Delta\boldsymbol{x}^{(0)}=-\boldsymbol{F}(\boldsymbol{x}^{(0)})$,即

$$\begin{bmatrix} -10 & 0 \\ 1 & -10 \end{bmatrix}\Delta\boldsymbol{x}^{(0)}=-\begin{bmatrix} 8 \\ 8 \end{bmatrix}.$$

求出 $\Delta\boldsymbol{x}^{(0)}$ 后,$\boldsymbol{x}^{(1)}=\boldsymbol{x}^{(0)}+\Delta\boldsymbol{x}^{(0)}=(0.8,0.88)^{\mathrm{T}}$. 同理计算 $\boldsymbol{x}^{(2)},\cdots$,计算结果列于表 7-10. 可见,Newton 法的收敛速度比例 7.11 中的迭代法(7.30)要快得多.

表 7-10

k	0	1	2	3	4
$x_1{}^{(k)}$	0	0.80	0.991 787 221	0.999 975 229	1.000 000 000
$x_2{}^{(k)}$	0	0.88	0.991 711 737	0.999 968 524	1.000 000 000

关于 Newton 法的收敛性,有下面的局部收敛性定理.

定理 7.11　设 $\boldsymbol{F}:D\subset\mathbf{R}^n\to\mathbf{R}^n$,$\boldsymbol{x}^*$ 满足 $\boldsymbol{F}(\boldsymbol{x}^*)=0$. 若有 $\boldsymbol{x}^*$ 的开邻域 $S_0\subset D$,$\boldsymbol{F}'(\boldsymbol{x})$ 在其上连续,$\boldsymbol{F}'(\boldsymbol{x}^*)$ 可逆,则

(1) 存在以 $\boldsymbol{x}^*$ 为中心,$\delta>0$ 为半径的闭球 $S=S(\boldsymbol{x}^*,\delta)\subset S_0$,使(7.33)式的 $\boldsymbol{\Phi}(\boldsymbol{x})$ 对所有 $\boldsymbol{x}\in S$ 有意义,并且 $\boldsymbol{\Phi}(\boldsymbol{x})\in S$;

(2) Newton 迭代序列 $\{\boldsymbol{x}^{(k)}\}$ 在 S 上收敛于 $\boldsymbol{x}^*$,且是超线性收敛;

(3) 若还有常数 $\alpha>0$,使

$$\|\boldsymbol{F}'(\boldsymbol{x})-\boldsymbol{F}'(\boldsymbol{x}^*)\|\leqslant\alpha\|\boldsymbol{x}-\boldsymbol{x}^*\|,\quad \forall\,\boldsymbol{x}\in S,$$

则 Newton 迭代序列 $\{\boldsymbol{x}^{(k)}\}$ 至少 2 阶收敛于 $\boldsymbol{x}^*$.

虽然 Newton 法具有 2 阶局部收敛性,但它要求 $\boldsymbol{F}'(\boldsymbol{x}^*)$ 非奇异. 如果矩阵 $\boldsymbol{F}'(\boldsymbol{x}^*)$ 奇异或病态,那么 $\boldsymbol{F}'(\boldsymbol{x}^k)$ 也可能奇异或病态,从而可能导致数值计算失败或产生数值不稳定. 这时可采用"阻尼 Newton 法",即把(7.34)式改成

$$(\boldsymbol{F}'(\boldsymbol{x}^{(k)})+\mu_k\boldsymbol{I})\Delta\boldsymbol{x}^{(k)}=-\boldsymbol{F}(\boldsymbol{x}^k),\quad k=0,1,\cdots.$$

其中的参数 μ_k 称为阻尼因子,$\mu_k\boldsymbol{I}$ 称为阻尼项. 解出 $\Delta\boldsymbol{x}^{(k)}$ 后,令 $\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}+\Delta\boldsymbol{x}^{(k)}$. 加进阻尼项的目的,是使线性方程的系数矩阵非奇异并良态. 当 μ_k 选得合适时,阻尼 Newton 法是线性收敛的.

例 7.14　用 Newton 法和阻尼 Newton 法求解方程 $\boldsymbol{F}(\boldsymbol{x})=0$,其中

$$\boldsymbol{F}(\boldsymbol{x})=\begin{pmatrix} x_1{}^2-10x_1+x_2{}^2+23 \\ x_1x_2{}^2+x_1-10x_2+2 \end{pmatrix}.$$

解　易知该方程有一个解是 $\boldsymbol{x}^*=(4,1)^{\mathrm{T}}$. 由于

$$\boldsymbol{F}'(\boldsymbol{x}^*)=\begin{bmatrix}-2 & 2\\ 2 & -2\end{bmatrix}$$

是奇异的,取阻尼因子 $\mu_k=10^{-5}$. 若取 $\boldsymbol{x}^{(0)}=(2.5,2.5)^{\mathrm{T}}$,按 Newton 法有 $\boldsymbol{x}^{(1)}=(3.538\,461\,538,1.438\,461\,538)^{\mathrm{T}},\cdots,\boldsymbol{x}^{(25)}=(4.000\,000\,025,1.000\,000\,025)^{\mathrm{T}}$. 再按阻尼 Newton 法计算有 $\boldsymbol{x}^{(1)}=(3.538\,463\,160,1.438\,461\,083)^{\mathrm{T}},\cdots,\boldsymbol{x}^{(29)}=(4.000\,000\,286,1.000\,000\,286)^{\mathrm{T}}$.

可见,即使矩阵 $\boldsymbol{F}'(\boldsymbol{x}^*)$奇异,只要 $\boldsymbol{F}'(\boldsymbol{x}^{(k)})$非奇异,Newton 法仍收敛,但收敛是线性的. 因为此例题的维数太小,Newton 法并没有出现奇异或数值稳定性问题,从而阻尼 Newton 法不仅没有显示出它的作用,反而使迭代次数更多. 但可以看出,阻尼 Newton 法是线性收敛的.

用迭代法求解非线性方程,特别是非线性方程组时,初始值的选取至关重要. 初值不仅影响迭代是否收敛,而且当方程多解时,不同的初值可能收敛到不同的解.

例 **7.15**　用 Newton 法求解 $\boldsymbol{F}(\boldsymbol{x})=0$,其中

$$\boldsymbol{F}(\boldsymbol{x})=\begin{pmatrix}x_1^{\,2}-x_2+1\\ (x_1-2)^2+(x_2-0.5)^2-1\end{pmatrix}.$$

解　该方程组的实数解是抛物线 $x_1^{\,2}-x_2-1=0$ 与圆 $(x_1-2)^2+(x_2-0.5)^2-1=0$ 的交点. 这两个实根是 $\boldsymbol{x}^*\approx(1.067\,346\,086,0.139\,227\,667)^{\mathrm{T}}$ 和 $\boldsymbol{x}^{**}\approx(1.546\,342\,883,1.391\,176\,313)^{\mathrm{T}}$. 如果取初始向量 $\boldsymbol{x}^{(0)}=(0,0)^{\mathrm{T}}$,那么有 $\boldsymbol{x}^{(1)}=(1.062\,5,-1.000\,0)^{\mathrm{T}},\cdots,\boldsymbol{x}^{(5)}=(1.067\,343\,609,0.139\,221\,092)^{\mathrm{T}}$,计算结果收敛到 $\boldsymbol{x}^*$. 若取初值 $\boldsymbol{x}^{(0)}=(2,2)^{\mathrm{T}}$,则有 $\boldsymbol{x}^{(1)}=(1.645\,833\,333,1.583\,333\,333)^{\mathrm{T}},\cdots,\boldsymbol{x}^{(5)}=(1.546\,342\,883,1.391\,176\,313)^{\mathrm{T}}$,计算结果收敛到 $\boldsymbol{x}^{**}$.

一般来说,为了保证迭代的收敛性,初始值应当取在所求解的足够小的邻域内. 有的实际问题可以凭经验取初值,有的则可以用某些方法预估一个近似解. 从数学的角度讲,这是个相当困难的问题.

7.4.3　非线性方程组的拟 Newton 法

Newton 法有较好的收敛性,但是每步都要计算 $\boldsymbol{F}'(\boldsymbol{x}^{(k)})$是很不方便的,特别是当 $\boldsymbol{F}(\boldsymbol{x})$的分量函数 $f_i(\boldsymbol{x})$比较复杂时,求导数值将是困难的. 所以,我们用较简单的矩阵 $\boldsymbol{A}_k$ 代替 Newton 法的 $\boldsymbol{F}'(\boldsymbol{x}^{(k)})$,迭代公式是

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}-\boldsymbol{A}_k^{-1}\boldsymbol{F}'(\boldsymbol{x}^{(k)}),\quad k=0,1,\cdots. \tag{7.35}$$

下一步是要确定 $\boldsymbol{A}_{k+1}$. 若是单个方程，割线法是将 Newton 法中的 $f'(x_{k+1})$用差商$\dfrac{f(x_{k+1})-f(x_k)}{x_{k+1}-x_k}$代替. 对于方程组的情形，$\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}$ 是向量，于是取具有性质

$$\boldsymbol{A}_{k+1}(\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)})=\boldsymbol{F}(\boldsymbol{x}^{(k+1)})-\boldsymbol{F}(\boldsymbol{x}^{(k)}) \tag{7.36}$$

的矩阵 $\boldsymbol{A}_{k+1}$代替 Newton 法中的 $\boldsymbol{F}'(\boldsymbol{x}^{(k+1)})$. 在多元情形下，当 $\boldsymbol{x}^{(k)}$和 $\boldsymbol{x}^{(k+1)}$已知时，由方程(7.36)不能确定矩阵 $\boldsymbol{A}_{k+1}$($n>1$ 个方程中含有 $n^2>n$ 个未知量). 因此，为了确定矩阵 $\boldsymbol{A}_{k+1}$，需要附加其他条件. 一个可行的途径是令

$$\boldsymbol{A}_{k+1}=\boldsymbol{A}_k+\Delta\boldsymbol{A}_k,\quad \operatorname{rank}(\Delta\boldsymbol{A}_k)=m\geqslant 1. \tag{7.37}$$

称 $\Delta\boldsymbol{A}_k$ 为增量矩阵，由此得到的迭代法(7.35)式称为拟 Newton 法，(7.36)式称为拟 Newton 方程. 通常取 $m=1$ 或 2，当 $m=1$ 时，称为秩 **1** 方法；当 $m=2$ 时，称为秩 **2** 方法.

下面以秩 1 的情形为例，说明确定增量矩阵 $\Delta\boldsymbol{A}_k$ 的方法.

秩为 1 的矩阵 $\Delta\boldsymbol{A}_k$ 总可表示为 $\Delta\boldsymbol{A}_k=\boldsymbol{u}_k\boldsymbol{v}_k{}^{\mathrm{T}}$，其中 $\boldsymbol{u}_k,\boldsymbol{v}_k\in\mathbf{R}^n$ 为列向量. 记

$$\boldsymbol{s}_k=\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)},\quad \boldsymbol{y}_k=\boldsymbol{F}(\boldsymbol{x}^{(k+1)})-\boldsymbol{F}(\boldsymbol{x}^{(k)}).$$

选择 $\boldsymbol{u}_k$ 和 $\boldsymbol{v}_k$，使得矩阵 $\boldsymbol{A}_{k+1}=\boldsymbol{A}_k+\Delta\boldsymbol{A}_k$ 满足拟 Newton 方程(7.36)，即

$$(\boldsymbol{A}_k+\boldsymbol{u}_k\boldsymbol{v}_k{}^{\mathrm{T}})\boldsymbol{s}_k=\boldsymbol{y}_k.$$

若 $\boldsymbol{v}_k{}^{\mathrm{T}}\boldsymbol{s}_k\neq 0$，则由此可解出

$$\boldsymbol{u}_k=\frac{1}{\boldsymbol{v}_k{}^{\mathrm{T}}\boldsymbol{s}_k}(\boldsymbol{y}_k-\boldsymbol{A}_k\boldsymbol{s}_k),$$

即 $\boldsymbol{u}_k$ 由 $\boldsymbol{v}_k$ 唯一确定. 向量 $\boldsymbol{v}_k$ 的一个自然取法是令 $\boldsymbol{v}_k=\boldsymbol{s}_k$，因为只要 $\boldsymbol{x}^{(k+1)}\neq\boldsymbol{x}^{(k)}$(即迭代尚未终止)，这时总有 $\boldsymbol{v}_k{}^{\mathrm{T}}\boldsymbol{s}_k=\|\boldsymbol{s}_k\|_2^2\neq 0$. 把上述 $\boldsymbol{v}_k$ 和 $\boldsymbol{u}_k$ 代入 $\Delta\boldsymbol{A}_k$ 有

$$\Delta\boldsymbol{A}_k=\frac{1}{\|\boldsymbol{s}_k\|_2^2}(\boldsymbol{y}_k-\boldsymbol{A}_k\boldsymbol{s}_k)\boldsymbol{s}_k{}^{\mathrm{T}}.$$

于是得到求解方程 $\boldsymbol{F}(\boldsymbol{x})=0$ 的迭代法

$$\begin{cases}\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}-\boldsymbol{A}_k{}^{-1}\boldsymbol{F}(\boldsymbol{x}^{(k)}),\\ \boldsymbol{s}_k=\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)},\\ \boldsymbol{y}_k=\boldsymbol{F}(\boldsymbol{x}^{(k+1)})-\boldsymbol{F}(\boldsymbol{x}^{(k)}),\\ \boldsymbol{A}_{k+1}=\boldsymbol{A}_k+\dfrac{1}{\|\boldsymbol{s}_k\|_2^2}(\boldsymbol{y}_k-\boldsymbol{A}_k\boldsymbol{s}_k)\boldsymbol{s}_k{}^{\mathrm{T}},\quad k=0,1,\cdots.\end{cases} \tag{7.38}$$

称之为 Broyden 秩 **1** 方法，其中的初始值 $\boldsymbol{x}^{(0)}$ 给定，$\boldsymbol{A}_0$ 可取为 $\boldsymbol{F}'(\boldsymbol{x}^{(0)})$或单位矩阵.

利用下面的引理，可以避免方程(7.38)中的矩阵求逆，从而可将解方程组的直接法的运算量 $O(n^3)$降为 $O(n^2)$. 引理的结论只要直接做矩阵运算即可证明.

引理　若矩阵 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 非奇异，$\boldsymbol{u},\boldsymbol{v}\in\mathbf{R}^{\mathrm{T}}$，则 $\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$ 非奇异的充分必要条件是 $1+\boldsymbol{v}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{u}\neq 0$，并且有

$$(\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}})^{-1}=\boldsymbol{A}^{-1}-\frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{A}^{-1}}{1+\boldsymbol{v}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{u}}. \tag{7.39}$$

在(7.38)式中，令 $\boldsymbol{B}_k=\boldsymbol{A}_k{}^{-1}$，有

$$\boldsymbol{A}_k{}^{-1}\boldsymbol{u}_k=\frac{1}{\|\boldsymbol{s}_k\|_2^2}\boldsymbol{B}_k(\boldsymbol{y}_k-\boldsymbol{A}_k\boldsymbol{s}_k)=\frac{1}{\|\boldsymbol{s}_k\|_2^2}(\boldsymbol{B}_k\boldsymbol{y}_k-\boldsymbol{s}_k),$$

$$1+\boldsymbol{v}_k{}^{\mathrm{T}}\boldsymbol{A}_k{}^{-1}\boldsymbol{u}_k=1+\frac{1}{\|\boldsymbol{s}_k\|_2^2}(\boldsymbol{v}_k{}^{\mathrm{T}}\boldsymbol{B}_k\boldsymbol{y}_k-\boldsymbol{v}_k{}^{\mathrm{T}}\boldsymbol{s}_k)=\frac{1}{\|\boldsymbol{s}_k\|_2^2}\boldsymbol{s}_k{}^{\mathrm{T}}\boldsymbol{B}_k\boldsymbol{y}_k.$$

如果 $\boldsymbol{s}_k{}^{\mathrm{T}}\boldsymbol{B}_k\boldsymbol{y}_k\neq 0$，那么利用(7.39)式有

$$\boldsymbol{B}_{k+1}=(\boldsymbol{A}_k+\boldsymbol{u}_k\boldsymbol{v}_k{}^{\mathrm{T}})^{-1}=\boldsymbol{B}_k-\frac{1}{\boldsymbol{s}_k{}^{\mathrm{T}}\boldsymbol{B}_k\boldsymbol{y}_k}(\boldsymbol{B}_k\boldsymbol{y}_k-\boldsymbol{s}_k)\boldsymbol{s}_k{}^{\mathrm{T}}\boldsymbol{B}_k.$$

于是，方程(7.38)可改写成

$$\begin{cases}\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}-\boldsymbol{B}_k\boldsymbol{F}(\boldsymbol{x}^{(k)}),\\ \boldsymbol{s}_k=\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)},\\ \boldsymbol{y}_k=\boldsymbol{F}(\boldsymbol{x}^{(k+1)})-\boldsymbol{F}(\boldsymbol{x}^{(k)}),\\ \boldsymbol{B}_{k+1}=\boldsymbol{B}_k+\dfrac{1}{\boldsymbol{s}_k{}^{\mathrm{T}}\boldsymbol{B}_k\boldsymbol{y}_k}(\boldsymbol{s}_k-\boldsymbol{B}_k\boldsymbol{y}_k)\boldsymbol{s}_k{}^{\mathrm{T}}\boldsymbol{B}_k,\quad k=0,1,\cdots.\end{cases}$$

称之为逆 Broyden 秩 1 方法，其中的初始值 $\boldsymbol{x}^{(0)}$ 给定，$\boldsymbol{B}_0$ 取为 $(\boldsymbol{F}'(\boldsymbol{x}^{(0)}))^{-1}$ 或单位矩阵. 逆 Broyden 方法是一种能有效地求解非线性方程组的拟 Newton 方法. 可以证明，在一定的条件下，它是超线性收敛的.

例 7.16　用逆 Broyden 方法解例 7.15 中的方程组.

解　对所给 $\boldsymbol{F}(\boldsymbol{x})$ 有

$$\boldsymbol{F}'(\boldsymbol{x})=\begin{pmatrix}2x_1 & -1\\ 2x_1-4 & 2x_2-1\end{pmatrix}.$$

取 $\boldsymbol{x}^{(0)}=(0,0)^{\mathrm{T}}$，有 $\boldsymbol{F}(\boldsymbol{x}^{(0)})=(-1,3.25)^{\mathrm{T}}$ 及

$$\boldsymbol{F}'(\boldsymbol{x}^{(0)})=\begin{bmatrix}0 & -1\\ -4 & -1\end{bmatrix},$$

$$\boldsymbol{B}_0=(\boldsymbol{F}'(\boldsymbol{x}^{(0)}))^{-1}=\begin{bmatrix}0.25 & -0.25\\ -1 & 0\end{bmatrix},$$

$$\boldsymbol{x}^{(1)}=\boldsymbol{x}^{(0)}-\boldsymbol{B}_0\boldsymbol{F}(\boldsymbol{x}^{(0)})=(1.062\,5,-1)^{\mathrm{T}},$$

$$\boldsymbol{s}_0=\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(0)}=\boldsymbol{x}^{(1)},$$

$$\boldsymbol{F}(\boldsymbol{x}^{(1)})=(1.128\,906\,25,2.128\,906\,25)^{\mathrm{T}},$$

$$\boldsymbol{y}_0=\boldsymbol{F}(\boldsymbol{x}^{(1)})-\boldsymbol{F}(\boldsymbol{x}^{(0)})=(2.128\,906\,25,-1.121\,093\,75)^{\mathrm{T}},$$

$$\boldsymbol{B}_1=\begin{bmatrix} 0.355\,744\,1 & -0.272\,193\,2 \\ -0.522\,499\,1 & 0.100\,216\,2 \end{bmatrix}.$$

接着再进行第 $k=1$ 步的计算，如此迭代 11 次后有

$$\boldsymbol{x}^{(11)}=(1.546\,342\,883\,32,1.391\,176\,312\,79)^{\mathrm{T}},$$

这是具有 12 位有效数字的近似解. 如果用 Newton 法求解，迭代到 $\boldsymbol{x}^{(7)}$ 便可得到同样精度的结果，比逆 Broyden 方法少迭代 4 次，但每步计算量却要多得多.

评　注

本章介绍非线性方程和非线性方程组的数值解法，主要方法有二分法、Steffensen方法、方程求根的 Newton 法和割线法、非线性方程组求解的 Newton 法和拟 Newton 法. 介绍了不动点迭代、局部收敛性和收敛阶等基本概念和理论.

单个方程的一阶迭代法比较容易构造，但用于实际计算的迭代法最好是超线性收敛的. Steffensen 方法可以把一阶方法加速为二阶的. Newton 法是实用的有效方法，它具有至少二阶的收敛性. 但 Newton 法需要求导数. 应用 Newton 法的关键在于选取足够精确的初值，如果初值选取不当，则 Newton 法可能发散. 尽管如此，Newton 法作为最经典的求解方法，至今仍是一个常用算法，并且很多新算法也是针对 Newton 法存在的缺点而加以改进所提出的. 应用 Newton 法时，一般还与计算多项式的秦九韶算法结合起来. 割线法和抛物线法是多点迭代法，它们属于所谓插值方法的范围.

非线性方程组的解法和理论是当今数值分析研究的重点之一，新的方法不断出现，本章介绍的只是一个简单的开头. 非线性方程组迭代法的概念和理论，与单个方程的情形是类似的，后者可以看成是前者的特殊情形. 但是，为了教学方便，本章先重点介绍了单个方程的情形.

对于非线性方程组，Newton 法的关键在于每步要解一个方程组，高维时的工作量很大. 防止 Newton 方程组奇异或病态的办法是加阻尼项. 互逆形式的拟 Newton 法不需要求导数，也不需要求解 Newton 方程组，计算效率比 Newton 法高，但只是超线性收敛的，而 Newton 法具有二阶收敛性. 本章只导出了秩 1 拟 Newton 法，还有秩 2 拟 Newton 法，有兴趣的读者可参看有关文献.

习　题　7

7.1　用二分法求方程 $e^x+10x-2=0$ 在区间$(0,1)$内的根，要求精确到

3 位小数.

7.2 证明方程 $1-x-\sin x=0$ 在区间$(0,1)$内有一根,用二分法经过多少次二分求得的近似根误差不大于 0.5×10^{-4}?

7.3 用二分法和 Newton 法求 $x-\tan x=0$ 的最小正根.

7.4 设有方程

(1) $x-\cos x=0$; (2) $3x^2-\mathrm{e}^x=0$.

确定区间$[a,b]$及迭代函数 $\varphi(x)$,使 $x_{k+1}=\varphi(x_k)$对任意 $x_0\in[a,b]$均收敛,并求各方程的根,误差不超过 10^{-4}.

7.5 设 $f(x)=0$ 有根,且 $0<m\leqslant f'(x)\leqslant M,-\infty<x<+\infty$. 试证明由 $x_{k+1}=x_k-\lambda f(x_k)$产生的序列$\{x_k\}$对任意的 x_0 和 $0<\lambda<\dfrac{2}{M}$均收敛于根.

7.6 已知在区间$[a,b]$内方程 $x=\varphi(x)$只有一根,且

$$|\varphi'(x)|\geqslant k>1.$$

试问如何将 $\varphi(x)$化为适合于迭代的形式? 求 $x=\tan x$ 在 $x_0=4.5$ 附近的根,准确到 4 位小数.

7.7 对于 $\varphi(x)=x+x^3$,$x^*=0$ 为 $\varphi(x)$的一个不动点. 验证 $x_{k+1}=\varphi(x_k)$迭代对 $x_0\neq0$ 不收敛,但改用 Steffensen 方法却是收敛的.

7.8 用下列方法求 $f(x)=x^3-3x-1=0$ 在 $x_0=2$ 附近的根. 根的准确值 $x^*=1.879\,385\,24\cdots$,要求计算结果准确到 4 位有效数字.

(1) 用 Newton 法,$x_0=2$;

(2) 用割线法,$x_0=2$,$x_1=1.9$;

(3) 用抛物线法,$x_0=1$,$x_1=3$,$x_2=2$.

7.9 讨论用 Newton 法求解方程 $f(x)=0$ 的收敛性,其中

(1) $f(x)=\begin{cases}\sqrt{x}, & x\geqslant0,\\ -\sqrt{-x}, & x<0;\end{cases}$ (2) $f(x)=\begin{cases}\sqrt[3]{x^2}, & x\geqslant0,\\ -\sqrt[3]{x^2}, & x<0.\end{cases}$

7.10 将 Newton 法用于解方程 $x^3-a=0$,讨论其收敛性.

7.11 设 x^* 是方程 $f(x)=0$ 的根,并且 $f'(x^*)\neq0$,$f''(x)$在 x^* 的邻域上连续. 试证:Newton 法的迭代序列$\{x_k\}$满足

$$\lim_{k\to\infty}\frac{x_k-x_{k-1}}{(x_{k-1}-x_{k-2})^2}=-\frac{f''(x^*)}{2f'(x^*)}.$$

7.12 构造一种不动点迭代法,求方程组

$$\begin{cases}x_1-0.7\sin x_1-0.2\cos x_2=0,\\ x_2-0.7\cos x_1+0.2\sin x_2=0,\end{cases}$$

在 $\boldsymbol{x}^{(0)}=(0.5,0.5)^{\mathrm{T}}$ 附近的解,分析方法的收敛性,迭代至 $\boldsymbol{x}^{(3)}$ 或达到误差限为

10^{-3}的精度.

7.13　用 Newton 法求解：

(1) 7.12 题的方程组，取 $\boldsymbol{x}^{(0)}=(0.5,0.5)^{\mathrm{T}}$；

(2) $\begin{cases}x^2+y^2=4,\\ x^2-y^2=1,\end{cases}$取 $\boldsymbol{x}^{(0)}=(1.6,1.2)^{\mathrm{T}}$.

7.14　用逆 Broyden 秩 1 方法求解 7.13 题中的两个方程组.

数值试验题 7

7.1　对方程 $x=1.6+0.99\cos x$ 的简单迭代法

$$x_{k+1}=1.6+0.99\cos x_k,\quad x_0=\frac{\pi}{2},\quad k=0,1,\cdots,$$

作计算，并与 Steffensen 加速方法比较，准确解为 $x^*=1.585\,471\,802\cdots$.

7.2　用迭代法求方程 $x^3+3x^2-1=0$ 的全部根，要求误差限为0.5×10^{-8}.

7.3　用适当的方法求方程 $x^9-522+\mathrm{e}^x=0$ 在 1.9 附近的根.

7.4　用迭代法 $\boldsymbol{x}^{(k+1)}=\boldsymbol{\Phi}(\boldsymbol{x}^{(k)})$求下列函数 $\boldsymbol{\Phi}(\boldsymbol{x})$在 D 中的不动点 $\boldsymbol{x}^*$，使结果有 8 位有效数字，其中

(1) $\boldsymbol{\Phi}(\boldsymbol{x})=\begin{pmatrix}\dfrac{7+x_2^2+4x_3}{12}\\ \dfrac{11-x_1^2+x_3}{10}\\ \dfrac{8-x_2^3}{10}\end{pmatrix}$,

$D=\{(x_1,x_2,x_3)^{\mathrm{T}}:\ 0\leqslant x_1,x_2,x_3\leqslant1.5\}$,　$\boldsymbol{x}^{(0)}=(0,0,0)^{\mathrm{T}}$；

(2) $\boldsymbol{\Phi}(\boldsymbol{x})=\begin{pmatrix}\dfrac{1}{3}\cos(x_2x_3)+\dfrac{1}{6}\\ \dfrac{1}{9}\sqrt{{x_1}^2+\sin x_3+1.06}-0.1\\ -\dfrac{1}{20}\mathrm{e}^{-x_1x_2}-\dfrac{1}{60}(10\pi-3)\end{pmatrix}$,

$D=\{(x_1,x_2,x_3)^{\mathrm{T}}:\ -1\leqslant x_i\leqslant1,i=1,2,3\}$,　$\boldsymbol{x}^{(0)}=(0,0,0)^{\mathrm{T}}$.

7.5　用 Newton 法和逆 Broyden 秩 1 方法求下列非线性方程组 $\boldsymbol{F}(\boldsymbol{x})=0$ 的解. 逆 Broyden 秩 1 方法的初始矩阵 $\boldsymbol{B}_0$ 分别取$(\boldsymbol{F}'(\boldsymbol{x}^{(0)}))^{-1}$和单位矩阵. 计算结果与上题进行比较.

(1) $\boldsymbol{F}(\boldsymbol{x})=\begin{pmatrix}12x_1-x_2^{\,2}-4x_3-7\\ x_1^{\,2}+10x_2-x_3-11\\ x_2^{\,3}+10x_3-8\end{pmatrix}$, $\boldsymbol{x}^{(0)}=(1,1,1)^{\mathrm{T}}$;

(2) $\boldsymbol{F}(\boldsymbol{x})=\begin{pmatrix}3x_1-\cos(x_2x_3)-\dfrac{1}{2}\\ x_1^{\,2}-81(x_2+0.1)^2+\sin x_3+1.06\\ \mathrm{e}^{-x_1x_2}+20x_3+\dfrac{1}{3}(10\pi-3)\end{pmatrix}$, $\boldsymbol{x}^{(0)}=(0,0,0)^{\mathrm{T}}$.

第 8 章　矩阵特征值问题的数值解法

对于矩阵 $\boldsymbol{A} \in \mathbf{R}^{n\times n}$（或 $\mathbf{C}^{n\times n}$），特征值问题是求 $\lambda \in \mathbf{C}$ 及非零向量 $\boldsymbol{x}$，使

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}.$$

称 λ 为矩阵 $\boldsymbol{A}$ 的特征值，$\boldsymbol{x}$ 为对应于 λ 的特征向量. 上述方程是一个非线性方程组，它有非零解 $\boldsymbol{x}$ 的充要条件是

$$\varphi(\lambda) = \det(\lambda \boldsymbol{I} - \boldsymbol{A}) = \lambda^n + c_1\lambda^{n-1} + \cdots + c_{n-1}\lambda + c_n = 0,$$

称 $\varphi(\lambda)$ 为特征多项式. 方程 $\varphi(\lambda) = 0$ 有 n 个根，包括重根和复根.

因为一般不能通过有限次运算准确求解方程 $\varphi(\lambda) = 0$ 的根，而且有的问题只需要求部分特征值和特征向量，因此特征值问题的数值方法通常采用迭代法. 本章先介绍特征值问题的一些有关性质和估计，再介绍一些数值求解方法.

8.1　特征值问题的性质与估计

定理 8.1　设 $\boldsymbol{A} = (a_{ij}) \in \mathbf{R}^{n\times n}$，$\lambda_i (i = 1,2,\cdots,n)$ 是 A 的特征值，则有

(1) $\prod\limits_{i=1}^{n} \lambda_i = \det(\boldsymbol{A})$；

(2) $\sum\limits_{i=1}^{n} \lambda_i = \sum\limits_{i=1}^{n} a_{ii} = \mathrm{tr}(\boldsymbol{A})$，称为 $\boldsymbol{A}$ 的迹.

定理 8.2（Gershgorin 圆盘定理）　设矩阵 $\boldsymbol{A} = (a_{ij}) \in \mathbf{C}^{n\times n}$，则 $\boldsymbol{A}$ 的每一个特征值

$$\lambda \in \bigcup_{i=1}^{n} D_i,$$

其中 D_i 为第 i 个圆盘

$$D_i = \{z: \mid z - a_{ii} \mid \leqslant \sum_{j=1,\ j\neq i}^{n} \mid a_{ii} \mid\}, \quad i = 1,2\cdots,n.$$

证　设 λ 为 $\boldsymbol{A}$ 的任意一个特征值，$\boldsymbol{x} \neq 0$ 为对应的特征向量，即

$$(\lambda \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = 0.$$

记 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$，$\mid x_i \mid = \max\limits_{k} \mid x_k \mid$，则 $x_i \neq 0$，

$$(\lambda - a_{ii})x_i = \sum_{j=1,\ j\neq i}^{n} a_{ij}x_j.$$

由于 $\left|\dfrac{x_j}{x_i}\right| \leqslant 1 (j \neq i)$，有

$$|\lambda - a_{ii}| \leqslant \sum_{j \neq i} |a_{ij}| \left|\frac{x_j}{x_i}\right| \leqslant \sum_{j \neq i} |a_{ij}|.$$

这说明 λ 属于 D_i. 定理得证.

从定理的证明可见，如果一个特征向量的第 i 个分量按模最大，则对应的特征值一定属于第 i 个圆盘中. 利用定理 8.2，我们可以由 $\boldsymbol{A}$ 的元素估计特征值的范围. $\boldsymbol{A}$ 的 n 个特征值均落在 n 个圆盘上，但不一定每个圆盘都有一个特征值.

定义 8.1 设 $\boldsymbol{A}$ 为 n 阶实对称矩阵，对于任一非零向量 $\boldsymbol{x}$，称

$$R(\boldsymbol{x}) = \frac{(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{x})}{(\boldsymbol{x}, \boldsymbol{x})}$$

为对应于向量 $\boldsymbol{x}$ 的 Rayleigh 商.

定理 8.3 设 $\boldsymbol{A}$ 为 n 阶实对称矩阵，其特征值都为实数，排列为

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n,$$

对应的特征向量 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ 组成正交向量组，则有

(1) 对任何非零向量 $\boldsymbol{x} \in \mathbf{R}^n$，有 $\lambda_n \leqslant R(\boldsymbol{x}) \leqslant \lambda_1$；

(2) $\lambda_1 = \max\limits_{0 \neq \boldsymbol{x} \in \mathbf{R}^n} R(\boldsymbol{x}) = R(\boldsymbol{x}_1)$；

(3) $\lambda_n = \min\limits_{0 \neq \boldsymbol{x} \in \mathbf{R}^n} R(\boldsymbol{x}) = R(\boldsymbol{x}_n)$.

证 设 $\boldsymbol{x} \neq 0$，则有表示式

$$\boldsymbol{x} = \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i,$$

$$(\boldsymbol{x}, \boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^2 > 0,$$

$$\lambda_n \sum_{i=1}^{n} \alpha_i^2 \leqslant \sum_{i=1}^{n} \alpha_i^2 \lambda_i = (\boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}) \leqslant \lambda_1 \sum_{i=1}^{n} \alpha_i^2.$$

由此可见，(1) 成立，(2) 和 (3) 是显然的. 定理得证.

对于复矩阵 $\boldsymbol{A} \in \mathbf{C}^{n \times n}$，亦有类似性质，但应注意“$\boldsymbol{A}$ 为对称矩阵”应改为“$\boldsymbol{A}$ 为 Hermite 阵”，即 $\boldsymbol{A}^{\mathrm{H}} = \boldsymbol{A}$，其特征值都是实数，特征向量也构成正交向量组.

8.2 幂法和反幂法

8.2.1 幂法和加速方法

设矩阵 $\boldsymbol{A} \in \mathbf{R}^{n \times n}$ 的 n 个特征值满足

$$|\lambda_1|>|\lambda_2|\geqslant\cdots\geqslant|\lambda_n|,\tag{8.1}$$

对应的 n 个特征向量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 线性无关. 称模最大的特征值 λ_1 为主特征值，称对应的特征向量 $\boldsymbol{x}_1$ 为主特征向量.

幂法用于求主特征值和主特征向量. 它的基本思想是任取一个非零的初始向量 $\boldsymbol{v}_0$，由矩阵 $\boldsymbol{A}$ 构造一向量序列

$$\boldsymbol{v}_k=\boldsymbol{A}\boldsymbol{v}_{k-1}=\boldsymbol{A}^k\boldsymbol{v}_0,\quad k=1,2,\cdots.$$

由假设，$\boldsymbol{v}_0$ 可表示为

$$\boldsymbol{v}_0=\alpha_1\boldsymbol{x}_1+\alpha_2\boldsymbol{x}_2+\cdots+\alpha_n\boldsymbol{x}_n.\tag{8.2}$$

若记$(\boldsymbol{v}_k)_i$ 为 $\boldsymbol{v}_k$ 的第 i 个分量，则有

$$\begin{aligned}\boldsymbol{v}_k=\boldsymbol{A}^k\boldsymbol{v}_0&=\sum_{i=1}^{n}\alpha_i\lambda_i^k\boldsymbol{x}_i\\&=\lambda_1^k\left(\alpha_1\boldsymbol{x}_1+\sum_{i=2}^{n}\alpha_i\left(\frac{\lambda_i}{\lambda_1}\right)^k\boldsymbol{x}_i\right)=\lambda_i^k(\alpha_1\boldsymbol{x}_1+\boldsymbol{\varepsilon}_k),\\\frac{(\boldsymbol{v}_{k+1})_i}{(\boldsymbol{v}_k)_i}&=\frac{\lambda_1(\alpha_1\boldsymbol{x}_1+\boldsymbol{\varepsilon}_{k+1})_i}{(\alpha_1\boldsymbol{x}_1+\boldsymbol{\varepsilon}_k)_i},\end{aligned}$$

其中 $\boldsymbol{\varepsilon}_k=\sum\limits_{i=2}^{n}\alpha_i\left(\dfrac{\lambda_i}{\lambda_1}\right)^k\boldsymbol{x}_i$. 若 $\alpha_1\neq0,(\boldsymbol{x}_1)_i\neq0$，由$\lim\limits_{k\to\infty}\boldsymbol{\varepsilon}_k=0$ 知

$$\lim_{k\to\infty}\frac{\boldsymbol{v}_k}{\lambda_1^k}=\alpha_1\boldsymbol{x}_1,\quad\lim_{k\to\infty}\frac{(\boldsymbol{v}_{k+1})_i}{(\boldsymbol{v}_k)_i}=\lambda_1.$$

可见，当 k 充分大时，$\boldsymbol{v}_k$ 近似于主特征向量，$\boldsymbol{v}_{k+1}$ 与 $\boldsymbol{v}_k$ 的对应非零分量的比值近似于主特征值.

在实际计算中，需要对计算结果进行规范化. 因为当 $|\lambda_1|<1$ 时，$\boldsymbol{v}_k$ 趋于零，当 $|\lambda_1|>1$ 时，$\boldsymbol{v}_k$ 的非零分量趋于无穷. 从而计算时会出现下溢或上溢. 为此，对 $\boldsymbol{z}=(z_1,z_2,\cdots,z_n)^{\mathrm{T}}\in\mathbf{R}^n$，记 $\max(\boldsymbol{z})=z_i$，其中 $|z_i|=\|\boldsymbol{z}\|_\infty$. 这样，我们有如下幂法的实用的计算公式：

$$\begin{cases}\boldsymbol{v}_0=\boldsymbol{u}_0\neq0,\\\boldsymbol{v}_k=\boldsymbol{A}\boldsymbol{u}_{k-1},\\\boldsymbol{u}_k=\dfrac{\boldsymbol{v}_k}{\max(\boldsymbol{v}_k)},\quad k=1,2,\cdots.\end{cases}\tag{8.3}$$

定理 8.4　设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 的特征值$\lambda_i(i=1,2,\cdots,n)$ 满足(8.1) 式，且有对应的 n 个线性无关的特征向量 $\boldsymbol{x}_i(i=1,2,\cdots,n)$. 给定初始向量 $\boldsymbol{v}_0=\sum\limits_{i=1}^{n}\alpha_i\boldsymbol{x}_i$，$\alpha_1\neq0$，则由(8.3) 式生成的向量序列有

$$\lim_{k\to\infty}\boldsymbol{u}_k=\frac{\boldsymbol{x}_1}{\max(\boldsymbol{x}_1)},\quad\lim_{k\to\infty}\max(\boldsymbol{v}_k)=\lambda_1.$$

证　由(8.3)式有

$$\boldsymbol{v}_k = \frac{\boldsymbol{A}^k \boldsymbol{v}_0}{\max(\boldsymbol{A}^{k-1}\boldsymbol{v}_0)}, \quad \boldsymbol{u}_k = \frac{\boldsymbol{A}^k \boldsymbol{v}_0}{\max(\boldsymbol{A}^k \boldsymbol{v}_0)}.$$

由(8.2)式有

$$\boldsymbol{A}^k \boldsymbol{v}_0 = \lambda_1^k \left(\alpha_1 \boldsymbol{x}_1 + \sum_{i=2}^{n} \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^k \boldsymbol{x}_i \right) = \lambda_1^k (\alpha_1 \boldsymbol{x}_1 + \boldsymbol{\varepsilon}_k),$$

$$\boldsymbol{u}_k = \frac{\boldsymbol{A}^k \boldsymbol{v}_0}{\max(\boldsymbol{A}^k \boldsymbol{v}_0)} = \frac{\lambda_1^k (\alpha_1 \boldsymbol{x}_1 + \boldsymbol{\varepsilon}_k)}{\max(\lambda_1^k (\alpha_1 \boldsymbol{x}_1 + \boldsymbol{\varepsilon}_k))}$$

$$= \frac{\alpha_1 \boldsymbol{x}_1 + \boldsymbol{\varepsilon}_k}{\max(\alpha_1 \boldsymbol{x}_1 + \boldsymbol{\varepsilon}_k)} \to \frac{\boldsymbol{x}_1}{\max(\boldsymbol{x}_1)} (k \to \infty).$$

同理,可得到

$$\boldsymbol{v}_k = \frac{\lambda_1^k (\alpha_1 \boldsymbol{x}_1 + \boldsymbol{\varepsilon}_k)}{\max(\lambda_1{}^{k-1} (\alpha_1 \boldsymbol{x}_1 + \boldsymbol{\varepsilon}_{k-1}))} = \frac{\lambda_1 (\alpha_1 \boldsymbol{x}_1 + \boldsymbol{\varepsilon}_k)}{\max(\alpha_1 \boldsymbol{x}_1 + \boldsymbol{\varepsilon}_{k-1})},$$

$$\max(\boldsymbol{v}_k) = \lambda_1 \frac{\max(\alpha_1 \boldsymbol{x}_1 + \boldsymbol{\varepsilon}_k)}{\max(\alpha_1 \boldsymbol{x}_1 + \boldsymbol{\varepsilon}_{k-1})} \to \lambda_1 (k \to \infty).$$

定理得证.

由定理的证明可见,幂法的收敛速度由 $\left| \frac{\lambda_2}{\lambda_1} \right|$ 的大小确定. 若 $\boldsymbol{A}$ 的特征值不满足(8.1)式,将有不同的情况. 如果 $\lambda_1 = \lambda_2 = \cdots = \lambda_r$,且 $|\lambda_r| > |\lambda_{r+1}|$,可以作类似的分析,对初始向量(8.2)和计算公式(8.3)有

$$\lim_{k \to \infty} \boldsymbol{u}_k = \frac{\sum_{i=1}^{r} \alpha_i \boldsymbol{x}_i}{\max(\sum_{i=1}^{r} \alpha_i \boldsymbol{x}_i)},$$

$$\lim_{k \to \infty} \max(\boldsymbol{v}_k) = \lambda_1.$$

可见,$\boldsymbol{u}_k$ 仍收敛于一个主特征向量. 对特征值的其他情况,讨论较为复杂. 完整的幂法程序要加上各种情况的判断.

例 **8.1**　用幂法求矩阵

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 0.5 \\ 1 & 1 & 0.25 \\ 0.5 & 0.25 & 2 \end{pmatrix}$$

的主特征值和主特征向量.

解　取初始向量 $\boldsymbol{u}_0 = (1,1,1)^{\mathrm{T}}$,按(8.3)式的计算结果如表 8-1.

表 8－1

k	$\boldsymbol{u}_k{}^T$	$\max(\boldsymbol{v}_k)$
0	(1.000 0,1.000 0,1)	
1	(0.909 1,0.818 2,1)	2.750 000 0
5	(0.765 1,0.667 4,1)	2.588 791 8
10	(0.749 4,0.650 8,1)	2.538 002 9
15	(0.748 3,0.649 7,1)	2.536 625 6
20	(0.748 2,0.649 7,1)	2.536 532 3

矩阵 $\boldsymbol{A}$ 的主特征值和主特征向量的准确值(8 位数字) 分别为 $\lambda_1 = 2.536\,525\,8$，$\boldsymbol{x}_1{}^* = (0.748\,221\,16, 0.649\,661\,16, 1)^{\mathrm{T}}$. 可见迭代 20 次后，所得的主特征值有 5 位有效数字.

从定理 8.4 的证明中易见，当 k 充分大时，有 $|\max(\boldsymbol{v}_k) - \lambda_1| \approx c\left|\dfrac{\lambda_2}{\lambda_1}\right|^k$. 因此，幂法是线性收敛的方法. 当 $|\lambda_2|$ 接近于 $|\lambda_1|$ 时，收敛很慢. 这时，一个补救的办法是采用加速收敛的办法. 下面简要地介绍两种加速方法.

(1) Aitken 外推法.

记 $m_k = \max(\boldsymbol{v}_k)$. 对幂法的计算结果进行外推加速

$$\widetilde{m}_k = m_m - \frac{(m_k - m_{k-1})^2}{m_k - 2m_{k-1} + m_{k-2}}, \quad k \geqslant 3,$$

$$(\widetilde{\boldsymbol{u}}_k)_j = (\boldsymbol{u}_k)_j - \frac{((\boldsymbol{u}_k)_j - (\boldsymbol{u}_{k-1})_j)^2}{(\boldsymbol{u}_k)_j - 2(\boldsymbol{u}_{k-1})_j + (\boldsymbol{u}_{k-2})_j}, \quad (\boldsymbol{u}_k)_j \neq 1.$$

(2) Rayleigh 商加速.

若 $\boldsymbol{A} \in \mathbf{R}^{n\times n}$ 为对称矩阵，则幂法所得的规范化向量 $\boldsymbol{u}_k$ 的 Rayleigh 商给出特征值 λ_1 的较好的近似值，

$$\frac{(\boldsymbol{A}\boldsymbol{u}_k, \boldsymbol{u}_k)}{(\boldsymbol{u}_k, \boldsymbol{u}_k)} = \lambda_1 + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{2k}\right).$$

8.2.2　反幂法和原点位移

反幂法用来计算矩阵按模最小的特征值及其特征向量. 设 $\boldsymbol{A} \in \mathbf{R}^{n\times n}$ 为非奇异矩阵，它的特征值满足

$$|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_{n-1}| > |\lambda_n| > 0, \tag{8.4}$$

则 $\boldsymbol{A}^{-1}$ 的特征值 $\lambda_1{}^{-1}, \lambda_2{}^{-1}, \cdots, \lambda_n{}^{-1}$ 满足

$$|\lambda_n{}^{-1}| > |\lambda_{n-1}{}^{-1}| \geqslant \cdots \geqslant |\lambda_1{}^{-1}|,$$

即 λ_n^{-1} 是 $\boldsymbol{A}^{-1}$ 的主特征值. 因此,对 $\boldsymbol{A}^{-1}$ 应用幂法可得矩阵 $\boldsymbol{A}$ 的按模最小的特征值及其特征向量,称为反幂法,我们有如下反幂法的实用计算公式为

$$\begin{cases} \boldsymbol{v}_0 = \boldsymbol{u}_0 \neq 0, \\ \boldsymbol{v}_k = \boldsymbol{A}^{-1}\boldsymbol{u}_{k-1}, \\ \boldsymbol{u}_k = \dfrac{\boldsymbol{v}_k}{\max(\boldsymbol{v}_k)}, \quad k = 1,2,\cdots. \end{cases} \tag{8.5}$$

在(8.5)式中,向量 $\boldsymbol{v}_k$ 可以通过解方程组 $\boldsymbol{A}\,\boldsymbol{v}_k = \boldsymbol{u}_{k-1}$ 得到. 这些方程组有相同的系数矩阵,为了节省计算工作量,可先对矩阵 $\boldsymbol{A}$ 进行三角分解 $\boldsymbol{A} = \boldsymbol{LU}$,再解三角形方程组

$$\boldsymbol{L}\boldsymbol{w}_k = \boldsymbol{u}_{k-1}, \quad \boldsymbol{U}\,\boldsymbol{v}_k = \boldsymbol{w}_k, \quad k = 1,2,\cdots.$$

定理 8.5 设非奇异矩阵 $\boldsymbol{A} \in \mathbf{R}^{n\times n}$ 的特征值 $\lambda_i (i = 1,2,\cdots,n)$ 满足(8.4)式,并且有对应的 n 个线性无关的特征向量 $\boldsymbol{x}_i (i = 1,2,\cdots,n)$. 给定初始向量 $\boldsymbol{v}_0 = \sum\limits_{i=1}^{n} \alpha_i \boldsymbol{x}_i, \alpha_n \neq 0$,则由(8.5)式生成的向量序列有

$$\lim_{k\to\infty} \boldsymbol{u}_k = \frac{\boldsymbol{x}_n}{\max(\boldsymbol{x}_n)}, \quad \lim_{k\to\infty} \max(\boldsymbol{v}_k) = \frac{1}{\lambda_n}.$$

反幂法的一个重要应用是利用"原点位移",求指定点附近的某个特征值和对应的特征向量.

如果矩阵 $(\boldsymbol{A} - p\boldsymbol{I})^{-1}$ 存在,显然其特征值为 $(\lambda_i - p)^{-1}, i = 1,2,\cdots,n$,对应的特征向量仍然是 $\boldsymbol{x}_i (i = 1,2,\cdots,n)$. 如果 p 是 $\boldsymbol{A}$ 的特征值 λ_j 的一个近似值,且

$$|\lambda_j - p| < |\lambda_i - p|, i \neq j, \tag{8.6}$$

即 $(\lambda_i - p)^{-1}$ 是 $(\boldsymbol{A} - p\boldsymbol{I})^{-1}$ 的主特征值,可用反幂法计算相应的特征值和特征向量,计算公式为

$$\begin{cases} \boldsymbol{u}_0 = \boldsymbol{v}_0 \neq 0, \\ \boldsymbol{v}_k = (\boldsymbol{A} - p\boldsymbol{I})^{-1}\boldsymbol{u}_{k-1}, \\ \boldsymbol{u}_k = \dfrac{\boldsymbol{v}_k}{\max(\boldsymbol{v}_k)}, \quad k = 1,2,\cdots. \end{cases} \tag{8.7}$$

定理 8.6 设 $\boldsymbol{A} \in \mathbf{R}^{n\times n}$ 的特征值 $\lambda_i (i = 1,2,\cdots,n)$ 对应的特征向量 $\boldsymbol{x}_i (i = 1,2,\cdots,n)$ 线性无关,p 为 λ_j 的近似值,满足(8.6)式,$(\boldsymbol{A} - p\boldsymbol{I})^{-1}$ 存在. 给定初始向量 $\boldsymbol{v}_0 = \sum\limits_{i=1}^{n} \alpha_i \boldsymbol{x}_i, \alpha_n \neq 0$,则由(8.7)式生成的向量序列有

$$\lim_{k\to\infty} \boldsymbol{u}_k = \frac{\boldsymbol{x}_j}{\max(\boldsymbol{x}_j)}, \quad \lim_{k\to\infty} \max(\boldsymbol{v}_k) = \frac{1}{\lambda_j - p}.$$

由该定理可知,$p + [\max(\boldsymbol{v}_k)]^{-1}$ 是特征值 λ_j 的近似值,对应的近似特征向

量为 $\boldsymbol{u}_k$. 迭代收敛速度由比值 $\sigma=\max\limits_{i\neq j}\left|\dfrac{\lambda_j-p}{\lambda_i-p}\right|$ 来确定.

反幂法迭代公式(8.7)中的 $\boldsymbol{v}_k$ 是通过解方程组

$$(\boldsymbol{A}-p\boldsymbol{I})\boldsymbol{v}_k=\boldsymbol{u}_{k-1}$$

求得的. 为了节省计算工作量,可以先将$(\boldsymbol{A}-p\boldsymbol{I})$ 进行三角分解

$$\boldsymbol{P}(\boldsymbol{A}-p\boldsymbol{I})=\boldsymbol{L}\boldsymbol{U},$$

其中 $\boldsymbol{P}$ 为排列阵.

只要选择的 p 是λ_j 的一个较好的近似且特征值分离情况较好,一般σ很小,收敛将是较快的. 实验表明,按下述方法选择 $\boldsymbol{v}_0=\boldsymbol{u}_0$ 是较好的:选 $\boldsymbol{u}_0$ 使

$$\boldsymbol{U}\boldsymbol{v}_1=\boldsymbol{L}^{-1}\boldsymbol{P}\boldsymbol{u}_0=(1,1,\cdots,1)^{\mathrm{T}},$$

用回代求解可得 $\boldsymbol{v}_1$.

例 8.2　用反幂法求下列矩阵的接近于 $p=1.267\,9$ 的特征值(精确特征值 $\lambda_3=3-\sqrt{3}$) 及其特征向量(用 5 位浮点数进行计算),

$$\boldsymbol{A}=\begin{pmatrix}2&1&0\\1&3&1\\0&1&4\end{pmatrix}.$$

解　用列选主元的三角分解将 $\boldsymbol{A}-p\boldsymbol{I}$ 分解为

$$\boldsymbol{P}(\boldsymbol{A}-p\boldsymbol{I})=\boldsymbol{L}\boldsymbol{U},$$

其中

$$\boldsymbol{P}=\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix},\quad \boldsymbol{L}=\begin{pmatrix}1&0&0\\0&1&0\\0.732\,1&-0.268\,07&1\end{pmatrix},$$

$$\boldsymbol{U}=\begin{pmatrix}1&1.732\,1&1\\0&1&2.732\,1\\0&0&0.294\,05\times10^{-3}\end{pmatrix}.$$

由 $\boldsymbol{U}\boldsymbol{v}_1=(1,1,1)^{\mathrm{T}}$ 得

$$\boldsymbol{v}_1=(12\,692,-9\,290.3,3\,400.8)^{\mathrm{T}},$$
$$\boldsymbol{u}_1=(1,-0.731\,98,0.267\,95)^{\mathrm{T}}.$$

由 $\boldsymbol{L}\boldsymbol{U}\boldsymbol{v}_2=\boldsymbol{P}\boldsymbol{u}_1$,得

$$\boldsymbol{v}_2=(20\,404,-14\,937,5\,467.4)^{\mathrm{T}},$$
$$\boldsymbol{u}_2=(1,-0.732\,06,0.267\,96)^{\mathrm{T}}.$$

由此可得特征值 $\lambda_3(=1.267\,949\,2)$ 的近似值为

$$1.267\,9+\frac{1}{20\,404}=1.267\,949.$$

λ_3 对应的特征向量是

$$\boldsymbol{x}_3 = (1, 1-\sqrt{3}, 2-\sqrt{3})^{\mathrm{T}} \approx (1, -0.732\,05, 0.267\,95)^{\mathrm{T}}.$$

由此可见，$\boldsymbol{u}_2$ 是 $\boldsymbol{x}_3$ 的相当好的近似.

例 **8.3** 设矩阵 $\boldsymbol{A}$ 的特征值都是实数，满足

$$\lambda_1 > \lambda_2 \geqslant \cdots \geqslant \lambda_n.$$

若取 $p = \dfrac{1}{2}(\lambda_2 + \lambda_n)$ 进行原点位移，则求 λ_1 的乘幂法收敛最快.

证 取 p 进行原点位移，矩阵 $\boldsymbol{B} = \boldsymbol{A} - p\boldsymbol{I}$ 的特征值为 $\lambda_1 - p, \lambda_2 - p, \cdots, \lambda_n - p$。为求 λ_1，必须选择 p，满足

$$|\lambda_1 - p| > |\lambda_j - p|, \quad j = 2, 3, \cdots, n.$$

根据已知条件，要使乘幂法收敛最快，应取 p 使

$$\max\left\{\left|\frac{\lambda_2 - p}{\lambda_1 - p}\right|, \left|\frac{\lambda_n - p}{\lambda_1 - p}\right|\right\}$$

达到最小. 因此，必有

$$|\lambda_2 - p| = |\lambda_n - p|.$$

由此解得 $p = \dfrac{1}{2}(\lambda_2 + \lambda_n)$. 此时有

$$\left|\frac{\lambda_2 - p}{\lambda_1 - p}\right| = \left|\frac{\lambda_n - p}{\lambda_1 - p}\right| = \frac{\lambda_2 - \lambda_n}{2\lambda_1 - \lambda_2 - \lambda_n} < 1.$$

8.3 Jacobi 方 法

我们知道，若矩阵 $\boldsymbol{A} \in \boldsymbol{R}^{n\times n}$ 为对称矩阵，则存在一正交阵 $\boldsymbol{P}$，使

$$\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{\mathrm{T}} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n) = \boldsymbol{D}.$$

$\boldsymbol{D}$ 的对角元 $\lambda_i (i = 1, 2, \cdots, n)$ 就是 $\boldsymbol{A}$ 的特征值，$\boldsymbol{P}^{\mathrm{T}}$ 的列向量就是对应于特征值的特征向量. 于是求实对称矩阵 $\boldsymbol{A}$ 的特征值问题就转为寻找正交矩阵 $\boldsymbol{P}$，使 $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{\mathrm{T}} = \boldsymbol{D}$ 为对角阵，而这个问题的主要困难在于如何构造 $\boldsymbol{P}$.

Jacobi 方法是用来计算实对称矩阵的全部特征值及对应的特征向量的一种变换方法，其基本思想是对矩阵作一系列正交相似变换，使其非对角线元素收敛到零. 所用的变换是 Jacobi 旋转变换. 下面先讨论 Jacobi 旋转变换及其性质.

在 $\mathbf{R}^n$ 中的 $\{x_i, x_j\}$ 平面内的平面旋转变换为

$$\begin{cases} y_i = x_i\cos\theta + x_j\sin\theta, \\ y_j = -x_i\sin\theta + x_j\cos\theta, \\ y_k = x_k, \quad k \neq i, j, \end{cases}$$

或写成 $\boldsymbol{y} = \boldsymbol{J}\boldsymbol{x}$，其中，$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$，$\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^{\mathrm{T}}$，

$$\boldsymbol{J}=\boldsymbol{J}(i,j)=\begin{pmatrix}1&&&&&&\\&\ddots&&&&&\\&&c&\cdots&s&&\\&&\vdots&\ddots&\vdots&&\\&&-s&\cdots&c&&\\&&&&&\ddots&\\&&&&&&1\end{pmatrix}.$$

称 $\boldsymbol{J}$ 为平面旋转矩阵，它只有在 $(i,i),(i,j),(j,i)$ 和 (j,j) 位置上的元素与单位矩阵不一样，分别为 $c=\cos\theta, s=\sin\theta, -\sin\theta$ 和 $\cos\theta$.

显然，矩阵 $\boldsymbol{J}$ 是正交矩阵，$\boldsymbol{JA}$ 只改变 $\boldsymbol{A}$ 的第 i 行与第 j 行的元素，$\boldsymbol{AJ}^{\mathrm{T}}$ 只改变 $\boldsymbol{A}$ 的第 i 列与第 j 列的元素，$\boldsymbol{JAJ}^{\mathrm{T}}$ 只改变 $\boldsymbol{A}$ 的第 i 行、第 j 行、第 i 列、第 j 列的元素.

设 $\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{n\times n}$ 为对称矩阵，$\boldsymbol{J}(i,j)$ 为一平面旋转矩阵，则 $\boldsymbol{B}=\boldsymbol{JAJ}^{\mathrm{T}}=(b_{ij})$ 的元素的计算公式为

$$\begin{aligned}
b_{ii}&=a_{ii}\cos^2\theta+a_{jj}\sin^2\theta+2a_{ij}\sin\theta\cos\theta,\\
b_{jj}&=a_{ii}\sin^2\theta+a_{jj}\cos^2\theta-2a_{ij}\sin\theta\cos\theta,\\
b_{ij}&=b_{ji}=\frac{1}{2}(a_{jj}-a_{ii})\sin2\theta+a_{ij}\cos2\theta,\\
b_{ik}&=b_{ki}=a_{ik}\cos\theta+a_{jk}\sin\theta,\quad k\neq i,j,\\
b_{jk}&=b_{kj}=a_{jk}\cos\theta-a_{ik}\sin\theta,\quad k\neq i,j,\\
b_{lm}&=b_{ml}=a_{lm},\quad l\neq i,j;\ m\neq i,j.
\end{aligned}$$

而且，不难验证

$$b_{ii}^2+b_{jj}^2+2b_{ij}^2=a_{ii}^2+a_{jj}^2+2a_{ij}^2. \tag{8.8}$$

定理 8.7　设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 为对称矩阵，若 $\boldsymbol{B}=\boldsymbol{PAP}^{\mathrm{T}}$，$\boldsymbol{P}$ 为正交阵，则有 $\|\boldsymbol{B}\|_{\mathrm{F}}=\|\boldsymbol{A}\|_{\mathrm{F}}$.

证　设 $\lambda_i(i=1,2,\cdots,n)$ 为 $\boldsymbol{A}$ 的特征值，则

$$\|\boldsymbol{A}\|_{\mathrm{F}}^2=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}^2=\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})=\mathrm{tr}(\boldsymbol{A}^2)=\sum_{i=1}^{n}\lambda_i^2.$$

另一方面，矩阵 $\boldsymbol{B}$ 的特征值也为 $\lambda_i(i=1,2,\cdots,n)$，

$$\|\boldsymbol{B}\|_{\mathrm{F}}^2=\mathrm{tr}(\boldsymbol{B}^2)=\sum_{i=1}^{n}\lambda_i^2.$$

因此，$\|\boldsymbol{B}\|_{\mathrm{F}}^2=\|\boldsymbol{A}\|_{\mathrm{F}}^2$. 定理得证.

设 $\boldsymbol{A}$ 的非对角元素 $a_{ij}\neq0$，我们可选择平面旋转矩阵 $\boldsymbol{J}(i,j)$，使 $\boldsymbol{B}=\boldsymbol{JAJ}^{\mathrm{T}}$ 的非对角元素 $b_{ij}=b_{ji}=0$. 为此，由矩阵 $\boldsymbol{B}$ 的元素的计算公式可知，可选择 θ，使

$$\tan 2\theta = \frac{2a_{ij}}{a_{ii} - a_{jj}}, \quad |\theta| \leqslant \frac{\pi}{4}. \tag{8.9}$$

如果用 $D(\boldsymbol{A})$ 表示 $\boldsymbol{A}$ 的对角线元素的平方和，用 $S(\boldsymbol{A})$ 表示 $\boldsymbol{A}$ 的非对角线元素的平方和，则对 $\boldsymbol{B} = \boldsymbol{J}\boldsymbol{A}\boldsymbol{J}^{\mathrm{T}}$，由(8.8)、(8.9)式和定理 8.7 可知

$$D(\boldsymbol{B}) = D(\boldsymbol{A}) + 2a_{ij}^2,$$

$$S(\boldsymbol{B}) = S(\boldsymbol{A}) - 2a_{ij}^2.$$

这说明 $\boldsymbol{B}$ 的对角线元素的平方和比 $\boldsymbol{A}$ 的对角线元素的平方和增加了 $2a_{ij}^2$，而 $\boldsymbol{B}$ 的非对角线元素的平方和比 $\boldsymbol{A}$ 的非对角线元素的平方和减少了 $2a_{ij}^2$. 这就是 Jacobi 方法求矩阵特征值和特征向量的依据. 下面说明 Jacobi 方法的计算过程.

先在 $\boldsymbol{A} = \boldsymbol{A}_0 = (a_{ij}^{(0)})$ 中选择非对角元中绝对值最大的 $a_{ij}^{(0)}$. 可设 $a_{ij}^{(0)} \neq 0$，否则 $\boldsymbol{A}$ 已经对角化了. 由(8.9)式选择平面旋转矩阵 $\boldsymbol{J}_1$，使 $\boldsymbol{J}_1\boldsymbol{A}_0\boldsymbol{J}_1^{\mathrm{T}} = \boldsymbol{A}_1$ 的元素 $a_{ij}^{(1)} = 0$. 计算出 $\boldsymbol{A}_1$，再类似地选择 $\boldsymbol{J}_2$，计算 $\boldsymbol{A}_2 = \boldsymbol{J}_2\boldsymbol{A}_1\boldsymbol{J}_2^{\mathrm{T}}$，继续这个过程，连续对 $\boldsymbol{A}$ 施行一系列平面旋转变换，消除非对角线绝对值最大的元素，直到将 $\boldsymbol{A}$ 的非对角线元素全化为充分小为止.

定理 8.8 设 $\boldsymbol{A} \in \mathbf{R}^{n\times n}$ 为对称矩阵，对 $\boldsymbol{A}$ 施行上述一系列平面旋转变换

$$\boldsymbol{A}_m = \boldsymbol{J}_m\boldsymbol{A}_{m-1}\boldsymbol{J}_m{}^{\mathrm{T}}, \quad m = 1,2,\cdots,$$

则有 $\lim\limits_{m\to\infty} S(\boldsymbol{A}_m) = 0$.

证 设 $|a_{ij}^{(m)}| = \max\limits_{l\neq k} |a_{lk}^{(m)}|$，由于

$$S(\boldsymbol{A}_{m+1}) = S(\boldsymbol{A}_m) - 2(a_{ij}^{(m)})^2,$$

$$S(\boldsymbol{A}_m) = \sum_{l\neq k} (a_{lk}^{(m)})^2 \leqslant n(n-1)(a_{ij}^{(m)})^2,$$

则有

$$S(\boldsymbol{A}_{m+1}) \leqslant S(\boldsymbol{A}_m)\left(1 - \frac{2}{n(n-1)}\right).$$

反复利用上式，即得

$$S(\boldsymbol{A}_{m+1}) \leqslant S(\boldsymbol{A}_0)\left(1 - \frac{2}{n(n-1)}\right)^{m+1}, \quad n > 2.$$

故 $\lim\limits_{m\to\infty} S(\boldsymbol{A}_m) = 0$，定理得证.

我们指出，可以证明 $\boldsymbol{A}_m$ 的对角线元素一定有极限.

设 m 充分大时，有

$$\boldsymbol{A}_m = \boldsymbol{J}_m\cdots\boldsymbol{J}_2\boldsymbol{J}_1\boldsymbol{A}\boldsymbol{J}_1^{\mathrm{T}}\boldsymbol{J}_2^{\mathrm{T}}\cdots\boldsymbol{J}_m^{\mathrm{T}} \approx D,$$

$\boldsymbol{D}$ 为对角阵，则 $\boldsymbol{A}_m$ 的对角线元素就是 $\boldsymbol{A}$ 的近似特征值，$\boldsymbol{Q}_m = \boldsymbol{J}_1^{\mathrm{T}}\boldsymbol{J}_2^{\mathrm{T}}\cdots\boldsymbol{J}_m^{\mathrm{T}}$ 的列向量就是对应的近似特征向量. 可用 $S(\boldsymbol{A}_m) < \varepsilon$ 控制迭代终止，其中 ε 是要求的精度.

例 8.4　用 Jacobi 方法计算矩阵

$$\boldsymbol{A}=\begin{pmatrix}2&-1&0\\-1&2&-1\\0&-1&2\end{pmatrix}$$

的特征值的特征向量.

解　先取$(i,j)=(1,2)$,按(8.9)式有 $\cot 2\theta=0$,$s=c=\dfrac{1}{\sqrt{2}}$,所以

$$\boldsymbol{J}_1=\begin{pmatrix}\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}&0\\-\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}&0\\0&0&1\end{pmatrix},\quad \boldsymbol{A}_1=\boldsymbol{J}_1\boldsymbol{A}\boldsymbol{J}_1^{\mathrm{T}}=\begin{pmatrix}1&0&-\frac{1}{\sqrt{2}}\\0&3&-\frac{1}{\sqrt{2}}\\-\frac{1}{\sqrt{2}}&-\frac{1}{\sqrt{2}}&2\end{pmatrix}.$$

再取$(i,j)=(1,3)$,有 $\cot 2\theta=\dfrac{1}{\sqrt{2}}$,$c=0.888\,08$,$s=0.459\,70$,

$$\boldsymbol{J}_2=\begin{pmatrix}0.888\,08&0&0.459\,70\\0&1&0\\-0.459\,70&0&0.888\,08\end{pmatrix},$$

$$\boldsymbol{A}_2=\boldsymbol{J}_2\boldsymbol{A}_1\boldsymbol{J}_2^{\mathrm{T}}=\begin{pmatrix}0.633\,98&-0.325\,05&0\\-0.325\,05&3&-0.627\,97\\0&-0.627\,97&2.366\,03\end{pmatrix}.$$

这里我们看到经变换后所得矩阵的非对角线元素的最大绝对值逐次变小. 继续做下去,可以得到

$$\boldsymbol{A}_9=\begin{pmatrix}0.585\,78&0.000\,00&0.000\,00\\0.000\,00&2.000\,00&0.000\,00\\0.000\,00&0.000\,00&3.414\,21\end{pmatrix},$$

$$\boldsymbol{Q}_9=\boldsymbol{J}_1^{\mathrm{T}}\boldsymbol{J}_2^{\mathrm{T}}\cdots\boldsymbol{J}_9^{\mathrm{T}}=\begin{pmatrix}0.500\,000&0.707\,10&0.500\,00\\0.707\,10&0.000\,00&-0.707\,10\\0.500\,00&-0.707\,10&0.500\,00\end{pmatrix}.$$

矩阵 $\boldsymbol{A}$ 的近似特征值和特征向量均已求出,即得矩阵 $\boldsymbol{A}$ 的近似特征值

$$\lambda_1\approx 0.585\,78,\quad \lambda_2\approx 2.000\,00,\quad \lambda_3\approx 3.414\,21,$$

相应的近似特征向量

$$x_1\approx(0.500\,00,0.707\,10,0.500\,00)^{\mathrm{T}},$$

$$x_2\approx(0.707\,10,0.000\,00,-0.707\,10)^{\mathrm{T}},$$

$$x_3 \approx (0.500\,00, -0.707\,10, 0.500\,00)^{\mathrm{T}}.$$

矩阵 $\boldsymbol{A}$ 的特征值的精确值为

$$\lambda_1 = 2-\sqrt{2}, \quad \lambda_2 = 2, \quad \lambda_3 = 2+\sqrt{2},$$

相应的特征向量为

$$x_1 = \left(\frac{1}{2}, \frac{1}{\sqrt{2}}, \frac{1}{2}\right)^{\mathrm{T}},$$

$$x_2 = \left(\frac{1}{\sqrt{2}}, 0, -\frac{1}{\sqrt{2}}\right)^{\mathrm{T}},$$

$$x_3 = \left(\frac{1}{2}, -\frac{1}{\sqrt{2}}, \frac{1}{2}\right)^{\mathrm{T}}.$$

由此可见，Jacobi 方法变换 9 次后的结果已经相当精确了。

用 Jacobi 方法求得的结果精度一般都比较高，特别是求得的特征向量正交性很好. 所以 Jacobi 方法是求实对称矩阵全部特征值和特征向量的一个较好的方法. 它的弱点是计算量大，对原矩阵是稀疏矩阵，旋转变换后不能保持其稀疏的性质.

由于 Jacobi 方法在每次寻找非对角元素的绝对值最大者时，要花费很多计算机时间，因此提出了不少改进的方法，常采用的一种就是 Jacobi 过关法. 这种方法是选取一个单调减小而趋于零的数列 $\{a_m\}$ 作为限值，这些限值称为“关”. 常用的取法是，对 $N \geqslant n$，取

$$a_1 = \frac{\sqrt{S(\boldsymbol{A})}}{N}$$

在 $\boldsymbol{A}$ 的非对角线元素中按行(或列) 扫描，碰到绝对值小于 a_1 的元素就跳过去，否则就做变换将其化零. 重复上述过程，可能要经过多遍扫描，直到所有的非对角线元素的绝对值都小于 a_1 为止. 再取 $a_2, a_3, \cdots$ 类似处理，直到所有的非对角线元素的绝对值都小于 a_m 时，迭代停止，这里的 a_m 应小于给定的精度要求.

8.4 QR 算 法

8.4.1 化矩阵为 Hessenberg 形

对于实对称矩阵，可通过正交相似变换约化为对角矩阵。那么，对于一般的实矩阵，通过正交相似变换可约化到什么程度呢?线性代数中有如下结果.

定理 8.9(实 Schur 定理) 对于任何矩阵 $\boldsymbol{A} \in \mathbf{R}^{n\times n}$，存在正交矩阵 $\boldsymbol{Q}$，使得

$$\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Q}=\begin{pmatrix}\mathbf{R}_{11} & \mathbf{R}_{12} & \cdots & \mathbf{R}_{1m}\\ & \mathbf{R}_{22} & \cdots & \mathbf{R}_{2m}\\ & & \ddots & \vdots\\ & & & \mathbf{R}_{mm}\end{pmatrix},$$

其中的对角块 $\mathbf{R}_{ii}(i=1,2,\cdots,m)$ 为一阶或二阶方阵，每一个一阶对角块即为 $\boldsymbol{A}$ 的实特征值，每一个二阶对角块的两个特征值是 $\boldsymbol{A}$ 的一对共轭复特征值.

我们称这种分块上三角阵为矩阵 $\boldsymbol{A}$ 的 Schur 分块上三角形，上三角阵和对角阵是它的特殊情形. 定理 8.9 并没有解决如何计算全部特征值的问题. 为了节省运算工作量，实用的方法是先将矩阵约化为与 Schur 分块上三角形很接近的 Hessenberg 形.

定义 8.2　若矩阵 $\boldsymbol{B}=(b_{ij})\in\mathbf{R}^{n\times n}$ 的次对角线以下的 $b_{ij}=0(i>j+1)$，则称 $\boldsymbol{B}$ 为上 Hessenberg 矩阵，简称 Hessenberg 形，即 $\boldsymbol{B}$ 的形状为

$$\boldsymbol{B}=\begin{pmatrix}* & * & \cdots & \cdots & *\\ * & * & \cdots & \cdots & *\\ & \ddots & \ddots & & \vdots\\ & & \ddots & \ddots & \vdots\\ & & & * & *\end{pmatrix}.$$

若 $\boldsymbol{B}$ 有一个次对角元如 $b_{k+1,k}=0(1\leqslant k\leqslant n-1)$，则 $\boldsymbol{B}$ 是可约的，否则是不可约的. 对于可约的 Hessenberg 形，可把求解特征值的问题，约简成求解较小的矩阵的特征值问题.

可以用平面旋转变换化矩阵为 Hessenberg 形. 下面介绍另一种正交变换.

定义 8.3　设向量 $\boldsymbol{w}\in\mathbf{R}^n$，$\|\boldsymbol{w}\|_2=1$，则称

$$H(\boldsymbol{w})=\boldsymbol{I}-2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}} \tag{8.10}$$

为(初等)镜面反射矩阵，或 Householder 变换矩阵.

Householder 矩阵 $\boldsymbol{H}=\boldsymbol{H}(\boldsymbol{w})$ 有如下性质：

(1) $\boldsymbol{H}$ 是对称正交阵，即 $\boldsymbol{H}=\boldsymbol{H}^{\mathrm{T}}=\boldsymbol{H}^{-1}$. 事实上，显然有 $\boldsymbol{H}^{\mathrm{T}}=\boldsymbol{H}$，又由 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}=\|\boldsymbol{w}\|_2=1$ 得知

$$\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{H}^2=\boldsymbol{I}-4\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}+4\boldsymbol{w}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w})\boldsymbol{w}^{\mathrm{T}}=\boldsymbol{I}.$$

(2) 对任何 $\boldsymbol{x}\in\mathbf{R}^{n\times n}$，记 $\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}$，有 $\|\boldsymbol{y}\|_2=\|\boldsymbol{x}\|_2$.

(3) 记 S 为与 $\boldsymbol{w}$ 垂直的平面，则几何上 $\boldsymbol{x}$ 与 $\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}$ 关于平面 S 对称. 事实上，由 $\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}=(\boldsymbol{I}-2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})\boldsymbol{x}$ 得知

$$\boldsymbol{x}-\boldsymbol{y}=2(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})\boldsymbol{w}.$$

上式表明向量 $\boldsymbol{x}-\boldsymbol{y}$ 与 $\boldsymbol{w}$ 平行，注意到 $\boldsymbol{y}$ 与 $\boldsymbol{x}$ 的长度相等，于是 $\boldsymbol{x}$ 经过变换后的象 $\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}$ 是 $\boldsymbol{x}$ 关于 S 对称的向量，如图 8-1 所示.

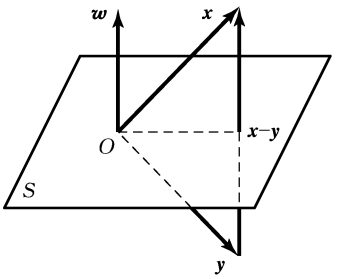

图 8－1

对应于性质(2),有下面的定理.

定理 8.10 设 $\boldsymbol{x},\boldsymbol{y}\in\mathbf{R}^n$,$\boldsymbol{x}\neq\boldsymbol{y}$ 且 $\|\boldsymbol{x}\|_2=\|\boldsymbol{y}\|_2$,则有镜面反射矩阵 $\boldsymbol{H}$,使 $\boldsymbol{Hx}=\boldsymbol{y}$.

证 令 $\boldsymbol{w}=\dfrac{\boldsymbol{x}-\boldsymbol{y}}{\|\boldsymbol{x}-\boldsymbol{y}\|_2}$,$\boldsymbol{H}=\boldsymbol{I}-2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}$. 则有 $\|\boldsymbol{w}\|_2=1$. 由 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}=\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}$ 知

$$2(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}\boldsymbol{x}=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}-2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}+\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}=(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y})=\|\boldsymbol{x}-\boldsymbol{y}\|_2^2.$$

由此可得

$$\boldsymbol{Hx}=\boldsymbol{x}-2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}=\boldsymbol{x}-\frac{2(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}\boldsymbol{x}}{\|\boldsymbol{x}-\boldsymbol{y}\|_2^2}=\boldsymbol{x}-(\boldsymbol{x}-\boldsymbol{y})=\boldsymbol{y}.$$

定理得证.

该定理的一个重要应用是对 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}\neq 0$,有镜面反射矩阵 $\boldsymbol{H}$,使得

$$\boldsymbol{Hx}=\sigma\boldsymbol{e}_1. \tag{8.11}$$

其中 $\sigma=-\operatorname{sign}(x_1)\|\boldsymbol{x}\|_2$,$\boldsymbol{e}_1=(1,0,\cdots,0)^{\mathrm{T}}$. 矩阵 $\boldsymbol{H}$ 的计算公式为

$$\begin{cases}\boldsymbol{u}=\boldsymbol{x}-\sigma\boldsymbol{e}_1,\\ \rho=\sigma(\sigma-x_1),\\ \boldsymbol{H}=\boldsymbol{I}-\rho^{-1}\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}.\end{cases} \tag{8.12}$$

关于 σ 的符号的选取,是为了使作为分母的 ρ 尽量大,从而有利于数值计算的稳定性. (8.11) 式的意义是对向量作消元运算. 与平面旋转不同的是,镜面反射变换可成批消去向量的非零元.

例 8.5 对于向量 $\boldsymbol{x}=(3,5,1,1)^{\mathrm{T}}$,构造镜面反射阵 $\boldsymbol{H}$,使

$$\boldsymbol{Hx}=-\operatorname{sign}(x_1)\|\boldsymbol{x}\|_2(1,0,0,0)^{\mathrm{T}}.$$

解 $\|\boldsymbol{x}\|_2=6$,$\sigma=-\operatorname{sign}(x_1)\|\boldsymbol{x}\|_2=-6$,$\boldsymbol{u}=\boldsymbol{x}-\sigma\boldsymbol{e}_1=(9,5,1,1)^{\mathrm{T}}$. $\|\boldsymbol{u}\|_2=108$,$\rho=54$. 按(8.12)式得

$$\boldsymbol{H}=\boldsymbol{I}-\rho^{-1}\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}=\frac{1}{54}\begin{pmatrix}-27 & -45 & -9 & -9\\ -45 & 29 & -5 & -5\\ -9 & -5 & 53 & -1\\ -9 & -5 & -1 & 53\end{pmatrix}.$$

定理 8.11　对于任何矩阵 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$，存在正交阵 $\boldsymbol{Q}$，使得

$$\boldsymbol{B}=\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Q} \tag{8.13}$$

为 Hessenberg 形.

证　设 $\boldsymbol{A}_1=\boldsymbol{A}$，$\boldsymbol{a}_1$ 为 $\boldsymbol{A}_1$ 的第 1 列对角线以下(不含对角线)的 $n-1$ 维向量. 根据(8.11)式，可构造 $n-1$ 阶对称正交阵 $\boldsymbol{H}_1$，使得 $\boldsymbol{H}_1\boldsymbol{a}_1=\sigma_1\boldsymbol{e}_1$，其中 $\boldsymbol{e}_1=(1,0,\cdots,0)^{\mathrm{T}}\in\mathbf{R}^{n-1}$. 记 $\boldsymbol{P}_1=\mathrm{diag}(1,\boldsymbol{H}_1)$，显然 $\boldsymbol{P}_1$ 是对称正交阵，$\boldsymbol{P}_1^{-1}=\boldsymbol{P}_1^{\mathrm{T}}=\boldsymbol{P}_1$. 用 $\boldsymbol{P}_1$ 对 $\boldsymbol{A}_1$ 作相似交换，由于 $\boldsymbol{P}_1\boldsymbol{A}\boldsymbol{P}_1^{-1}=\boldsymbol{P}_1\boldsymbol{A}\boldsymbol{P}_1$ 不改变矩阵 $\boldsymbol{P}_1\boldsymbol{A}$ 的第 1 列，而 $\boldsymbol{P}_1\boldsymbol{A}$ 的第 1 列中的第 2 个元素以后的元素全为零，易知

$$\boldsymbol{A}_2=\boldsymbol{P}_1\boldsymbol{A}_1\boldsymbol{P}_1=\begin{pmatrix}*&*&\cdots&*\\ \sigma_1&*&\cdots&*\\ &*&\cdots&*\\ &\vdots&&\vdots\\ &*&\cdots&*\end{pmatrix}.$$

记 $\boldsymbol{a}_2$ 为 $\boldsymbol{A}_2$ 的第 2 列对角线以下(不含对角线)的 $n-2$ 维向量，那么同理可构造 $n-2$ 阶对称正交阵 $\boldsymbol{H}_2$，使得 $\boldsymbol{H}_2\boldsymbol{a}_2=\sigma_2\boldsymbol{e}_1$，其中 $\boldsymbol{e}_1=(1,0,\cdots,0)^{\mathrm{T}}\in\mathbf{R}^{n-2}$. 记 $\boldsymbol{I}_2$ 为 2 阶单位向量，$\boldsymbol{P}_2=\mathrm{diag}(\boldsymbol{I}_2,\boldsymbol{H}_2)$，显然 $\boldsymbol{P}_2$ 是对称正交阵. 用 $\boldsymbol{P}_2$ 对 $\boldsymbol{A}_2$ 作相似变换，有

$$\boldsymbol{A}_3=\boldsymbol{P}_2\boldsymbol{A}_2\boldsymbol{P}_2=\begin{pmatrix}*&*&*&\cdots&*\\ \sigma_1&*&*&\cdots&*\\ &\sigma_2&*&\cdots&*\\ &&*&\cdots&*\\ &&\vdots&&\vdots\\ &&*&\cdots&*\end{pmatrix}.$$

如此类推，经 $n-2$ 步对称正交相似交换，得到 Hessenberg 形矩阵

$$\boldsymbol{A}_{n-1}=\boldsymbol{P}_{n-2}\boldsymbol{A}_{n-2}\boldsymbol{P}_{n-2}=\boldsymbol{P}_{n-2}\cdots\boldsymbol{P}_2\boldsymbol{P}_1\boldsymbol{A}\boldsymbol{P}_1\boldsymbol{P}_2\cdots\boldsymbol{P}_{n-2}.$$

若记 $\boldsymbol{B}=\boldsymbol{A}_{n-1}$，$\boldsymbol{Q}=\boldsymbol{P}_1\boldsymbol{P}_2\cdots\boldsymbol{P}_{n-2}$，则有(8.13)式. 定理得证.

推论 8.1　对于任何对称矩阵 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$，存在正交阵 $\boldsymbol{Q}$，使得 $\boldsymbol{B}=\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Q}$ 为对称三角阵.

上述定理 8.11 的证明是构造性的，即可以用镜面反射变换化矩阵为 Hessenberg 形. 此定理也可用平面旋转变换来证明，即也可用平面旋转变换化

矩阵为 Hessenberg 形. 对于阶数大于 3 的矩阵，第 1 列被消元的向量 $\boldsymbol{a}_1$ 的维数大于 2. 这时，可以连续使用平面旋转变换，把 $\boldsymbol{a}_1$ 的从第 2 个分量开始的非零元素逐个化为零. 如此类推，最后得到的正交矩阵 $\boldsymbol{Q}$，是平面旋转矩阵的乘积.

8.4.2 QR 算法及其收敛性

QR 算法可以用来求任意非奇异实矩阵的全部特征值，是目前计算这类问题最有效的方法之一. 它基于对任何非奇异实矩阵都可以分解为正交矩阵 $\boldsymbol{Q}$ 和上三角矩阵 $\mathbf{R}$ 的乘积.

定理 8.12（QR 分解定理） 设 $\boldsymbol{A} \in \mathbf{R}^{n\times n}$ 为非奇异矩阵，则存在正交阵 $\boldsymbol{Q}$ 与上三角阵 $\boldsymbol{R}$，使得 $\boldsymbol{A} = \boldsymbol{QR}$，且当 $\boldsymbol{R}$ 的对角元素均取正时，分解是唯一的.

证 类似于定理 8.11 的证明，对矩阵 $\boldsymbol{A}$ 左乘一系列正交变换矩阵，可以将 $\boldsymbol{A}$ 化为上三角形矩阵，因此，可得 $\boldsymbol{A}$ 的 QR 分解. 下面证明分解的唯一性. 设有两种分解

$$\boldsymbol{A} = \boldsymbol{Q}_1\boldsymbol{R}_1 = \boldsymbol{Q}_2\boldsymbol{R}_2,$$

而且 $\boldsymbol{R}_1, \boldsymbol{R}_2$ 的对角元均为正数. 由此可得

$$\boldsymbol{Q}_2{}^{\mathrm{T}}\boldsymbol{Q}_1 = \boldsymbol{R}_2\boldsymbol{R}_1{}^{-1}.$$

上式左边为正交阵，所以右边为正交阵，即

$$(\boldsymbol{R}_2\boldsymbol{R}_1{}^{-1})^{\mathrm{T}} = (\boldsymbol{R}_2\boldsymbol{R}_1{}^{-1})^{-1}.$$

这个式子左边是下三角阵，而右边是上三角阵，所以只能是对角阵. 设

$$\boldsymbol{D} = \boldsymbol{R}_2\boldsymbol{R}_1{}^{-1} = \mathrm{diag}(d_1, d_2, \cdots, d_n),$$

则有 $\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}} = \boldsymbol{D}^2 = \boldsymbol{I}$，且 $d_i > 0, i = 1, 2, \cdots, n$. 故有 $\boldsymbol{D} = \boldsymbol{I}$，从而 $\boldsymbol{R}_2 = \boldsymbol{R}_1$，进而 $\boldsymbol{Q}_2 = \boldsymbol{Q}_1$，定理得证.

一般按平面旋转变换或镜面反射变换作出的分解 $\boldsymbol{A} = \boldsymbol{QR}$，$\boldsymbol{R}$ 的对角元不一定是正的. 设 $\boldsymbol{R} = (r_{ij})$，只要令

$$\boldsymbol{D} = \mathrm{diag}\left(\frac{r_{11}}{|r_{11}|}, \frac{r_{22}}{|r_{22}|}, \cdots, \frac{r_{mm}}{|r_{mm}|}\right),$$

$\overline{\boldsymbol{Q}} = \boldsymbol{QD}$ 为正交阵，$\overline{\boldsymbol{R}} = \boldsymbol{D}^{-1}\boldsymbol{R}$ 为对角元是 $|r_{ii}|$ 的上三角阵，这样，$\boldsymbol{A} = \overline{\boldsymbol{Q}}\,\overline{\boldsymbol{R}}$ 就是符合定理 8.12 的唯一 QR 分解.

设有 $\boldsymbol{A}$ 的 QR 分解，即 $\boldsymbol{A} = \boldsymbol{QR}$，那么令 $\boldsymbol{B} = \boldsymbol{RQ}$，则有 $\boldsymbol{B} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{AQ}$. 这说明 $\boldsymbol{B}$ 与 $\boldsymbol{A}$ 有相同的特征值. 对 $\boldsymbol{B}$ 继续作 QR 分解，又可得一新的矩阵. 令 $\boldsymbol{A}_1 = \boldsymbol{A}$，得如下的算法：

$$\begin{cases} \boldsymbol{A}_k = \boldsymbol{Q}_k\boldsymbol{R}_k, \\ \boldsymbol{A}_{k+1} = \boldsymbol{R}_k\boldsymbol{Q}_k, \quad k = 1, 2, \cdots. \end{cases} \tag{8.14}$$

由(8.14)式得到矩阵序列$\{\boldsymbol{A}_k\}$的方法称为 QR 算法，或称为基本 QR 算法.

定理 8.13　QR 算法产生的序列$\{\boldsymbol{A}_k\}$满足：

(1) $\boldsymbol{A}_{k+1}=\boldsymbol{Q}_k{}^{\mathrm{T}}\boldsymbol{A}_k\boldsymbol{Q}_k$；

(2) $\boldsymbol{A}^k=\overline{\boldsymbol{Q}}_k\overline{\boldsymbol{R}}_k$，

其中$\overline{\boldsymbol{Q}}_k=\boldsymbol{Q}_1\boldsymbol{Q}_2\cdots\boldsymbol{Q}_k$，$\overline{\boldsymbol{R}}_k=\boldsymbol{R}_k\cdots\boldsymbol{R}_2\boldsymbol{R}_1$.

证　容易证(1). 从它递推得

$$\begin{aligned}\boldsymbol{A}_k&=\boldsymbol{Q}_{k-1}{}^{\mathrm{T}}\boldsymbol{A}_{k-1}\boldsymbol{Q}_{k-1}\\&=(\boldsymbol{Q}_1\boldsymbol{Q}_2\cdots\boldsymbol{Q}_{k-1})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{Q}_1\boldsymbol{Q}_2\cdots\boldsymbol{Q}_{k-1})=\overline{\boldsymbol{Q}}_{k-1}{}^{\mathrm{T}}\boldsymbol{A}\overline{\boldsymbol{Q}}_{k-1},\\\overline{\boldsymbol{Q}}_k\overline{\boldsymbol{R}}_k&=\boldsymbol{Q}_1\boldsymbol{Q}_2\cdots\boldsymbol{Q}_k\boldsymbol{R}_k\cdots\boldsymbol{R}_2\boldsymbol{R}_1=\overline{\boldsymbol{Q}}_{k-1}\boldsymbol{A}_k\overline{\boldsymbol{R}}_{k-1}\\&=\overline{\boldsymbol{Q}}_{k-1}\overline{\boldsymbol{Q}}_{k-1}{}^{\mathrm{T}}\boldsymbol{A}\overline{\boldsymbol{Q}}_{k-1}\overline{\boldsymbol{R}}_{k-1}=\boldsymbol{A}\overline{\boldsymbol{Q}}_{k-1}\overline{\boldsymbol{R}}_{k-1}.\end{aligned}$$

由此递推及$\overline{\boldsymbol{Q}}_1\overline{\boldsymbol{R}}_1=\boldsymbol{Q}_1\boldsymbol{R}_1=\boldsymbol{A}_1=\boldsymbol{A}$，即证得(2). 定理得证.

一般情形下，QR 算法的收敛性比较复杂. 若矩阵列$\{\boldsymbol{A}_k\}$的对角元均收敛，且严格下三角部分元素均收敛到零，则对求 $\boldsymbol{A}$ 的特征值而言已足够了. 此时，我们称$\{\boldsymbol{A}_k\}$基本收敛到上三角阵. 下面对最简单的情形给出收敛性定理.

定理 8.14　设矩阵$\boldsymbol{A}\in\mathbf{R}^{n\times n}$的特征值满足

$$|\lambda_1|>|\lambda_2|>\cdots>|\lambda_n|>0,$$

λ_i 对应特征向量$\boldsymbol{x}_i$，$i=1,2,\cdots,n$. 若矩阵$\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$的逆可分解为$\boldsymbol{X}^{-1}=\boldsymbol{L}\boldsymbol{U}$，其中$\boldsymbol{L}$为单位下三角阵，$\boldsymbol{U}$为上三角阵，则 QR 算法产生的序列$\{\boldsymbol{A}_k\}$基本收敛到上三角阵，其对角元极限为

$$\lim_{k\to\infty}a_{ii}^{(k)}=\lambda_i,\quad i=1,2,\cdots,n.$$

更一般地，在一定条件下，由 QR 算法生成的序列$\{\boldsymbol{A}_k\}$收敛为 Schur 分块上三角形，对角块按特征值的模从大到小排列，上述定理是它的特殊情形. 当收敛结果为 Schur 分块上三角形时，序列$\{\boldsymbol{A}_k\}$的对角块以上的元素以及 2 阶块的元素不一定收敛，但这不影响求全部特征值：

例 8.6　用 QR 方法求下列矩阵的全部特征值：

$$(1)\ \boldsymbol{A}=\begin{pmatrix}-3&-5&-1\\13&13&1\\-5&-5&1\end{pmatrix};\quad(2)\ \boldsymbol{A}=\begin{pmatrix}4&1&-3\\-2&1&1\\2&1&-1\end{pmatrix}.$$

解　先用镜面反射变换矩阵 $\boldsymbol{A}$ 为 Hessenberg 形矩阵$\boldsymbol{A}_1$，然后用平面旋转变换作 QR 分解进行迭代，生成序列$\{\boldsymbol{A}_k\}$. (1) 的计算结果为

$$\boldsymbol{A}_1=\begin{pmatrix}-3.0000&4.3077&-2.7282\\-13.9284&12.7938&-5.5361\\&0.4639&1.2062\end{pmatrix},$$

$$\boldsymbol{A}_2=\begin{pmatrix}10.1133 & 17.7711 & 0.8265\\ -1.5511 & -0.6362 & -0.9381\\ & 0.4656 & 1.5229\end{pmatrix},$$

$$\boldsymbol{A}_{16}=\begin{pmatrix}6.0001 & 16.9820 & -9.2165\\ 0.0000 & 3.0019 & -1.6317\\ & 0.0012 & 1.9980\end{pmatrix},$$

$$\boldsymbol{A}_{23}=\begin{pmatrix}6.0000 & 16.9712 & -9.2364\\ 0.0000 & 3.0001 & -1.6329\\ & 0.0000 & 1.9999\end{pmatrix}.$$

该矩阵 $\boldsymbol{A}$ 非对称，从计算结果来看，收敛于上三角阵.

(2) 的计算结果为

$$\boldsymbol{A}_1=\begin{pmatrix}4.0000 & -2.8284 & -1.4142\\ 2.8284 & -1.0000 & -1.0000\\ & -1.0000 & 1.0000\end{pmatrix},$$

$$\boldsymbol{A}_2=\begin{pmatrix}2.3333 & -1.9379 & -5.1121\\ 0.7454 & 1.2667 & 0.3266\\ & -0.4899 & 0.4000\end{pmatrix},$$

$$\boldsymbol{A}_{25}=\begin{pmatrix}2.0003 & -0.8171 & 3.6516\\ 0.0002 & -0.3336 & 3.7263\\ & -0.7456 & 2.3333\end{pmatrix},$$

$$\boldsymbol{A}_{26}=\begin{pmatrix}2.0002 & -2.9999 & -2.2374\\ 0.0001 & 2.9996 & 2.2366\\ & -2.2349 & -0.9998\end{pmatrix}.$$

从计算结果来看，迭代收敛于 Schur 分块上三角形，对角块分别是 1 阶和 2 阶子矩阵. 事实上，矩阵 $\boldsymbol{A}_{25}$ 和 $\boldsymbol{A}_{26}$ 的右下角的 2 阶子矩阵的特征值都是 $0.9999\pm1.0000\mathrm{i}$，迭代已接近收敛.

一般在实际使用 QR 方法之前，先用镜面反射变换将 $\boldsymbol{A}$ 化为 Hessenberg 形矩阵 $\boldsymbol{H}$，然后对 $\boldsymbol{H}$ 作 QR 迭代，这样可以大大节省运算工作量. 因为上 Hessenberg 阵 $\boldsymbol{H}$ 的次对角线以下元素均为零，所以用平面旋转变换作 QR 分解较为方便。

对 $i=1,2,\cdots,n-1$，依次用平面旋转矩阵 $\boldsymbol{J}(i,i+1)$ 左乘 $\boldsymbol{H}$，使 $\boldsymbol{J}(i,i+1)\boldsymbol{H}$ 的第 $i+1$ 行第 i 列元素为零. 左乘 $\boldsymbol{J}(i,i+1)$ 后，矩阵 $\boldsymbol{H}$ 的第 i 行与第 $i+1$ 行零元素位置上仍为零，其他行不变. 这样，共 $n-1$ 次左乘正交矩阵后得到上三角阵 $\boldsymbol{R}$. 即 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{R}$，$\boldsymbol{U}^{\mathrm{T}}=\boldsymbol{J}(n-1,n)\boldsymbol{J}(n-2,n-1)\cdots\boldsymbol{J}(1,2)$. 可以验证 $\boldsymbol{U}^{\mathrm{T}}$ 是一

个下 Hessenberg 阵，即 $\boldsymbol{U}$ 是一个上 Hessenberg 阵. 这样，得到 $\boldsymbol{H}$ 的 QR 分解 $\boldsymbol{H}=\boldsymbol{UR}$. 在作 QR 迭代时，下一步计算 $\boldsymbol{RU}$，容易验证 $\boldsymbol{RU}$ 是一个上 Hessenberg 阵. 以上说明了 QR 算法保持了 $\boldsymbol{H}$ 的上 Hessenberg 结构形式。

例 **8.7**　求 Hessenberg 形矩阵

$$\boldsymbol{H}=\begin{pmatrix}5 & -2 & -5 & -1\\ 1 & 0 & -3 & 2\\ 0 & 2 & 2 & -3\\ 0 & 0 & 1 & -2\end{pmatrix}$$

的特征值.

解　令 $\boldsymbol{H}_1=\boldsymbol{H}$，根据 $\boldsymbol{H}_1$ 的次对角线非零元素，有平面旋转矩阵 $\boldsymbol{J}(1,2)$，$\boldsymbol{J}(2,3)$，$\boldsymbol{J}(3,4)$，使得

$$\boldsymbol{H}_1=\boldsymbol{U}_1\boldsymbol{R}_1=\begin{pmatrix}0.980\,6 & -0.037\,7 & 0.192\,3 & -0.103\,8\\ 0.196\,1 & 0.188\,7 & -0.880\,4 & -0.419\,2\\ 0 & 0.981\,3 & 0.176\,1 & 0.074\,0\\ 0 & 0 & 0.396\,2 & -0.898\,9\end{pmatrix}\cdot\begin{pmatrix}5.099\,2 & -1.961\,2 & -5.491\,2 & -0.392\,2\\ 0 & 2.038\,1 & 1.585\,2 & -2.528\,8\\ 0 & 0 & 2.524\,2 & -3.273\,6\\ 0 & 0 & 0 & 0.782\,2\end{pmatrix},$$

其中 $\boldsymbol{U}_1{}^{\mathrm{T}}=\boldsymbol{J}(3,4)\boldsymbol{J}(2,3)\boldsymbol{J}(1,2)$. 然后将求得的 $\boldsymbol{U}_1$ 和 $\boldsymbol{R}_1$ 逆序相乘，求出 $\boldsymbol{H}_2$，

$$\boldsymbol{H}_2=\boldsymbol{R}_1\boldsymbol{U}_1=\begin{pmatrix}4.615\,7 & 5.950\,8 & 1.592\,2 & 0.239\,0\\ 0.399\,7 & 1.940\,1 & -2.517\,1 & 1.536\,1\\ 0 & 2.477\,0 & -0.852\,5 & 3.129\,4\\ 0 & 0 & 0.309\,9 & -0.703\,1\end{pmatrix}.$$

重复上面的过程，计算 10 次得

$$\boldsymbol{H}_{12}=\begin{pmatrix}4.000\,0 & * & * & *\\ & 1.878\,9 & -3.591\,0 & *\\ & 1.329\,0 & 0.121\,1 & *\\ & & & -1.000\,0\end{pmatrix}.$$

至此，不难看出，一个特征值是 4，另一个特征值是 -1，其他两个特征值是方程

$$\begin{vmatrix}1.878\,9-\lambda & -3.591\,0\\ 1.329\,0 & 0.121\,1-\lambda\end{vmatrix}=0$$

的根，求得为 $1\pm 2\mathrm{i}$. 事实上，可以求得矩阵 $\boldsymbol{H}$ 的特征方程为

$$\lambda^4-5\lambda^3+7\lambda^2-7\lambda-20=0,$$

上述用 QR 方法求得的特征值是该特征方程的准确解.

8.4.3 带原点位移的 QR 算法

前面我们介绍了在反幂法中应用原点位移的策略,这种思想方法也可用于 QR 算法. 一般我们针对上 Hessenberg 矩阵讨论 QR 算法,并且假设每次 QR 迭代中产生的 $\boldsymbol{A}_k$ 都是不可约的,否则,可以将问题分解为较小型的问题. 这样,带原点位移的 QR 算法可以描述为

$$\begin{cases} \text{取 } \boldsymbol{A}_1 \text{ 为 } \boldsymbol{A} \text{ 的 Hessenberg 形(初始化)}, \\ \boldsymbol{A}_k - s_k\boldsymbol{I} = \boldsymbol{Q}_k\boldsymbol{R}_k \quad \text{(QR 分解)}, \\ \boldsymbol{A}_{k+1} = \boldsymbol{R}_k\boldsymbol{Q}_k + s_k\boldsymbol{I} \quad \text{(正交相似变换)}, k = 1,2,\cdots. \end{cases}$$

这里,$\boldsymbol{A}_k$ 到 $\boldsymbol{A}_{k+1}$ 的变换称为原点位移的 **QR** 变换.

由于 $\boldsymbol{R}_k\boldsymbol{Q}_k + s_k\boldsymbol{I} = \boldsymbol{Q}_k^{\mathrm{T}}\boldsymbol{Q}_k\boldsymbol{R}_k\boldsymbol{Q}_k + s_k\boldsymbol{Q}_k^{\mathrm{T}}\boldsymbol{Q}_k = \boldsymbol{Q}_k^{\mathrm{T}}(\boldsymbol{Q}_k\boldsymbol{R}_k + s_k\boldsymbol{I})\boldsymbol{Q}_k$,所以,$\boldsymbol{A}_{k+1} = \boldsymbol{Q}_k^{\mathrm{T}}\boldsymbol{A}_k\boldsymbol{Q}_k$. 即每个 $\boldsymbol{A}_k$ 都与 $\boldsymbol{A}_1$ 相似,从而与原矩阵 $\boldsymbol{A}$ 相似. 实际计算时,用不同的位移 $s_1, s_2, \cdots, s_k \cdots$,反复应用上述变换就产生一正交相似于 Hessenberg 阵的序列$\{\boldsymbol{A}_k\}$. 设 $\boldsymbol{A}_k = (a_{ij}^{(k)})_{n\times n}$,$\boldsymbol{B}_k$ 是 $\boldsymbol{A}_k$ 的 $n-1$ 阶顺序主子矩阵,若选取 $s_k = a_{nn}^{(k)}$,则在一定的条件下,$\boldsymbol{A}_k$ 基本收敛于三角矩阵,并且 $a_{nn}^{(k)}$ 作为 $\boldsymbol{A}$ 的近似特征值. 采用收缩方法,即对 $\boldsymbol{B}_k \in \mathbf{R}^{(n-1)\times(n-1)}$ 应用 QR 算法,就可逐步求出 $\boldsymbol{A}$ 的其余近似特征值.

判别 $a_{n,n-1}^{(k)}$ 充分小的准则可以是

$$|a_{n,n-1}^{(k)}| \leqslant \varepsilon \|\boldsymbol{A}_1\|_\infty$$

或者将 $a_{n,n-1}^{(k)}$ 与相邻元素进行比较,取准则

$$|a_{n,n-1}^{(k)}| \leqslant \varepsilon(|a_{n-1,n-1}^{(k)}| + |a_{nn}^{(k)}|),$$

其中 ε 为给定的精度要求. 满足上述准则时,可认为 $a_{n,n-1}^{(k)} = 0$,$a_{nn}^{(k)}$ 就作为 $\boldsymbol{A}$ 的一个近似特征值.

根据 QR 算法的收敛性质,位移量有下列两种算法:

(1) $s_k = a_{nn}^{(k)}$;

(2) s_k 取为矩阵

$$\begin{pmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)} & a_{nn}^{(k)} \end{pmatrix}$$

的特征值中与 $a_{nn}^{(k)}$ 最接近的一个.

具体计算时,用平面旋转变换对 Hessenberg 矩阵 $\boldsymbol{A}_1$ 进行原点位移的 QR 变换可表达为

$$\boldsymbol{P}_{n-1,n}\cdots\boldsymbol{P}_{23}\boldsymbol{P}_{12}(\boldsymbol{A}_1 - s_1\boldsymbol{I}) = \boldsymbol{R}_1,$$
$$\boldsymbol{A}_2 = \boldsymbol{R}_1\boldsymbol{P}_{12}^{\mathrm{T}}\boldsymbol{P}_{23}^{\mathrm{T}}\cdots\boldsymbol{P}_{n-1,n}^{\mathrm{T}} + s_1\boldsymbol{I}.$$

容易验证，$\boldsymbol{A}_2$ 仍为上 Hessenberg 矩阵.

例 8.8　用带原点位移的 QR 算法求下列矩阵的特征值

$$\boldsymbol{A}=\begin{pmatrix}-1 & 2 & 1\\ 2 & -4 & 1\\ 1 & 1 & -6\end{pmatrix}.$$

解　先用镜面反射变换把 $\boldsymbol{A}$ 化为上 Hessenberg 矩阵. 按(8.12) 式有

$$\boldsymbol{H}=\boldsymbol{I}-2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}=\begin{pmatrix}1 & 0 & 0\\ 0 & -\dfrac{2}{\sqrt{5}} & -\dfrac{1}{\sqrt{5}}\\ 0 & -\dfrac{1}{\sqrt{5}} & \dfrac{2}{\sqrt{5}}\end{pmatrix},$$

$$\boldsymbol{A}_1=\boldsymbol{H}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{H}=\begin{pmatrix}-1 & -\sqrt{5} & 0\\ -\sqrt{5} & -3.6 & 0.2\\ 0 & 0.2 & -6.4\end{pmatrix}.$$

若按第一种方法取位移量，即取位移量为右下角元素，则有 $s_1=-6.4$，$\theta_1=-0.392\,590\,761$，$\theta_2=0.114\,997\,409$，

$$\boldsymbol{P}_1=\begin{pmatrix}\cos\theta_1 & \sin\theta_1 & 0\\ -\sin\theta_1 & \cos\theta_1 & 0\\ 0 & 0 & 1\end{pmatrix},\quad \boldsymbol{P}_2=\begin{pmatrix}1 & 0 & 0\\ 0 & \cos\theta_2 & \sin\theta_2\\ 0 & -\sin\theta_2 & \cos\theta_2\end{pmatrix},$$

$$\begin{aligned}\boldsymbol{R}_1&=\boldsymbol{P}_2\boldsymbol{P}_1(\boldsymbol{A}_1+6.4\boldsymbol{I})\\ &=\begin{pmatrix}5.844\,655\,679 & -3.137\,183\,510 & -0.076\,516\,671\\ & 1.743\,008\,79 & 0.183\,563\,711\\ & & -0.021\,202\,899\end{pmatrix},\\ \boldsymbol{A}_2&=\boldsymbol{R}_1\boldsymbol{P}_1^{\mathrm{T}}\boldsymbol{P}_1^{\mathrm{T}}-6.4\boldsymbol{I}\\ &=\begin{pmatrix}0.200\,234\,192 & -0.666\,846\,149 & 0\\ -0.666\,846\,149 & -4.779\,171\,336 & -0.002\,432\,908\\ 0 & -0.002\,432\,908 & -6.421\,062\,856\end{pmatrix}.\end{aligned}$$

同理可得

$$\boldsymbol{A}_3=\begin{pmatrix}0.283\,205\,88 & -0.157\,002\,612 & 0\\ -0.157\,002\,612 & -4.862\,139\,274 & 0.000\,000\,006\\ 0 & 0.000\,000\,006 & -6.421\,066\,615\end{pmatrix},$$

$$\boldsymbol{A}_4=\begin{pmatrix}0.287\,735\,078 & -0.036\,401\,350 & 0\\ -0.036\,401\,350 & -4.866\,668\,465 & 0\\ 0 & 0 & -6.421\,066\,615\end{pmatrix}.$$

故有特征值 $\lambda_3 = -6.421\,066\,615$. 求左上角 2×2 矩阵的特征值，得 $\lambda_1 = 0.287\,992\,138$，$\lambda_2 = -4.866\,925\,525$.

若按第二种方法取位移量，即取右上角 2×2 矩阵的特征值，则有 $s_1 = -6.469\,693\,846$，类似于上面的计算可得

$$\boldsymbol{A}_2 = \begin{pmatrix} 0.194\,154\,158 & -0.689\,146\,437 & 0 \\ -6.891\,464\,37 & -4.773\,105\,873 & 0.005\,374\,767 \\ 0 & 0.005\,374\,767 & -6.421\,048\,287 \end{pmatrix},$$

$$\boldsymbol{A}_3 = \begin{pmatrix} 0.282\,852\,106 & -0.162\,696\,110 & 0 \\ -0.162\,696\,110 & -4.861\,785\,493 & 0.000\,011\,074 \\ 0 & 0.000\,011\,074 & -6.421\,066\,615 \end{pmatrix},$$

$$\boldsymbol{A}_4 = \begin{pmatrix} 0.287\,716\,058 & -0.037\,723\,983 & 0 \\ -0.037\,723\,983 & -4.866\,649\,445 & 0 \\ 0 & 0 & -6.421\,066\,615 \end{pmatrix}.$$

由此可得特征值 $\lambda_1 = 0.287\,992\,139$，$\lambda_2 = -4.866\,925\,526$，$\lambda_3 = -6.421\,066\,615$.

该问题如果不用带原点位移的 QR 算法，而是用基本 QR 算法，则收敛速度很慢，计算结果为

$$\boldsymbol{A}_2 = \begin{pmatrix} -4.838\,383\,838 & 0.552\,770\,798 & 0 \\ 0.552\,770\,798 & -0.439\,393\,939 & -2.004\,127\,972 \\ 0 & -2.004\,127\,972 & -5.727\,272\,727 \end{pmatrix},$$

$$\boldsymbol{A}_7 = \begin{pmatrix} -5.282\,161\,439 & 0.687\,687\,671 & 0 \\ 0.687\,687\,671 & -6.005\,830\,703 & -0.000\,000\,190 \\ 0 & -0.000\,000\,190 & 0.287\,992\,139 \end{pmatrix},$$

$$\boldsymbol{A}_{30} = \begin{pmatrix} -6.421\,054\,217 & 0.004\,369\,012\,0 & 0 \\ 0.004\,390\,120 & -4.866\,937\,928 & 0 \\ 0 & 0 & 0.287\,992\,139 \end{pmatrix}.$$

不过，在上述计算结果中，$\boldsymbol{A}_7$ 的第 3 个对角元已稳定，可以认为 $\lambda_3 \approx 0.287\,992\,139$，$\lambda_1$ 和 λ_2 可认为是 $\boldsymbol{A}_7$ 的左上角 2×2 矩阵的特征值，可解得 $\lambda_1 \approx -6.421\,066\,617$，$\lambda_2 \approx 4.866\,925\,525$.

评　　注

本章介绍了矩阵特征问题的幂法、Jacobi 方法和 QR 算法，它们是求矩阵特征值和特征向量的常用数值方法. 本章用到较多的线性代数知识和方法，其中一

些是一般线性代数教科书上没有提到的. 圆盘定理给出了特征值的大致估计. 平面旋转变换和镜面反射变换是两种有力的正交相似变换工具,可以化简矩阵和作 QR 分解等,用于构造和分析数值方法.

幂法用于求矩阵的主特征值和主特征向量,特别适用于大型稀疏矩阵. 它计算简单,但收敛速度往往不能令人满意. 可以用反幂法结合位移技巧加速收敛,或求某一指定的特征值. 幂法以及它的变形显然适用于对称矩阵,因为这种矩阵的特征值都是实数并且可对角化,结合矩阵的收缩方法可以求出它的全部特征值和特征向量.

Jacobi 方法是古典的方法,用于求对称矩阵的全部特征值和特征向量. 一般情况下,它和对称的 QR 方法相比,已经没有多大优越性. 但是在求几乎接近对角形的矩阵的特征值时,它还是有效的方法.

QR 方法是求全部特征值的方法,它是 20 世纪 60 年代发展起来的. 1976 年 G. Strang 在他的著作 *Linear Algebra and its Applications* 中,称 QR 方法是"数值数学最值得注意的算法之一",从理论分析到实际应用都使这种观点得到广泛的认同。QR 方法具有收敛快、精度高的特点. 特别是对称的 QR 方法,可以写成很简洁的算法. 在中小型稠密矩阵的特征值问题计算中,目前它仍然是最有效的方法之一.

习　题　8

8.1　用幂法求下列矩阵的主特征值和主特征向量

$$\boldsymbol{A}=\begin{pmatrix}3 & -2 & -4\\ -2 & 6 & -2\\ -4 & -2 & 3\end{pmatrix}.$$

当特征值有 3 位小数稳定时迭代终止,再对计算结果用 Aitken 外推加速.

8.2　用反幂法求下列矩阵模最小的特征值和对应的特征向量

$$\boldsymbol{A}=\begin{pmatrix}3 & -4 & 3\\ -4 & 6 & 3\\ 3 & 3 & 1\end{pmatrix}.$$

8.3　用反幂法求矩阵

$$\boldsymbol{A}=\begin{pmatrix}6 & 2 & 1\\ 2 & 3 & 1\\ 1 & 1 & 1\end{pmatrix}$$

的最接近于 6 的特征值和特征向量.

8.4 用 Jacobi 方法计算下列矩阵的全部特征值和特征向量

$$\boldsymbol{A}_1=\begin{pmatrix}4&0&0\\0&3&1\\0&1&3\end{pmatrix},\quad \boldsymbol{A}_2=\begin{pmatrix}1.0&1.0&0.5\\1.0&1.0&0.25\\0.5&0.25&2.0\end{pmatrix}.$$

8.5 设 $\boldsymbol{x}=(1,1,1,1)^{\mathrm{T}}$，用下列两种方法分别求正交矩阵 $\boldsymbol{P}$，使得 $\boldsymbol{Px}=\pm\|\boldsymbol{x}\|_2\boldsymbol{e}_1$.

(1) $\boldsymbol{P}$ 为平面旋转矩阵的乘积；

(2) $\boldsymbol{P}$ 为镜面反射矩阵.

8.6 (1) 设矩阵 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 对称，λ 和 $\boldsymbol{x}(\|\boldsymbol{x}\|_2=1)$ 是 $\boldsymbol{A}$ 的一个特征值及对应的特征向量. 试证：若有正交矩阵 $\boldsymbol{P}$ 使得 $\boldsymbol{Px}=\boldsymbol{e}_1$，则有

$$\boldsymbol{PAP}^{\mathrm{T}}=\begin{pmatrix}\lambda&0\\0&\boldsymbol{B}\end{pmatrix};$$

(2) 已知矩阵

$$\boldsymbol{A}=\begin{pmatrix}2&10&2\\10&5&-8\\2&-8&11\end{pmatrix}$$

的一个特征值 $\lambda=9$ 和对应的特征向量 $\boldsymbol{x}=\left(\frac{2}{3},\frac{1}{3},\frac{2}{3}\right)^{\mathrm{T}}$. 试求镜面反射矩阵 $\boldsymbol{P}$，使得 $\boldsymbol{Px}=\boldsymbol{e}_1$，并计算 $\boldsymbol{PAP}^{\mathrm{T}}$.

8.7 用正交相似变换将下列矩阵化为对称三对角阵

$$\boldsymbol{A}=\begin{pmatrix}1&3&4\\3&1&2\\4&2&1\end{pmatrix}.$$

8.8 用镜面反射变换求下列矩阵的 QR 分解

$$\boldsymbol{A}=\begin{pmatrix}1&1&1\\2&-1&-1\\2&-4&5\end{pmatrix}.$$

8.9 用平面旋转变换对下列 Hessenberg 形矩阵作一步 QR 变换

(1) $\boldsymbol{A}=\begin{pmatrix}0&2&-2\\-1&2&-2\\&-1&1\end{pmatrix}$；　(2) $\boldsymbol{A}=\begin{pmatrix}3&1&\\1&4&2\\&2&1\end{pmatrix}$.

8.10 设矩阵 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 为 Hessenberg 形. 对 QR 变换

$$\boldsymbol{A}=\boldsymbol{QR},\quad \boldsymbol{B}=\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{AQ}=\boldsymbol{RQ},$$

证明矩阵 $\boldsymbol{Q}$ 和 $\boldsymbol{B}$ 都是 Hessenberg 形矩阵.

数值试验题 8

8.1　对于矩阵

$$\boldsymbol{A}=\begin{pmatrix}0&1&2&3\\2&3&0&1\\3&0&1&2\\1&2&3&0\end{pmatrix},$$

(1) 用幂法计算 $\boldsymbol{A}$ 的主特征值和对应原特征向量. 当特征值有 6 位小数稳定时迭代终止;

(2) 以幂法迭代几次所得主特征值的近似值为位移量 p,用反幂法求接近于 p 的特征值及对应的特征向量.

8.2　对于适当阶数(例如 10～100 阶)的矩阵

$$\boldsymbol{A}=\begin{pmatrix}4&1&&&\\1&4&1&&\\&\ddots&\ddots&\ddots&\\&&1&4&1\\&&&1&4\end{pmatrix},$$

用 Jacobi 方法求它的全部特征值和特征向量.

8.3　求多项式方程 $f(x)=x^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0$ 根的问题,可以化为求矩阵

$$\boldsymbol{A}=\begin{pmatrix}-a_{n-1}&a_{n-2}&\cdots&-a_1&-a_0\\1&0&\cdots&0&0\\0&1&\cdots&0&0\\\vdots&\vdots&&\vdots&\vdots\\0&0&\cdots&1&0\end{pmatrix}.$$

的特征值问题. 给定高次方程:

(1) $x^3+x^2-5x+3=0$;

(2) $x^3-3x-1=0$;

(3) $x^{41}+x^3+1=0$.

试用幂法求出方程的模最大的根,或用 QR 算法求出高次方程的一切根.

第 9 章　常微分方法的数值解法

科学技术与工程问题常常需要建立微分方程形式的数学模型，下面是这类问题的例子.

设 $N(t)$ 为某物种的数量，α 为该物种的出生率与死亡率之差，β 为生物的食物供给及它们所占空间的限制，描述该物种增长率的数学模型是

$$\begin{cases} \dfrac{\mathrm{d}N}{\mathrm{d}t} = \alpha N(t) - \beta N^2(t), \\ N(t_0) = N_0. \end{cases}$$

设 Q 是电容器上的带电量，C 为电容，R 为电阻，E 为电源的电动势，描述该电容器充电过程的数学模型是

$$\begin{cases} \dfrac{\mathrm{d}Q}{\mathrm{d}t} = E - \dfrac{Q(t)}{RC}, \\ Q(t_0) = Q_0. \end{cases}$$

以上两个例子是常微分方程的初值问题，下面是一个两点边值问题的例子.

设一根长为 L 的矩形截面的梁，两端固定. E 是弹性模量，S 是端点作用力，$I(x)$ 是惯性矩，q 是均匀载荷强度，梁的桡度 $y(x)$ 满足如下方程

$$\begin{cases} \dfrac{\mathrm{d}^2 y}{\mathrm{d}x^2} = \dfrac{S}{EI(x)} y(x) + \dfrac{qx}{2EI(x)}(x - L), \\ y(0) = y(L) = 0. \end{cases}$$

针对实际问题建立的数学模型，要找出模型解的解析表达式往往是困难的，甚至是不可能的. 因此，需要研究和掌握微分方程的数值解法，即计算域内离散点上解的近似值的方法. 本章讨论常微分方程数值解的基本方法和理论.

9.1　Euler　方　法

9.1.1　Euler 方法及其有关的方法

考虑一阶常微分方程的初值问题

$$\begin{cases} y' = f(x, y), \\ y(x_0) = y_0. \end{cases} \tag{9.1}$$

设 $f(x, y)$ 是连续函数，对 y 满足 Lipschitz 条件，即存在正数 L，使得对于任意两

点(x,y)与$(x,\bar{y})$,有$|f(x,y)-f(x,\bar{y})|\leqslant L|y-\bar{y}|$,这样初值问题的解是存在唯一的,而且连续依赖于初始条件.

为了求得离散点上的函数值,将微分方程的连续问题(9.1)式进行离散化.一般是引入点列$\{x_n\}$,这里$x_n=x_{n-1}+h_n,n=1,2,\cdots$.称$h_n$为步长,经常考虑定长的情形,即$h_n=h,x_n=x_0+nh,n=0,1,\cdots$.

记$y(x_n)$为初始问题(9.1)的准确解$y(x)$在x_n处的值,用均差近似代替(9.1)式中的导数得

$$\frac{y(x_n+h)-y(x_n)}{h}\approx f(x_n,y(x_n)),$$

$$\frac{y(x_n+h)-y(x_n)}{h}\approx f(x_{n+1},y(x_{n+1})).$$

令y_n为$y(x_n)$的近似值,将上面两个近似式写成等式,整理后得

$$y_{n+1}=y_n+hf(x_n,y_n),\quad n=0,1,\cdots, \tag{9.2}$$

$$y_{n+1}=y_n+hf(x_{n+1},y_{n+1}),\quad n=0,1,\cdots. \tag{9.3}$$

从x_0处的初值y_0开始,按(9.2)式可逐步计算以后各点上的近似值.称(9.2)式为显式 Euler 公式.由于(9.3)式的右端隐含有待求函数值y_{n+1},不能逐步显式计算,称(9.3)式为隐式 Euler 公式或后退 Euler 公式.

如果将(9.2)式和(9.3)式作算术平均,就得梯形公式

$$y_{n+1}=y_n+\frac{h}{2}[f(x_n,y_n)+f(x_{n+1},y_{n+1})],\quad n=0,1,\cdots. \tag{9.4}$$

梯形公式也是隐式公式.

以上公式都是由y_n去计算y_{n+1},故称它们为单步法.

例 9.1　取$h=0.1$,用 Euler 方法、隐式 Euler 方法和梯形方法解

$$\begin{cases}y'=x-y+1,\\ y(0)=1.\end{cases}$$

解　本题有$f(x,y)=x-y+1,y_0=1$.如果用 Euler 方法,由(9.2)式并代入$h=0.1$得

$$y_{n+1}=0.1x_n+0.9y_n+0.1.$$

同理,用隐式 Euler 方法有

$$y_{n+1}=\frac{1}{1.1}(0.1x_{n+1}+y_n+0.1).$$

用梯形公式有

$$y_{n+1}=\frac{1}{1.05}(0.1x_n+0.95y_n+0.105).$$

3 种方法及准确解$y(x)=x+\mathrm{e}^{-x}$的数值结果如表 9-1 所示.从表中可看到,在

$x_n = 0.5$ 处，Euler 方法和隐式 Euler 方法的误差 $|y(x_n) - y_n|$ 分别是 1.6×10^{-2} 和 1.4×10^{-2}，而梯形方法的误差却是 2.5×10^{-4}.

表 9-1

x_n	Euler 方法	隐式 Euler 方法	梯形法	准确解
0	1	1	1	1
0.1	1.000 000	1.009 091	1.004 762	1.004 837
0.2	1.010 000	1.026 446	1.018 594	1.018 731
0.3	1.029 000	1.051 315	1.040 633	1.040 818
0.4	1.056 100	1.083 013	1.070 096	1.070 320
0.5	1.090 490	1.120 921	1.106 278	1.106 531

在例 9.1 中，由于 $f(x,y)$ 对 y 是线性的，所以对隐式公式也可方便地计算 y_{n+1}. 但是，当 $f(x,y)$ 是 y 的非线性函数时，如 $y' = 5x + \sqrt[3]{y}$，其隐式 Euler 公式为 $y_{n+1} = y_n + h(5x_{n+1} + \sqrt[3]{y_{n+1}})$. 显然，它是 y_{n+1} 的非线性方程，可以选择非线性方程求根的迭代法求解 y_{n+1}. 以梯形公式为例，可用显式 Euler 公式提供迭代初值 $y_{n+1}^{(0)}$，用迭代公式

$$\begin{cases} y_{n+1}^{(0)} = y_n + hf(x_n, y_n), \\ y_{n+1}^{(k+1)} = y_n + \dfrac{h}{2}(f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(k)})), \quad k = 0, 1\cdots, \end{cases}$$

反复迭代，直到

$$|y_{n+1}^{(k+1)} - y_{n+1}^{(k)}| < \varepsilon,$$

其中，步长 h 成为迭代参数，它需要满足一定的条件迭代公式才能收敛. 若将 (9.4) 式减去该迭代公式，得

$$y_{n+1} - y_{n+1}^{(k+1)} = \frac{h}{2}(f(x_{n+1}, y_{n+1}) - f(x_{n+1}, y_{n+1}^{(k)})).$$

假设 $f(x,y)$ 关于 y 满足 Lipschitz 条件，则有

$$|y_{n+1} - y_{n+1}^{(k+1)}| \leqslant \frac{hL}{2}|y_{n+1} - y_{n+1}^{(k)}|,$$

这里，L 是 Lipschitz 常数. 由上式可见，当 $\dfrac{hL}{2} < 1$ 即 $h < \dfrac{2}{L}$ 时，迭代序列 $\{y_{n+1}^{(k)}\}$ 收敛于 y_{n+1}.

对于隐式公式，通常采用预估－校正技术，即先用显式公式计算，得到预估值，然后以预估值为隐式公式的迭代初值，用隐式公式迭代一次得到校正值，称为预估—校正技术. 例如，用显式 Euler 公式作预估，用梯形公式作校正，即

$$\begin{cases}\bar{y}_{n+1} = y_n + hf(x_n, y_n), \\ y_{n+1} = y_n + \dfrac{h}{2}(f(x_n, y_n) + f(x_{n+1}, \bar{y}_{n+1})), n = 0, 1, \cdots. \end{cases} \tag{9.5}$$

称该公式为改进的 Euler 公式. 它等价的显式公式为

$$y_{n+1} = y_n + \frac{h}{2}(f(x_n, y_n) + f(x_{n+1}, y_n + hf(x_n, y_n))), \tag{9.6}$$

也可表示为下列平均化的形式

$$\begin{cases} y_p = y_n + hf(x_n, y_n), \\ y_q = y_n + hf(x_{n+1}, y_p), \\ y_{n+1} = \dfrac{1}{2}(y_p + y_q). \end{cases}$$

例 9.2　取 $h = 0.1$,用改进的 Euler 方法解

$$\begin{cases} y' = y - \dfrac{2x}{y}, \\ y(0) = 1. \end{cases}$$

解　按(9.5) 式,改进的 Euler 公式为

$$\begin{cases} \bar{y}_{n+1} = y_n + h\left(y_n - \dfrac{2x_n}{y_n}\right), \\ y_{n+1} = y_n + \dfrac{h}{2}\left(\left(y_n - \dfrac{2x_n}{y_n}\right) + \left(\bar{y}_{n+1} - \dfrac{2x_{n+1}}{\bar{y}_{n+1}}\right)\right), \quad n = 0, 1, \cdots. \end{cases}$$

由于 $y_0 = 1, h = 0.1$ 得计算结果如表 9-2. 该初值问题的准确解为 $y(x) = \sqrt{1 + 2x}$.

表 9-2

x_n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
y_n	1.095 9	1.184 1	1.266 2	1.343 4	1.416 4	1.486 0	1.552 5	1.615 3
$y(x_n)$	1.095 4	1.183 2	1.264 9	1.341 6	1.414 2	1.483 2	1.549 2	1.616 5

9.1.2　局部误差和方法的阶

初值问题(9.1) 式的单步法可以写成如下统一形式

$$y_{n+1} = y_n + h\varphi(x_n, x_{n+1}, y_n, y_{n+1}, h), \tag{9.7}$$

其中 φ 与 f 有关. 若 φ 中不含 y_{n+1},则方法是显式的,否则是隐式的,所以一般显式单步法表示为

$$y_{n+1} = y_n + h\varphi(x_n, y_n, h). \tag{9.8}$$

例如,Euler 方法中,有 $\varphi(x, y, h) = f(x, y)$.

对于不同的方法，计算值 y_n 与准确解 $y(x_n)$ 的误差各不相同. 所以有必要讨论方法的截断误差。我们称 $e_n = y(x_n) - y_n$ 为某一方法在 x_n 点的整体截断误差. 显然，e_n 不单与 x_n 这步的计算有关，它与以前各步的计算都有关，所以误差称为整体的. 分析和估计整体截断误差 e_n 是复杂的. 为此，我们假设 x_n 处的 y_n 没有误差，即 $y_n = y(x_n)$，考虑从 x_n 到 x_{n+1} 这一步的误差，这就是如下的局部误差的概念.

定义 9.1 设 $y(x)$ 是初值问题(9.1)式的准确解，则称

$$T_{n+1} = y(x_{n+1}) - y(x_n) - h\varphi(x_n, x_{n+1}, y(x_n), y(x_{n+1}), h)$$

为单步法(9.7)式的局部截断误差.

定义 9.2 如果给定方法的局部截断误差 $T_{n+1} = O(h^{p+1})$，其中 $p \geqslant 1$ 为整数，则称该方法是 p 阶的，或者具有 p 阶精度. 若一个 p 阶单步法的局部截断误差为

$$T_{n+1} = g(x_n, y(x_n))h^{p+1} + O(h^{p+2}),$$

则称其第一个非零项 $g(x_n, y(x_n))h^{p+1}$ 为该方法的局部截断误差的主项.

对于 Euler 方法，由 Taylor 展开式有

$$\begin{aligned} T_{n+1} &= y(x_{n+1}) - y(x_n) - hf(x_n, y(x_n)) \\ &= y(x_{n+1}) - y(x_n) - hy'(x_n) \\ &= \frac{h^2}{2}y''(x_n) + \frac{h^3}{6}y'''(x_n) + O(h^4) = O(h^2). \end{aligned}$$

所以，Euler 方法是一种一阶方法，其局部截断误差的主项为 $\frac{h^2}{2}y''(x_n)$.

对于隐式 Euler 方法，其局部截断误差为

$$\begin{aligned} T_{n+1} &= y(x_{n+1}) - y(x_n) - hf(x_{n+1}, y(x_{n+1})) \\ &= y(x_{n+1}) - y(x_n) - hy'(x_{n+1}) \\ &= -\frac{h^2}{2}y''(x_n) + O(h^3) = O(h^2). \end{aligned}$$

所以，隐式 Euler 方法也是一种一阶方法，该方法的局部截断误差的主项为 $-\frac{h^2}{2}y''(x_n)$，仅与显式 Euler 方法的局部截断误差的主项相差一个符号.

梯形方法也是一种隐式单步法，类似可得其局部截断误差

$$\begin{aligned} T_{n+1} &= y(x_{n+1}) - y(x_n) - \frac{h}{2}(f(x_n, y(x_n)) + f(x_{n+1}, y(x_{n+1}))) \\ &= -\frac{h^3}{12}y'''(x_n) + O(h^4) = O(h^3). \end{aligned}$$

可见，梯形方法是 2 阶精度的.

例 9.3　证明

$$y_{n+1} = y_n + \frac{h}{6}(4f(x_n, y_n) + 2f(x_{n+1}, y_{n+1}) + hf'(x_n, y_n))$$

定义的隐式单步法公式是 3 阶的.

证　设 $y_n = y(x_n)$,则由

$$f(x_n, y_n) = y'(x_n), \quad f'(x_n, y_n) = y''(x_n),$$

$$\begin{aligned} f(x_{x+1}, y_{n+1}) &= y'(x_{n+1}) \\ &= y'(x_n) + hy''(x_n) + \frac{h^2}{2}y'''(x_n) + \frac{h^3}{6}y^{(4)}(x_n) + O(h^4), \end{aligned}$$

可得

$$\begin{aligned} y_{n+1} &= y(x_n) + \frac{h}{6}(4f(x_n, y_n) + 2f(x_{n+1}, y_{n+1}) + hf'(x_n, y_n)) \\ &= y(x_n) + hy'(x_n) + \frac{h^2}{2}y''(x_n) + \frac{h^3}{3!}y'''(x_n) + \frac{h^4}{18}y^{(4)}(x_n) + O(h^5). \end{aligned}$$

将此式与 $y(x_{n+1})$ 在 x_n 处的 Taylor 展开式比较得

$$y(x_{n+1}) - y_{n+1} = -\frac{1}{72}h^4 y^{(4)}(x_n) + O(h^5) = O(h^4).$$

由此可见,所给公式是 3 阶的.

9.2　Runge-Kutta 方法

9.2.1　Runge-Kutta 方法的基本思想

显式 Euler 方法是最简单的单步法,它是一阶的,它可以看作 Taylor 展开后取前两项. 因此,得到高阶方法的一个直接想法是用 Taylor 展开,如果能计算 $y(x)$ 的高阶导数,则可写出 p 阶方法的计算公式

$$y_{n+1} = y_n + hy'_n + \frac{h^2}{2}y''_n + \cdots + \frac{h^p}{p!}y_n^{(p)},$$

其中 $y_n^{(j)}$ 是 $y^{(j)}(x_n)$ 的近似值,$j = 0, 1, 2, \cdots, p$. 若将 $f(x, y)$,$\dfrac{\partial f}{\partial x}$,$\dfrac{\partial f}{\partial y}$,$\cdots$,分别记成 f,f_x,f_y,$\cdots$,则对于 2 阶和 3 阶导数可表示为

$$y'' = f_x + f_y f,$$

$$y''' = f_{xx} + 2f_{xy}f + f_x f_y + f_{yy}f^2 + {f_y}^2 f.$$

这个方法并不实用,因为一般情况下,求 $f(x, y)$ 的导数相当麻烦. 从计算高阶导数的公式知道,方法的截断误差提高一阶,需要增加的计算量很大.

例 9.4　取步长 $h = 0.25$,用 2 阶 Taylor 展开法求初值问题

$$\begin{cases} y' = x^2 + y^2, \\ y(1) = 1 \end{cases}$$

的解在 $x = 1.5$ 时的近似值.

解　2 阶 Taylor 展开公式为

$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{h^2}{2}y''(x_n) + O(h^3).$$

用 $y' = x^2 + y^2$, $y'' = 2x + 2yy' = 2x + 2y(x^2 + y^2)$ 代入上式并略去高阶项 $O(h^3)$, 则得求解公式

$$y_{n+1} = y_n + h({x_n}^2 + {y_n}^2) + \frac{h^2}{2}(2x_n + 2y_n({x_n}^2 + {y_n}^2)).$$

由 $y(1) = y_0 = 1$, 计算得

$$y(1.25) \approx y_1 = 1.687\,5,$$

$$y(1.50) \approx y_2 = 3.333\,298.$$

尽管如此, Taylor 展开法启发我们用区间上若干个点的导数 f, 而不是高阶导数, 将它们作线性组合得到平均斜率, 将其与解的 Taylor 展开式相比较, 使前面若干项吻合, 从而得到具有一定阶的方法. 这就是 Runge-Kutta 方法的基本思想, 其一般形式为

$$y_{n+1} = y_n + h\sum_{i=1}^{L}\lambda_i K_i, \tag{9.9}$$

其中

$$K_1 = f(x_n, y_n),$$

$$K_i = f(x_n + c_i h, y_n + c_i h\sum_{j=1}^{i-1}a_{ij}K_j), \quad i = 2,3,\cdots,L,$$

$c_i \leqslant 1$, $\sum_{i=1}^{L}\lambda_i = 1$, $\sum_{j=1}^{i-1}a_{ij} = 1$. 它的局部截断误差是

$$T_{n+1} = y(x_{n+1}) - y(x_n) - h\sum_{i=1}^{L}\lambda_i K_i^*, \tag{9.10}$$

其中, K_i^* 与 K_i 的区别在于: 用微分方程准确解 $y(x_n)$ 代替 K_i 中的 y_n 就得到 K_i^*. 参数 λ_i, c_i 和 a_{ij} 待定, 确定它们的原则和方法是: 将(9.10) 式中的 $y(x_{n+1})$ 在 x_n 处作 Taylor 展开, 将 K_i^* 在 $(x_n, y(x_n))$ 处作二元 Taylor 展开, 将展开式按 h 的幂次整理后, 令 T_{n+1} 中 h 的低次幂的系数为零, 使 T_{n+1} 首项中 h 的幂次尽量高, 比如使 $T_{n+1} = O(h^{p+1})$, 则称(9.9) 式为 L 级 p 阶 Runge-Kutta 方法(简称 R-K 法).

类似于显式 R-K 公式(9.9), 稍加改变, 就得到隐式 R-K 公式

$$y_{n+1} = y_n + h\sum_{i=1}^{L}\lambda_i K_i,$$

其中

$$K_i = f(x_n + c_i h, y_n + c_i h\sum_{j=1}^{L} a_{ij}K_j), \quad i = 1,2,\cdots,L.$$

它与显式 R-K 公式的区别在于：显式公式中，对系数 a_{ij} 求和的上限是 $i-1$，从而 a_{ij} 构成的矩阵是一个严格下三角阵. 而在隐式公式中，对系数 a_{ij} 求和的上限是 L，从而 a_{ij} 构成的矩阵是方阵，需要用迭代法求出近似斜率 $K_i(i = 1,2,\cdots,L)$. 推导隐式公式的思路和方法与显式公式类似.

9.2.2　几类显式 Runge-Kutta 方法

对于 $L = 2$，则

$$\begin{aligned} K_1 &= f(x_n, y_n),\\ K_2 &= f(x_n + c_2 h, y_n + c_2 h K_1),\\ y_{n+1} &= y_n + h(\lambda_1 K_1 + \lambda_2 K_2), \end{aligned}$$

其局部截断误差是

$$T_{n+1} = y(x_{n+1}) - y(x_n) - h(\lambda_1 K_1^* + \lambda_2 K_2^*). \tag{9.11}$$

将 T_{n+1} 中的各项作 Taylor 展开，并利用 $y'(x_n) = f(x_n, y(x_n))$，$y'' = f_x + f_y f$，则有

$$\begin{aligned} y(x_{n+1}) &= y(x_n) + hy'(x_n) + \frac{h^2}{2}y''(x_n) + \frac{h^3}{6}y'''(x_n) + O(h^4),\\ K_1^* &= f(x_n, y(x_n)) = y'(x_n),\\ K_2^* &= f(x_n + c_2 h, y(x_n) + c_2 h y'(x_n))\\ &= y'(x_n) + c_2 h y''(x_n) + \frac{c_2^2 h^2}{2}(f_{xx} + 2f_{xy}f + f_{yy}f^2) + O(h^3), \end{aligned}$$

将它们代入(9.11)式，整理后得

$$\begin{aligned} T_{n+1} = {} & (1 - \lambda_1 - \lambda_2)hy'(x_n) + \left(\frac{1}{2} - \lambda_2 c_2\right)h^2 y''(x_n)\\ & + h^3\left(\frac{1}{6}y'''(x_n) - \frac{\lambda_2 c_2^2}{2}(f_{xx} + 2f_{xy}f + f_{yy}f^2)\right) + O(h^4). \end{aligned}$$

选取 λ_1，λ_2 和 c_2，使方法的阶尽可能高，就是使 h 和 h^2 的系数为零，因为 h^3 的系数一般不为零. 于是得到方程组

$$\begin{cases} \lambda_1 + \lambda_2 = 1,\\ \lambda_2 c_2 = \dfrac{1}{2}. \end{cases}$$

显然,该方程组有无穷多组解,从而得到一族二级二阶 R-K 方法.

若以 c_2 为自由参数,取 $c_2 = \frac{1}{2}$ 得中点公式

$$y_{n+1} = y_n + hf\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}f(x_n, y_n)\right), \tag{9.12}$$

取 $c_2 = \frac{2}{3}$ 得 Heun 公式

$$y_{n+1} = y_n + \frac{h}{4}\left(f(x_n, y_n) + 3f\left(x_n + \frac{2}{3}h, y_n + \frac{2}{3}hf(x_n, y_n)\right)\right), \tag{9.13}$$

取 $c_2 = 1$ 得改进的 Euler 公式(9.6).

对于 $L = 3$ 的情形,要计算 3 个斜率的近似值:

$$\begin{aligned} K_1 &= f(x_n, y_n), \\ K_2 &= f(x_n + c_2 h, y_n + c_2 h K_1), \\ K_3 &= f(x_n + c_3 h, y_n + c_3 h(a_{31}K_1 + a_{32}K_2)). \end{aligned}$$

类似于二阶方法的推导,可以得三阶的方法,所得系数应满足的方程组是

$$\begin{cases} \lambda_1 + \lambda_2 + \lambda_3 = 1, \quad a_{21} = 1, \\ \lambda_2 c_2 + \lambda_3 c_3 = \frac{1}{2}, \quad \lambda_2 c_2^2 + \lambda_3 c_3^2 = \frac{1}{3}, \\ \lambda_3 c_2 c_3 a_{32} = \frac{1}{6}, \quad a_{31} + a_{32} = 1. \end{cases}$$

该方程组的解也是不唯一的. 常见的一种三级三阶方法是

$$\begin{aligned} K_1 &= f(x_n, y_n), \\ K_2 &= f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}K_1\right), \\ K_3 &= f(x_n + h, y_n - hK_1 + 2hK_2), \\ y_{n+1} &= y_n + \frac{h}{6}(K_1 + 4K_2 + K_3). \end{aligned}$$

对于 $L = 4$ 的情况,可进行类似推导. 最常用的四级四阶方法是如下的经典 R-K 方法

$$\begin{aligned} K_1 &= f(x_n, y_n), \\ K_2 &= f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}K_1\right), \\ K_3 &= f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}K_2\right), \\ K_4 &= f(x_n + h, y_n + hK_3), \\ y_{n+1} &= y_n + \frac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4). \end{aligned} \tag{9.14}$$

为了分析经典 R-K 公式的计算量和计算精度，将四阶经典 R-K 公式(9.14)与一阶显式 Euler 公式(9.2)及 2 阶改进的 Euler 公式(9.6)相比较. 一般来说，公式的级数越大，计算右端项 f 的次数越多，计算量越大. 在同样步长的情况下，Euler 方法每步只计算一个函数值，而经典方法要计算 4 个函数值. 4 阶 R-K 法的计算量差不多是改进的 Euler 公式的 2 倍，是显式 Euler 公式的 4 倍. 下面的例子中 Euler 方法用步长 h_1，2 阶改进的 Euler 法用步长 $2h_1$，而四阶的经典公式用步长 $4h_1$. 这样，从 x_n 到 x_n+4h_1，3 种方法都计算了 4 个函数值，计算量大体相当.

例 9.5　考虑初值问题

$$\begin{cases} y' = -y+1, \\ y(0)=0. \end{cases}$$

其解析解为 $y(x)=1-\mathrm{e}^{-x}$. 分别用 $h=0.025$ 的显式 Euler 方法，$h=0.05$ 的改进 Euler 法和 $h=0.1$ 的经典 R-K 方法计算到 $x=0.5$. 3 种方法在 x 方向每前进 0.1 都要计算 4 个右端函数值，计算量相当. 计算结果列于表 9-3. 从计算结果看，在工作量大致相同的情况下，还是经典 R-K 方法比其他两种方法的结果好得多. 在 $x=0.5$ 处，3 种方法的误差分别是 3.8×10^{-3}，1.3×10^{-4}，2.8×10^{-7}. 经典 R-K 方法对多数好条件问题($f_y<0$，参见下节单步法的稳定性)，能获得好的效果.

表 9－3

x_n	Euler 法 $h=0.025$	改进 Euler 法 $h=0.05$	经典 R-K 方法 $h=0.1$	准确解 $y(x_n)$
0.1	0.096 312	0.095 123	0.095 162 50	0.095 162 58
0.2	0.183 348	0.181 193	0.181 269 10	0.181 269 25
0.3	0.262 001	0.259 085	0.259 181 58	0.259 181 78
0.4	0.333 079	0.329 563	0.329 679 71	0.329 679 95
0.5	0.397 312	0.393 337	0.393 469 06	0.393 469 34

在微分方程数值解法的实际计算中，有个如何选择步长的问题. 因为单从每一步看，步长越小，截断误差就越小. 但随着步长的缩小，在一定求解范围内所要完成的步数就增加了. 步数的增加不但引起计算量的增大，而且可能导致舍入误差的严重积累.

在选择步长时，我们需要衡量和检验计算结果的精度，并依据所获得的精度处理步长. 下面以经典 R-K 方法为例进行说明.

从节点 x_n 出发，先以 h 为步长求出一个近似值 $y_{n+1}^{(h)}$，由于公式的局部截断误差为 $O(h^5)$，故有

$$y(x_{n+1}) - y_{n+1}^{(h)} \approx ch^5.$$

然后将步长折半，即 $\frac{h}{2}$ 为步长，从 x_n 跨两步到 x_{n+1}，再求得一个近似值 $y_{n+1}^{(\frac{h}{2})}$，每跨一步的截断误差约为 $c\left(\frac{h}{2}\right)^5$，因此有

$$y(x_{n+1}) - y_{n+1}^{(\frac{h}{2})} \approx 2c\left(\frac{h}{2}\right)^5.$$

比较上述两式，有

$$\frac{y(x_{n+1}) - y_{n+1}^{(\frac{h}{2})}}{y(x_{n+1}) - y_{n+1}^{(h)}} \approx \frac{1}{16}.$$

由此易得下列事后估计式

$$y(x_{n+1}) - y_{n+1}^{(\frac{h}{2})} \approx \frac{1}{15}(y_{n+1}^{(\frac{h}{2})} - y_{n+1}^{(h)}).$$

这样，我们可以通过检查步长折半前后再次计算结果的偏差

$$\Delta = |y_{n+1}^{(\frac{h}{2})} - y_{n+1}^{(h)}|$$

来判定所选的步长是否合适.

具体地说，对于给定的精度 ε，将按两种情况处理：如果 $\Delta > \varepsilon$，我们反复将步长折半进行计算，直到 $\Delta < \varepsilon$ 为止，这时取最终得到的 $y_{n+1}^{(\frac{h}{2})}$ 作为结果；如果 $\Delta < \varepsilon$，我们反复将步长加倍，直到 $\Delta > \varepsilon$ 为止，这时再将前一次步长折半的结果作为所要的结果. 这种通过加倍或折半处理步长的方法称作变步长方法. 虽然为了选择步长，每一步的计算量有所增加，但总体考虑还是值得的.

9.3 单步法的收敛性和稳定性

9.3.1 单步法的收敛性

数值解法的基本思想就是要通过某种离散化方法，将微分方程转化为某种差分方程（例如(9.8)式）来求解. 这种转化是否合理，还要看差分方程的解 y_n 是否收敛到微分方程的准确解 $y(x_n)$.

定义 9.3 对于任意固定的 $x_n = x_0 + nh$，若对于初值问题(9.1)的显式单步法(9.8)产生的近似解 y_n，均有 $y_n \to y(x_n)$ $(h \to 0$，同时 $n \to \infty)$，则称该方法是收敛的.

在定义中，x_n 是固定的点，$h \to 0$ 时有 $n \to \infty$，n 不是固定的. 显然，若方法是

收敛的，则在固定点 x_n 处的整体截断误差 $e_n = y(x_n) - y_n$ 趋于零. 下面给出方法收敛的条件.

定理 9.1　设初值问题(9.1)的单步法(9.8)是 p 阶的($p \geqslant 1$)，且函数 φ 满足对 y 的 Lipschitz 条件，即存在常数 $L > 0$，使

$$|\varphi(x, y_1, h) - \varphi(x, y_2, h)| \leqslant L|y_1 - y_2|,$$

对一切 $y_1, y_2 \in \mathbf{R}$ 成立，则方法(9.8)收敛，且 $y(x_n) - y_n = O(h^p)$.

证　仍记 $e_n = y(x_n) - y_n$，根据局部截断误差的定义

$$y(x_{n+1}) = y(x_n) + h\varphi(x_n, y(x_n), h) + T_{n+1}.$$

将此式与(9.8)式相减得

$$e_{n+1} = e_n + h(\varphi(x_n, y(x_n), h) - \varphi(x_n, y_n, h)) + T_{n+1}.$$

因为(9.8)式是 p 阶的，所以存在 h_0，当 $0 < h \leqslant h_0$ 时有 $|T_{n+1}| \leqslant ch^{p+1}$. 再用 φ 的 Lipschitz 条件有

$$|e_{n+1}| \leqslant |e_n| + hL|e_n| + ch^{p+1}.$$

为了方便，记 $\alpha = 1 + hL$，$\beta = ch^{p+1}$，即有 $|e_{n+1}| \leqslant \alpha|e_n| + \beta$. 由此可推得

$$\begin{aligned} |e_n| &\leqslant \alpha|e_{n-1}| + \beta \leqslant \alpha^2|e_{n-2}| + \alpha\beta + \beta \leqslant \cdots \\ &\leqslant \alpha^n|e_0| + \beta(\alpha^{n-1} + \alpha^{n-2} + \cdots + \alpha + 1). \end{aligned}$$

利用关系式

$$e^{Lh} = 1 + Lh + \frac{(Lh)^2}{2} + \cdots \geqslant 1 + Lh,$$

$$\alpha^n = (1 + Lh)^n \leqslant e^{nLh} = e^{L(x_n - x_0)},$$

可以得到

$$|e_n| \leqslant |e_0| e^{L(x_n - x_0)} + (e^{L(x_n - x_0)} - 1)ch^p L^{-1}.$$

现在取 $y_0 = y(x_0)$，有 $e_0 = 0$，于是 $e_n = O(h^p)$. 定理得证.

容易证明，如果(9.1)式的 f 满足对 y 的 Lipschitz 条件，且初值是准确的，则显式 Euler 法、改进的 Euler 法和 R-K 方法是收敛的.

由定理 9.1 说明，f 关于 y 满足 Lipschitz 条件是使单步收敛的充分条件，而且，还说明一个方法的整体截断误差比局部截断误差低一阶. 所以，常常通过求出局部截断误差去了解整体截断误差的大小.

单步法的显式形式(9.8)可写成

$$\varphi(x_n, y_n, h) = \frac{y_{n+1} - y_n}{h}. \tag{9.15}$$

称 $\varphi(x_n, y_n, h)$ 为增量函数. 对于收敛的方法，固定 $x = x_n$，有 $y_n \to y(x_n)$ $(h \to 0)$，从而 $\frac{y_{n+1} - y_n}{h} \to y'(x_n)(h \to 0)$. 对于(9.15)式，我们自然要考虑

$\varphi(x_n,y_n,h)\to f(x_n,y(x_n))(h\to 0)$ 是否成立. 这就是相容性问题.

定义 9.4 若方法(9.8)的增量函数 φ 满足 $\varphi(x,y,0)=f(x,y)$,则称方法(9.8)与初值问题(9.1)是相容的.

相容性说明数值计算的差分方程(9.15)趋于(9.1)式中微分方程. 我们本章讨论的数值方法都是与原初值问题相容的.

9.3.2 单步法的稳定性

对于一种收敛的相容的差分方程,由于计算过程中舍入误差总会存在,我们需要讨论其数值稳定性. 一个不稳定的差分方程会使计算解失真或计算失败.

为了讨论方便起见,将(9.1)式中的 $f(x,y)$ 在解域内某一点 (a,b) 作 Taylor 展开并局部线性化,即

$$\begin{aligned} y' &= f(x,y)\approx f(a,b)+(x-a)f_x(a,b)+(y-b)f_y(a,b)\\ &= f_y(a,b)y+c_1x+c_2. \end{aligned}$$

令 $\lambda=f_y(a,b)$

$$u=y+\frac{c_1}{\lambda}x+\frac{c_1}{\lambda^2}+\frac{c_2}{\lambda},$$

利用线性化的关系,可得 $u'\approx\lambda u$. 因此,我们通过如下的试验方程

$$y'=\lambda y \tag{9.16}$$

讨论数值方法的稳定性. 当某一步 y_n 有舍入误差时,若以后的计算中不会逐步扩大,则称这种稳定性为绝对稳定性.

现讨论显式 Euler 法的稳定性. 将显式 Euler 法用于试验方程(9.16),有

$$y_{n+1}=(1+\lambda h)y_n.$$

当 y_n 有舍入误差时,其近似值为 $\tilde{y}_n$,从而有

$$\tilde{y}_{n+1}=(1+\lambda h)\tilde{y}_n.$$

令 $\varepsilon_n=y_n-\tilde{y}_n$,得到误差传播方程

$$\varepsilon_{n+1}=(1+\lambda h)\varepsilon_n.$$

令 $E(\lambda h)=1+\lambda h$,只要 $|E(\lambda h)|\leqslant 1$,则显式 Euler 方法的解和误差都不会恶性发展,即 $-2\leqslant\lambda h\leqslant 0$ 时,显式 Euler 方法是稳定的,即是条件稳定的.

对于梯形方法,应用于试验方程后,有

$$y_{n+1}=\frac{1+\lambda\dfrac{h}{2}}{1-\lambda\dfrac{h}{2}}y_n,$$

同理,有误差方程 $\varepsilon_{n+1}=E(\lambda h)\varepsilon_n$,其中 $E(\lambda h)=\dfrac{1+\lambda\dfrac{h}{2}}{1-\lambda\dfrac{h}{2}}$. 因此当 $\lambda\leqslant 0$ 时,梯形

方法是稳定的.

一般地,在试验方程(9.16)中,我们只考虑$\lambda<0$的情形,而对$\lambda=f_y>0$的情形,我们认为微分方程是不稳定的.比如,将显式 Euler 方法用于(9.1)式中的方程,有

$$\varepsilon_{n+1}=\varepsilon_n+h(f(x_n,y_n)-f(x_n,\tilde{y}_n))=(1+hf_y(x_n,\eta))\varepsilon_n.$$

当$f_y(x_n,\eta)>0$时,有$1+hf_y(x_n,\lambda\eta)>1$.

对于每一种单步法应用于试验方程(9.16),可得

$$y_{n+1}=E(\lambda\eta)y_n, \tag{9.17}$$

然而,对于不同的单步法,$E(\lambda\eta)$有不同的表达式.

定义 9.5　若(9.17)式中的$|E(\lambda\eta)|\leqslant 1$,则称对应的单步法是绝对稳定的.在复平面上,$\lambda\eta$满足$|E(\lambda\eta)|\leqslant 1$的区域,称为方法的绝对稳定区域,它与实轴的交称为绝对稳定区间.

一些单步法的$E(\lambda\eta)$表达式和它们的绝对稳定区间列于表 9-4.从表中可见,隐式方法比显式方法的绝对稳定性好.

表 9-4

方　法	$E(\lambda\eta)$	绝对稳定区间
Euler 法	$1+\lambda h$	$-2\leqslant\lambda h\leqslant 0$
改进的 Euler 法	$1+\lambda h+\frac{(\lambda h)^2}{2}$	$-2\leqslant\lambda h\leqslant 0$
三阶 R-K 法	$1+\lambda h+\frac{(\lambda h)^2}{2}+\frac{(\lambda h)^3}{6}$	$-2.51\leqslant\lambda h\leqslant 0$
四阶 R-K 法	$1+\lambda h+\frac{(\lambda h)^2}{2}+\frac{(\lambda h)^3}{6}+\frac{(\lambda h)^4}{24}$	$-2.785\leqslant\lambda h\leqslant 0$
隐式 Euler 法	$\frac{1}{1-\lambda h}$	$-\infty<\lambda h\leqslant 0$
梯形式	$\frac{1+\lambda\frac{h}{2}}{1-\lambda\frac{h}{2}}$	$-\infty<\lambda\leqslant 0$

例 9.6　分别取$h=1,2,4$,用经典 R-K 方法求解

$$\begin{cases}y'=-y+x-\mathrm{e}^{-1},\\ y(1)=0,\end{cases}$$

其准确解为$y(x)=\mathrm{e}^{-x}+x-1-\mathrm{e}^{-1}$.

解　本题$\lambda=-1$,λh分别为$-1,-2,-4$.由表 9-4 可知,当$h\leqslant 2.785$时,该方法才稳定.计算结果列于表 9-5.

表 9－5

x_n	$h=1$ 的解	$h=2$ 的解	$h=4$ 的解	准确解
5	3.639 4	3.673 0	5.471 5	3.638 9
9	7.632 3	7.636 7	16.829 1	7.632 2
13	11.632 1	11.632 6	57.617 1	11.632 1

由表 9－5 可见，$h=1$ 和 $h=2$ 时，计算结果确实稳定，$h=4$，结果发散. 此外，h 为 1 的计算精度比 h 为 2 的计算精度高. 因为 h 越小，方法的截断误差越小. 但若 h 过分小的话，计算步数非常多，其累积误差会增加. 所以，实际计算时，应选取合适的步长，常常采用自动变步长的 R-K 方法.

9.4 线性多步法

常微分方程初值问题(9.1)的数值解法中，除了像 Runge-Kutta 型公式等单步法之外，还有另一种类型的解法，即某一步解的公式不仅与前一步解的值有关，而且与前若干步解的值有关，利用前面多步的信息预测下一步的值，这就是多步法的基本思想，可能期望获得较高的精度. 构造多步法有多种途径，下面先讨论基于数值积分的方法.

9.4.1 基于数值积分的方法

将(9.1)式中的方程在区间$[x_n, x_{n+1}]$上积分，可以得到

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(x, y(x))\mathrm{d}x. \tag{9.18}$$

如推导 Newton-Cotes 求积公式一样，用等距节点的插值多项式来替代被积函数，再对插值多项式积分，这样就得到一系列求积公式.

例如，用梯形方法计算积分项

$$\int_{x_n}^{x_{n+1}} f(x, y(x))\mathrm{d}x \approx \frac{h}{2}(f(x_n, y(x_n)) + f(x_{n+1}, y(x_{n+1}))),$$

代入(9.18)式有

$$y(x_{n+1}) \approx y(x_n) + \frac{h}{2}(f(x_n, y(x_n)) + f(x_{n+1}, y(x_{n+1}))).$$

据此即可导出(9.4)式.

一般地，设由 $r+1$ 个数据点(x_n, f_n)，(x_{n-1}, f_{n-1})，…，(x_{n-r}, f_{n-r}) 构造插值多项式 $P_r(x)$，这里 $f_k = f(x_k, y_k)$，$x_k = x_0 + kh$. 运用插值公式有

$$l_j(x)=\prod_{k=0,k\neq j}^{r}\frac{x-x_{n-k}}{x_{n-j}-x_{n-k}},$$

$$P_r(x)=\sum_{j=0}^{r}f_{n-j}l_j(x).$$

将(9.18)式离散化即得下列计算公式

$$y_{n+1}=y_n+h\sum_{j=0}^{r}\alpha_{rj}f_{n-j},\tag{9.19}$$

其中

$$\alpha_{rj}=\frac{1}{h}\int_{x_n}^{x_{n+1}}l_j(x)\mathrm{d}x=\int_0^1\prod_{k=0,k\neq j}^{r}\frac{t+k}{k-j}\mathrm{d}t,\quad j=0,1,\cdots,r.$$

由此可得(9.19)式中的系数,其具体数值见表 9-6.(9.19)式是一个 $r+1$ 步的显式公式,称为 Adams 显式公式.$r=0$ 时,即为 Euler 公式.

表 9-6

j	0	1	2	3	4
α_{0j}	1				
$2\alpha_{1j}$	3	−1			
$12\alpha_{2j}$	23	−16	5		
$24\alpha_{3j}$	55	−59	37	−9	
$720\alpha_{4j}$	1 901	−2 774	2 616	−1 274	251

在上述 Adams 显式公式的推导中,选用了 $x_n,x_{n-1},\cdots,x_{n-r}$ 作为插值节点.这样的插值多项式 $P_r(x)$ 在求积区间$[x_n,x_{n+1}]$上逼近 $f(x,y(x))$ 是一个外推结果.为了改善逼近效果,我们变外推为内插,即改用 $x_{n+1},x_n,\cdots,x_{n-r+1}$ 为插值节点,用数据点$(x_{n+1},f_{n+1}),(x_n,f_n),\cdots,(x_{n-r+1},f_{n-r+1})$构造插值多项式$P_r(x)$,则有

$$l_j(x)=\prod_{k=0,k\neq j}^{r}\frac{x-x_{n-k+1}}{x_{n-j+1}-x_{n-k+1}},$$

$$P_r(x)=\sum_{j=0}^{r}f_{n-j+1}l_j(x).$$

于是,我们有如下的计算公式

$$y_{n+1}=y_n+h\sum_{j=0}^{r}\beta_{rj}f_{n-j+1},\tag{9.20}$$

其中

$$\beta_{rj}=\frac{1}{h}\int_{x_n}^{x_{n+1}}l_j(x)\mathrm{d}x=\int_{-1}^{0}\prod_{k=0,k\neq j}^{r}\frac{t+k}{k-j}\mathrm{d}t,\quad j=0,1,\cdots,r,$$

其具体数值见表 9－7. (9.20) 式是隐式公式，称为 Adams 隐式公式. $r=0,1$ 时，分别为隐式 Euler 公式和梯形公式.

表 9－7

j	0	1	2	3	4
β_{0j}	1				
$2\beta_{1j}$	1	1			
$12\beta_{2j}$	5	8	−1		
$24\beta_{3j}$	9	19	−5	1	
$720\beta_{4j}$	251	646	−264	106	−19

对于隐式公式(9.20)，需要用迭代求解. 确定 y_{n+1} 的迭代公式为

$$y_{n+1}^{(s+1)}=y_n+h(\beta_{r0}f(x_{n+1},y_{n+1}^{(s)})+\sum_{j=1}^{r}\beta_{rj}f_{n-j+1}),\quad s=0,1,\cdots,$$

迭代收敛条件为 $h\mid\beta_{r0}\mid L<1$，其中 L 为 f 关于 y 的 Lipschitz 常数.

利用插值多项式的余项，可以求出 Adams 方法的局部截断误差. 当然也可以从得到的显式和隐式 Adams 公式，由局部截断误差的定义来求出方法的局部截断误差. 表 9－8 中列出了它们的局部截断误差的主项，由表 9－8 可以看出，Adams 隐式方法的局部截断误差要小.

表 9－8

r	0	1	2	3
Adams 显式公式	$\frac{1}{2}h^2y''(x_n)$	$\frac{5}{12}h^3y'''(x_n)$	$\frac{3}{8}h^4y^{(4)}(x_n)$	$\frac{251}{720}h^5y^{(5)}(x_n)$
Adams 隐式公式	$-\frac{1}{2}h^2y''(x_n)$	$-\frac{1}{12}h^3y'''(x_n)$	$-\frac{1}{24}h^4y^{(4)}(x_n)$	$-\frac{19}{720}h^5y^{(5)}(x_n)$

9.4.2 基于 Taylor 展开的方法

基于数值积分可以构造出一系列求解常微分方程的数值计算公式，计算公式由插值多项式唯一确定. 下面介绍基于 Taylor 展开的待定系数法，它可灵活地构造出线性多步法. 对固定的步数，可以选取待定系数使线性多步法的阶尽可能高. 还可以根据需要，确定显式还是隐式.

设构造如下具有 p 阶精度的线性多步公式

$$y_{n+1}=\alpha_0y_n+\alpha_1y_{n-1}+\cdots+\alpha_ry_{n-r}+h(\beta_{-1}f_{n+1}+\beta_0f_n+\cdots+\beta_rf_{n-r}).\tag{9.21}$$

当 $\beta_{-1}=0$ 时，则(9.21)式为显式多步式. 当 $\beta_{-1}\neq 0$ 时，(9.21)式为隐式多步式. 它们的局部截断误差为

$$T_{n+1}=y(x_{n+1})-\Big(\sum_{k=0}^{r}\alpha_k y(x_{n-k})+h\sum_{k=-1}^{r}\beta_k f(x_{n-k},y(x_{n-k}))\Big),$$

利用原微分方程，有

$$T_{n+1}=y(x_{n+1})-\Big(\sum_{k=0}^{r}\alpha_k y(x_{n-k})+h\sum_{k=-1}^{r}\beta_k y'(x_{n-k})\Big). \tag{9.22}$$

现利用 Taylor 展开定理，确定线性多步公式(9.21)中的待定参数 α_k,β_k，使它达到 p 阶精度，即 $T_{n+1}=O(h^{p+1})$.

对(9.22)式的右端各项在 x_n 点处作 Taylor 展开有

$$y(x_{n-k})=\sum_{j=0}^{p}\frac{(-kh)^j}{j!}y^{(j)}(x_n)+\frac{(-kh)^{p+1}}{(p+1)!}y^{(p+1)}(x_n)+O(h^{p+2}),$$

$$y'(x_{n-k})=\sum_{j=1}^{p}\frac{(-kh)^j}{(j-1)!}y^{(j)}(x_n)+\frac{(-kh)^{p+1}}{p!}y^{(p+1)}(x_n)+O(h^{p+1}).$$

将它们代入(9.22)式整理后得

$$\begin{aligned}T_{n+1}=&(1-\sum_{k=0}^{r}\alpha_k)y(x_n)+\sum_{j=1}^{p}\frac{h^j}{j!}(1-\sum_{k=1}^{r}(-k)^j\alpha_k\\&-j\sum_{k=-1}^{r}(-k)^{j-1}\beta_k)y^{(j)}(x_n)+\frac{h^{p+1}}{(p+1)!}(1-\sum_{k=1}^{r}(-k)^{p+1}\alpha_k\\&-(p+1)\sum_{k=-1}^{r}(-k)^p\beta_k)y^{(p+1)}(x_n)+O(h^{p+2}).\end{aligned}$$

使 $y(x_n),h,h^2,\cdots,h^p$ 的系数为零，得到关于 α_k 和 β_k 的线性方程组

$$\begin{cases}\displaystyle\sum_{k=0}^{r}\alpha_k=1,\\[2ex]\displaystyle\sum_{k=1}^{r}(-k)^j\alpha_k+j\sum_{k=-1}^{r}(-k)^{j-1}\beta_k=1,\quad j=1,2,\cdots,p.\end{cases} \tag{9.23}$$

而且得到线性多步法的局部截断误差

$$\begin{aligned}T_{n+1}=&\frac{h^{p+1}}{(p+1)!}\Big(1-\sum_{k=1}^{r}(-k)^{p+1}\alpha_k-(p+1)\sum_{k=-1}^{r}(-k)^p\beta_k\Big)y^{(p+1)}(x_n)\\&+O(h^{p+2}).\end{aligned}$$

当参数 α_k 和 β_k 满足(9.23)式时，线性多步法公式(9.21)达到 p 阶精度，满足(9.23)式的 α_k 和 β_k 可能有多组解. 下面我们构造几个著名的四阶线性多步公式，考虑下列形式的公式

$$\begin{aligned}y_{n+1}=&\alpha_0 y_n+\alpha_1 y_{n-1}+\alpha_2 y_{n-2}+\alpha_3 y_{n-3}\\&+h(\beta_{-1}f_{n+1}+\beta_0 f_n+\beta_1 f_{n-1}+\beta_2 f_{n-2}+\beta_3 f_{n-3}),\end{aligned} \tag{9.24}$$

$$T_{n+1}=\frac{h^5}{5!}\Big(1-\sum_{k=1}^{3}(-k)^5\alpha_k-5\sum_{k=-1}^{3}(-k)^4\beta_k\Big)y^{(5)}(x_n)+O(h^6). \quad (9.25)$$

由于$r=3,p=4$,由(9.23)式得到5个方程,而(9.24)式中有9个未知量,因此,(9.24)式中有4个自由度.

若取$\beta_{-1}=0,\alpha_1=\alpha_2=\alpha_3=0$,由(9.23)式得到其他5个待定参数的方程组,解之得

$$\alpha_0=1,\quad \beta_0=\frac{55}{24},\quad \beta_1=-\frac{59}{24},\quad \beta_2=\frac{37}{24},\quad \beta_3=-\frac{9}{24}.$$

代入(9.24)式和(9.25)式,得到常用的四步四阶显式 Admas 公式和它的余项

$$y_{n+1}=y_n+\frac{h}{24}(55f_n-59f_{n-1}+37f_{n-2}-9f_{n-3}), \quad (9.26)$$

$$T_{n+1}=\frac{251}{720}h^5y^{(5)}(x_n)+O(h^6). \quad (9.27)$$

若取$\beta_{-1}=0,\alpha_0=\alpha_1=\alpha_2=0$,由(9.23)式得到其他5个待定参数的方程组,解之得

$$\alpha_3=1,\quad \beta_0=\frac{8}{3},\quad \beta_1=-\frac{4}{3},\quad \beta_2=\frac{8}{3},\quad \beta_3=0.$$

由此构造成著名的四步四阶显式 Milne 公式和它的余项

$$y_{n+1}=y_{n-3}+\frac{4}{3}h(2f_n-f_{n-1}+2f_{n-2}), \quad (9.28)$$

$$T_{n+1}=\frac{14}{45}h^5y^{(5)}(x_n)+O(h^6). \quad (9.29)$$

若取$\alpha_1=\alpha_2=\alpha_3=0,\beta_3=0$,由(9.23)式得到其他5个待定参数的方程组,从而得三步四阶隐式 Admas 公式及其余项

$$y_{n+1}=y_n+\frac{h}{24}(9f_{n+1}+19f_n-5f_{n-1}+f_{n-2}), \quad (9.30)$$

$$T_{n+1}=-\frac{19}{720}h^5y^{(5)}(x_n)+O(h^6) \quad (9.31)$$

若取$\alpha_1=\alpha_3=0,\beta_2=\beta_3=0$,求解(9.23)式得著名的三步四阶隐式 Hamming 公式及其余项

$$y_{n+1}=\frac{1}{8}(9y_n-y_{n-2})+\frac{3}{8}h(f_{n+1}+2f_n-f_{n-1}), \quad (9.32)$$

$$T_{n+1}=-\frac{1}{40}h^5y^{(5)}(x_n)+O(h^6). \quad (9.33)$$

若取$\alpha_0=\alpha_2=\alpha_3=0,\beta_3=0$,求解(9.23)式得到隐式 Simpson 公式及其余项

$$y_{n+1} = y_{n-1} + \frac{h}{3}(f_{n+1} + 4f_n + f_{n-1}),$$

$$T_{n+1} = -\frac{1}{90}h^5 y^{(5)}(x_n) + O(h^6).$$

例 9.7　分别取 $h = 0.2, 2$，用 4 阶显式 Milne 公式和 4 阶隐式 Hamming 公式求解例 9.6 所给的初值问题.

解　我们用单步法提供多步法的开始值. 由 4 阶经典 R-K 公式为 Milne 公式提供开始值 y_1, y_2, y_3，为 Hamming 公式提供开始值 y_1, y_2. $h = 0.2$ 和 $h = 2$ 时的计算结果及与准确解之间的误差分别列于表 9－9 和表 9－10.

表 9－9　($h = 0.2$)

x_n	Milne 方法	误差	Hamming 方法	误差
2.2	0.942 942 68	-1.8×10^{-5}	0.942 919 55	4.2×10^{-6}
2.4	1.122 833 49	5.0×10^{-6}	1.122 833 86	4.6×10^{-6}
2.6	1.306 432 14	-3.8×10^{-5}	1.306 389 30	4.8×10^{-6}
2.8	1.492 916 25	1.4×10^{-5}	1.492 925 82	4.8×10^{-6}
3.0	1.681 954 50	-4.7×10^{-5}	1.681 902 99	4.6×10^{-6}

表 9－10　($h = 2$)

x_n	Milne 方法	误差	Hamming 方法	误差
7	5.645 745	-1.3×10^{-2}	5.645 745	-1.3×10^{-2}
9	7.382 325	2.5×10^{-1}	7.637 126	-4.9×10^{-3}
11	10.905 316	-1.3	9.635 636	-3.5×10^{-3}
13	4.143 831	7.5	11.632 261	-1.4×10^{-4}
15	58.310 717	-4.5×10^{1}	13.632 240	-1.1×10^{-3}
17	−249.662 672	2.7×10^{2}	15.631 690	4.3×10^{-4}

从表 9－9 看出，两种多步法的计算精度都很高，Hamming 公式比 Milne 公式更精确. 这是因为 Hamming 公式的截断误差主项的系数比 Milne 公式小. 从表 9－10 看到，当计算步长变大后，显式多步法 Milne 公式的计算结果误差增大，不稳定，而隐式多步法 Hamming 公式的计算结果仍然是稳定的，这说明隐式公式的稳定性比同阶的显式公式好.

经典 R-K 法和上述四阶线性多步公式都是 4 阶精度，但每前进一步，前者要计算 4 次微分方程的右端函数值，而后者只要计算一次新的右端函数值，计算量减小了.

9.4.3 预估 — 校正算法

显式多步法容易计算，但其精度和稳定性没有相应的隐式方法好. 然而，隐式多步法需解方程，如果初值选得不当，则计算量较大. 因此，设法选取好的迭代初值是必要的. 初值的自然选取是采用同阶显式多步法计算得到的解作为隐式方法迭代的初值. 这样，迭代次数不会多. 若只迭代一次，则这样的算法就是预估 — 校正算法. 对于线性多步法，常用的预估 — 校正方法有四阶 Admas 显隐式预估 — 校正公式和 Milne-Hamming 方法.

(1) Adams 预估 — 校正公式.

由(9.26)式作为预估公式，由(9.30)式作为校正公式，构成 Adams 预估 — 校正公式

$$y_{n+1}^{p} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}),$$

$$y_{n+1} = y_n + \frac{h}{24}(9f(x_{n+1}, y_{n+1}^{p}) + 19f_n - 5f_{n-1} + f_{n-2}).$$

若需作进一步的修正，则记上式所得的 $y_{n+1} = y_{n+1}^{c}$ 由(9.27)式和(9.31)式有

$$T_{n+1}^{p} = y(x_{n+1}) - y_{n+1}^{p} \approx \frac{251}{720}h^5 y^{(5)}(x_n),$$

$$T_{n+1}^{c} = y(x_{n+1}) - y_{n+1}^{c} \approx -\frac{19}{720}h^5 y^{(5)}(x_n).$$

于是得到

$$y(x_{n+1}) - y_{n+1}^{p} \approx -\frac{251}{270}(y_{n+1}^{p} - y_{n+1}^{c}),$$

$$y(x_{n+1}) - y_{n+1}^{c} \approx \frac{19}{270}(y_{n+1}^{p} - y_{n+1}^{c}).$$

由此可见，若记

$$\bar{y}_{n+1}^{p} = y_{n+1}^{p} + \frac{251}{270}(y_{n+1}^{c} - y_{n+1}^{p}),$$

$$\bar{y}_{n+1}^{c} = y_{n+1}^{c} - \frac{19}{270}(y_{n+1}^{c} - y_{n+1}^{p}),$$

则 $\bar{y}_{n+1}^{p}$，$\bar{y}_{n+1}^{c}$ 分别比 y_{n+1}^{p}，y_{n+1}^{c} 更好. 但注意到，$\bar{y}_{n+1}^{p}$ 的表达式中，y_{n+1}^{c} 是未知的，因此改为

$$\bar{y}_{n+1}^{p} = y_{n+1}^{p} + \frac{251}{270}(y_n^{c} - y_n^{p}).$$

这样，得到下面修正的 Adams 预估 — 校正公式

预估：$y_{n+1}^{p}=y_n+\frac{h}{24}(55f_n-59f_{n-1}+37f_{n-2}-9f_{n-3})$，

修正：$\bar{y}_{n+1}^{p}=y_{n+1}^{p}+\frac{251}{270}(y_n^c-y_n^p)$，

校正：$y_{n+1}^{c}=y_n+\frac{h}{24}(9f(x_{n+1},\bar{y}_{n+1}^{p})+19f_n-5f_{n-1}+f_{n-2})$，

修正：$y_{n+1}=y_{n+1}^{c}-\frac{19}{270}(y_{n+1}^{c}-y_{n+1}^{p})$.

在计算时，可调节计算步长 h，使 $\left|-\frac{19}{270}(y_{n+1}^{c}-y_{n+1}^{p})\right|<\varepsilon$，其中 ε 是要求达到的计算精度. 开始值 y_1,y_2,y_3 由同阶单步法提供，当计算 y_4 时，可取 $y_3^c=y_3^p$.

(2) 修正 Hamming 公式.

将 Milne 公式(9.28) 和 Hamming 公式(9.32) 结合，构成 Milne-Hamming 预估 — 校正公式

$$y_{n+1}^{p}=y_{n-3}+\frac{4}{3}h(2f_n-f_{n-1}+2f_{n-2}),$$

$$y_{n+1}=\frac{1}{8}(9y_n-y_{n-2})+\frac{3}{8}h(f(x_{n+1},y_{n+1}^{p})+2f_n-f_{n-1}).$$

若需作进一步的修正，则记上式所得的 $y_{n+1}=y_{n+1}^{c}$，由(9.29) 式和(9.33) 式有

$$T_{n+1}^{p}=y(x_{n+1})-y_{n+1}^{p}\approx\frac{14}{45}h^5y^{(5)}(x_n),$$

$$T_{n+1}^{c}=y(x_{n+1})-y_{n+1}^{c}\approx-\frac{1}{40}h^5y^{(5)}(x_n).$$

于是得到

$$y(x_{n+1})-y_{n+1}^{p}\approx\frac{112}{121}(y_{n+1}^{c}-y_{n+1}^{p}),$$

$$y(x_{n+1})-y_{n+1}^{c}\approx-\frac{9}{121}(y_{n+1}^{c}-y_{n+1}^{p}).$$

由此分别得 Milne 和 Hamming 公式的修正公式

$$\bar{y}_{n+1}^{p}=y_{n+1}^{p}+\frac{112}{121}(y_{n+1}^{c}-y_{n+1}^{p}),$$

$$\bar{y}_{n+1}^{c}=y_{n+1}^{c}-\frac{9}{121}(y_{n+1}^{c}-y_{n+1}^{p}).$$

从而构成如下的修正 Hamming 公式

预估：$y_{n+1}^{p}=y_{n-3}+\frac{4}{3}h(2f_n-f_{n-1}+2f_{n-2})$，

修正：$\bar{y}_{n+1}^{p} = y_{n+1}^{p} + \frac{112}{121}(y_n^c - y_n^p)$，

校正：$y_{n+1}^{c} = \frac{1}{8}(9y_n - y_{n-2}) + \frac{3}{8}h(f(x_{n+1}, \bar{y}_{n+1}^{p}) + 2f_n - f_{n-1})$，

修正：$y_{n+1} = y_{n+1}^{c} - \frac{9}{121}(y_{n+1}^{c} - y_{n+1}^{p})$.

在计算时，可调节计算步长 h，使 $\left| -\frac{9}{121}(y_{n+1}^{c} - y_{n+1}^{p}) \right| < \varepsilon$. 开始值 y_1, y_2 由同阶单步法提供，当计算 y_3 时，可取 $y_2^c = y_2^p$.

例 9.8 取 $h = 0.2$，用 Milne-Hamming 预估—校正公式和修正 Hamming 公式求解例 9.6 所给的初值问题.

解 用经典 R-K 法提供开始值，计算结果列于表 9-11. 将表 9-9 与表 9-11 所示的计算结果进行比较，它们的计算精度排列次序是：修正 Hamming 公式的精度最好，其次是隐式 Hamming 公式，再次是 Milne-Hamming 预估—校正公式，最后是 Milne 公式.

表 9-11

x_n	Milne-Hamming	误差	修正 Hamming	误差
2.2	0.942 916 25	7.5×10^{-6}	0.942 924 49	-7.8×10^{-7}
2.4	1.122 828 72	9.8×10^{-6}	1.122 839 55	-1.0×10^{-6}
2.6	1.306 382 71	1.1×10^{-5}	1.306 395 37	-1.2×10^{-6}
2.8	1.492 918 16	1.2×10^{-5}	1.492 931 84	-1.2×10^{-6}
3.0	1.681 894 67	1.3×10^{-5}	1.681 908 79	-1.2×10^{-6}

9.5 一阶方程组的数值解法

9.5.1 一阶方程组和高阶方程

考虑一阶常微分方程组的初值问题

$$\begin{cases} y_i' = f_i(x, y_1, y_2, \cdots, y_N), \\ y_i(x_0) = y_{0i}, \quad i = 1, 2, \cdots, N. \end{cases} \tag{9.34}$$

若将其中的未知函数、方程的右端项都表示成向量形式

$$\boldsymbol{y} = (y_1, y_2, \cdots, y_N)^{\mathrm{T}}, \boldsymbol{f} = (f_1, f_2, \cdots, f_N)^{\mathrm{T}},$$

初始条件表示成

$$\boldsymbol{y}(x_0) = \boldsymbol{y}_0 = (y_{01}, y_{02}, \cdots, y_{0N})^{\mathrm{T}},$$

那么,(9.34) 式可以写成

$$\begin{cases} \boldsymbol{y}' = \boldsymbol{f}(x,\boldsymbol{y}), \\ \boldsymbol{y}(x_0) = \boldsymbol{y}_0. \end{cases} \tag{9.35}$$

可见,(9.35) 式在形式上与一个方程的初值问题一样. 关于一个方程的初值问题的数值方法均适用于方程组. 相应的理论问题也可类似地讨论. 下面仅写出两种数值方法作说明.

梯形方法

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \frac{h}{2}(\boldsymbol{f}(x_n,\boldsymbol{y}_n) + \boldsymbol{f}(x_{n+1},\boldsymbol{y}_{n+1})),$$

或表达为

$$y_{n+1,i} = y_{ni} + \frac{h}{2}(f_i(x_n,\boldsymbol{y}_n) + f_i(x_{n+1},\boldsymbol{y}_{n+1})), \quad i = 1,2,\cdots,N,$$

其中 y_{ni} 是第 i 个因变量 $y_i(x)$ 在节点 x_n 处的近似值,相应地,$f_i(x_n,\boldsymbol{y}_n) = f_i(x_n,y_{n1},y_{n2},\cdots,y_{nN})$.

经典 R-K 方法

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \frac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4),$$

其中

$$K_1 = \boldsymbol{f}(x_n,\boldsymbol{y}_n), \quad K_2 = \boldsymbol{f}\left(x_n + \frac{h}{2},\boldsymbol{y}_n + \frac{h}{2}K_1\right),$$

$$K_3 = \boldsymbol{f}\left(x_n + \frac{h}{2},\boldsymbol{y}_n + \frac{h}{2}K_2\right), \quad K_4 = \boldsymbol{f}(x_n + h,\boldsymbol{y}_n + hK_3),$$

或表达为

$$y_{n+1,i} = y_{ni} + \frac{h}{6}(K_{1i} + 2K_{2i} + 2K_{3i} + K_{4i}), \quad i = 1,2,\cdots,N,$$

其中

$$K_{1i} = f_i(x_n,y_{n1},y_{n2},\cdots,y_{nN}),$$

$$K_{2i} = f_i\left(x_n + \frac{h}{2},y_{n1} + \frac{h}{2}K_{11},y_{n2} + \frac{h}{2}K_{12},\cdots,y_{nN} + \frac{h}{2}K_{1N}\right),$$

$$K_{3i} = f_i\left(x_n + \frac{h}{2},y_{n1} + \frac{h}{2}K_{21},y_{n2} + \frac{h}{2}K_{22},\cdots,y_{nN} + \frac{h}{2}K_{2N}\right),$$

$$K_{4i} = f_i(x_n + h,y_{n1} + hK_{31},y_{n2} + hK_{32},\cdots,y_{nN} + hK_{3N}).$$

对于高阶方程,可把它转化为一阶方程组. 例如,考察下列 m 阶微分方程

$$\begin{cases} y^{(m)} = f(x,y,y',\cdots,y^{(m-1)}), \\ y^{(k)}(x_0) = y_0^{(k)}, \quad k = 0,1,\cdots,m-1. \end{cases} \tag{9.36}$$

只要引进新的变量

$$y_1 = y,\quad y_2 = y',\quad \cdots,\quad y_m = y^{(m-1)},$$

则可将 m 阶方程(9.36)化为如下的一阶方程组

$$\begin{cases} y_i' = y_{i+1}, & i = 1,2,\cdots,m-1, \\ y_m' = f(x,y_1,y_2,\cdots,y_m), \\ y_k(x_0) = y_0^{(k-1)}, & k = 1,2,\cdots,m. \end{cases} \tag{9.37}$$

因此,可用求解方程组形式的方法来求解(9.37).

例如,对 2 阶微分方程初值问题

$$\begin{cases} y'' = f(x,y,y'), \\ y(x_0) = y_0, \quad y'(x_0) = y_0', \end{cases}$$

令 $z = y'$,则可将该初值问题化为一阶微分方程组的初值问题

$$\begin{cases} y' = z, \\ z' = f(x,y,z), \\ y(x_0) = y_0, \quad z(x_0) = y_0'. \end{cases}$$

对此应用 Euler 方法有

$$\begin{bmatrix} y_{n+1} \\ z_{n+1} \end{bmatrix} = \begin{bmatrix} y_n \\ z_n \end{bmatrix} + h\begin{bmatrix} z_n \\ f(x_n,y_n,z_n) \end{bmatrix},$$

即得数值计算公式

$$\begin{cases} y_{n+1} = y_n + hy_n', \\ y'_{n+1} = y_n' + nf(x_n,y_n,y_n'), \quad n = 0,1,2,\cdots. \end{cases}$$

9.5.2 刚性方程组

先考虑两个简单的初值问题.

问题 1

$$\begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} 2\sin x \\ 2(\cos x - \sin x) \end{bmatrix}, \quad \begin{bmatrix} u(0) \\ v(0) \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}.$$

问题 2

$$\begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 998 & -999 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} 2\sin x \\ 999(\cos x - \sin x) \end{bmatrix}, \quad \begin{bmatrix} u(0) \\ v(0) \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}.$$

这两个问题有同样的解

$$\begin{bmatrix} u(x) \\ v(x) \end{bmatrix} = 2\mathrm{e}^{-x}\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \sin x \\ \cos x \end{bmatrix}.$$

采用 4 阶经典 R-K 方法来计算上面的两个问题,以相同的误差要求来自动选取步长,计算从 $x = 0$ 到 $x = 10$. 第一个问题可用相当大的步长,而第二个问

题能使用的步长小到难以接受. 如果改用某种低阶隐式公式,那么这两个问题均可用较大的步长,计算出大致符合要求的解来. 上述显示出来的现象称为刚性. 问题 2 是刚性的,问题 1 是非刚性的. 由于这两个问题的解是相同的,因此这种现象不是问题的解的作用,而是方程组的一种特性所引起的. 基于这个事实,较为正确的应称之为刚性方程组而不是刚性问题.

考虑方程组的通解. 对于问题 1,方程组的系数矩阵的特征值为 $\lambda_1 = -1$ 和 $\lambda_2 = -3$,其通解为

$$\begin{bmatrix} u(x) \\ v(x) \end{bmatrix} = \alpha_1 \mathrm{e}^{-x} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \alpha_2 \mathrm{e}^{-3x} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \sin x \\ \cos x \end{bmatrix}, \tag{9.38}$$

其中 α_1, α_2 为任意常数. 对于问题 2,方程组的系数矩阵的特征值为 $\lambda_1 = -1$ 和 $\lambda_2 = -1\,000$,其通解为

$$\begin{bmatrix} u(x) \\ v(x) \end{bmatrix} = \beta_1 \mathrm{e}^{-x} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \beta_2 \mathrm{e}^{-1\,000x} \begin{bmatrix} 1 \\ -998 \end{bmatrix} + \begin{bmatrix} \sin x \\ \cos x \end{bmatrix}, \tag{9.39}$$

其中 β_1, β_2 为任意常数.

数值计算中出现的现象可以用稳定性来解释. 两个问题的特征值都是实的,因此可只考虑绝对稳定区间. 经典 R-K 方法的绝对稳定区间近似为 $(-2.785, 0)$. 对于问题 1,如果 $-3h \in (-2.785, 0)$ 或 $h < 0.928$ 时,可以是稳定的. 对于问题 2,要求 $-1\,000h \in (-2.785, 0)$ 或 $h < 0.002\,785$,才能保证稳定. 由上可以看出,一定精度范围内,h 完全由绝对稳定性决定.

由通解(9.39) 可见,当 $x \to \infty$ 时,(9.39) 式右边的第一项和第二项都趋于零,这两项称为瞬态解. 趋于零的快慢取决于特征值的大小. 显然,第二项很快趋于零,此项称为快瞬态解,而第一项称为慢瞬态解. (9.39) 式右边的第三项称为稳态解. 实际计算表明,当第二项很快趋于零以后,要使整个方程组的解趋于稳态解,必须由第一项来决定计算终止与否. 因此在计算中,要在一个很长的区间上处处用小步长来计算,这就是刚性现象. 由(9.38) 式和(9.39) 式可见,计算的步数与量 $\left|\dfrac{\lambda_2}{\lambda_1}\right|$ 有关.

一般地,考虑非齐次常系数方程组

$$\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{\varphi}(x), \tag{9.40}$$

其中,$\boldsymbol{y}, \boldsymbol{\varphi} \in \mathbf{R}^m$,$\boldsymbol{A} \in \mathbf{R}^{m \times m}$ 为常数矩阵. 设 $\boldsymbol{A}$ 有不同的特征值 $\lambda_k \in \mathbf{C}$,$k = 1, 2, \cdots, m$,它们对应的特征向量 $\boldsymbol{u}_k \in \mathbf{C}^m$,$k = 1, 2, \cdots, m$,则方程组(9.40) 的通解为

$$\boldsymbol{y}(x) = \sum_{k=1}^{m} \alpha_k \mathrm{e}^{\lambda_k x} \boldsymbol{u}_k + \boldsymbol{\Psi}(x),$$

其中 $\alpha_k (k = 1, 2, \cdots, m)$ 为任意常数,$\boldsymbol{\Psi}$ 为(9.40) 式的特解.

假定特征值 λ_k 的实部为负，即

$$\mathrm{Re}(\lambda_k) < 0, \quad k = 1,2,\cdots,m.$$

当 $k \to \infty$ 时，

$$\sum_{k=1}^{m} \alpha_k \mathrm{e}^{\lambda_k x} \boldsymbol{u}_k$$

趋向于零，此项称为瞬态解，而 $\boldsymbol{\Psi}$ 则称为稳态解. 如果 $|\mathrm{Re}(\lambda_k)|$ 大，那么对应的项 $\alpha_k \mathrm{e}^{\lambda_k x}\boldsymbol{u}_k$ 当 x 增加时快速衰减，此项称为快瞬态解. 如果 $|\mathrm{Re}(\lambda_k)|$ 小，那么对应的项 $\alpha_k \mathrm{e}^{\lambda_k x}\boldsymbol{u}_k$ 当 x 增加时衰减慢，称其为慢瞬态解.

现设 $\boldsymbol{A}$ 的特征值按其实部的绝对值大小排列

$$|\mathrm{Re}(\lambda_1)| \leqslant |\mathrm{Re}(\lambda_2)| \leqslant \cdots \leqslant |\mathrm{Re}(\lambda_m)|.$$

当我们计算稳态解时，必须求到 $\alpha_1 \mathrm{e}^{\lambda_1 x}\boldsymbol{u}_1$ 可以忽略为止，所以 $|\mathrm{Re}(\lambda_1)|$ 越小，计算的区间越长. 另一方面，为使 $\lambda_k h (k = 1,2,\cdots,m)$ 均在绝对稳定性区域内，显然，$|\mathrm{Re}(\lambda_m)|$ 的值大时，必须采用很小的步长 h. 因此，引入微分方程组(9.40)的刚性比

$$S = \frac{|\mathrm{Re}(\lambda_m)|}{|\mathrm{Re}(\lambda_1)|}, \tag{9.41}$$

那么，我们似乎可以用刚性比来描述刚性方程组，即方程组(9.40)中 $\boldsymbol{A}$ 的全部特征值有负的实部并刚性比 S 是大的，那么(9.40)式是刚性的.

上述描述性定义有时也会发生一些不妥，比如，该定义未能包含实际问题中常常出现和特征值实部为小的正数或等于零的情况. 因此，我们引入下面的定义.

定义 **9.6** 当具有有限的绝对稳定区域的数值方法应用到一个任意初始条件方程组时，如果在求解区间上必须用非常小的步长，则称此方程组在该区间上是刚性的.

刚性方程组有其自身的特点，一般显式方法难于应用. 梯形方法、隐式 Euler 法对 h 不限制，可适用于一定类型的刚性方程组的求解. 这里，我们不详细讨论方程的求解.

9.6 边值问题的数值解法

在具体求解常微分方程时，必须附加某种定解条件. 定解条件通常有两种：一种是初始条件；另一种是边界条件. 与边界条件相应的定解问题称为边值问题. 本节介绍求解两点边值问题

$$\begin{cases} y'' = f(x,y,y'), \\ y(a) = \alpha, \quad y(b) = \beta, \end{cases} \tag{9.42}$$

的数值解法. 当 f 关于 y 和 y' 是线性时,(9.42) 式为线性两点边值问题

$$\begin{cases} y'' + p(x)y' + q(x)y = f(x), \\ y(a) = \alpha, \quad y(b) = \beta. \end{cases} \tag{9.43}$$

9.6.1　打靶法

打靶法的基本原理是将两点边值问题(9.42) 转化为下列形式的初值问题

$$\begin{cases} y'' = f(x, y, y'), \\ y(a) = \alpha, \quad y'(a) = s_k. \end{cases} \tag{9.44}$$

这里的 s_k 为 y 在 a 处的斜率. 令 $z = y'$,上述 2 阶方程降为一阶方程组

$$\begin{cases} y' = z, \\ z' = f(x, y, z), \\ y(a) = \alpha, \quad z(a) = s_k. \end{cases} \tag{9.45}$$

因此,边值问题变成求合适的 s_k,使上述方程组初值问题的解满足原边值问题的右端边界条件 $y(b) = \beta$,从而得到边值问题的解. 这样,把一个两点边值问题的数值解问题转化成为一阶方程组初值问题的数值解问题. 方程组初值问题的所有数值方法在这里都可以使用. 问题的关键是如何去找合适的初始斜率的试探值 s_k.

对给定的 s_k,设初值问题(9.44) 的解为 $y(x, s_k)$,它是 s_k 的隐函数. 假设 $y(x, s_k)$ 随 s_k 是连续变化的,记为 $y(x, s)$,于是我们要找的 s_k 就是方程

$$y(b, x) - \beta = 0$$

的根. 可以用方程求根的迭代法求上述方程的根. 比如用割线法有

$$s_k = s_{k-1} - \frac{s_{k-1} - s_{k-2}}{y(b, s_{k-1}) - y(b, s_{k-2})}(y(b, s_{k-1}) - \beta), \quad k = 2, 3, \cdots. \tag{9.46}$$

这样,可以按下面简单的计算过程进行求解. 先给定两个初始斜率 s_0,s_1,分别作为初值问题(9.45) 的初始条件. 用一阶方程组的数值方法求解它们,分别得到区间右端点的函数的计算值 $y(b, s_0)$ 和 $y(b, s_1)$. 如果 $|y(b, s_0) - \beta| < \varepsilon$ 或 $|y(b, s_1) - \beta| < \varepsilon$,则以 $y(x, s_0)$ 或 $y(x, s_1)$ 作为两点边值问题的解. 否则用割线法(9.46) 求 s_2,同理得到 $y(b, s_2)$,再判断它是否满足精度要求 $|y(b, s_2) - \beta| < \varepsilon$. 如此重复,直到某个 s_k 满足 $|y(b, s_k) - \beta| < \varepsilon$,此时得到的 $y(x_i)$ 和 $y_i' = z(x_i)$ 就是边值问题的解函数值和它的一阶导数值. 上述过程好比打靶,s_k 作为斜率为子弹的发射,$y(b)$ 为靶心,故称为打靶法.

值得指出的是,对于线性边值问题(9.43),一个简单又实用的方法是用解析的思想,将它转化为两个初值问题

$$\begin{cases} y_1'' + p(x)y_1' + q(x)y_1 = f(x), \\ y_1(a) = \alpha, \quad y_1'(a) = 0; \end{cases}$$

$$\begin{cases} y_2'' + p(x)y_2' + q(x)y_2 = 0, \\ y_2(a) = 0, \quad y_2'(a) = 1. \end{cases}$$

求得这两个初值问题的解 $y_1(x)$ 和 $y_2(x)$，若 $y_2(b) \neq 0$，则容易验证

$$y(x) = y_1(x) + \frac{\beta - y_1(b)}{y_2(b)} y_2(x) \tag{9.47}$$

为线性两点边值问题(9.43)的解.

例 **9.9** 用打靶法求解线性边值问题

$$\begin{cases} y'' + xy' - 4y = 12x^2 - 3x, \quad 0 < x < 1, \\ y(0) = 0, \quad y(1) = 2, \end{cases}$$

其解的解析表达式为 $y(x) = x^4 + x$.

解 先将线性边值问题转化为两个初值问题

$$\begin{cases} y_1'' + xy_1' - 4y_1 = 12x^2 - 3x, \\ y_1(0) = 0, \quad y_1'(0) = 0, \end{cases}$$

$$\begin{cases} y_2'' + xy_2' - 4y_2 = 0, \\ y_2(0) = 0, \quad y_2'(0) = 1. \end{cases}$$

令 $z_1 = y_1'$，$z_2 = y_2'$，将上述两个初值问题分别降为一阶方程组初值问题

$$\begin{cases} y_1' = z_1, \\ z_1' = -xz_1 + 4y_1 + 12x^2 - 3x, \\ y_1(0) = 0, \quad z_1(0) = 0, \end{cases}$$

$$\begin{cases} y_2' = z_2, \\ z_2' = -xz_2 + 4y_2, \\ y_2(0) = 0, \quad z_2(0) = 1. \end{cases}$$

取 $h = 0.02$，用经典 R-K 法分别求这两个方程组的解 $y_1(x)$ 和 $y_2(x)$ 的计算值 y_{1i} 和 y_{2i}，然后按(9.47)式得精确解

$$y(x) = y_1(x) + \frac{2 - y_1(1)}{y_2(1)} y_2(x)$$

的打靶法计算值 y_i，部分点上的计算值、精确值和误差列于表 9－12.

表 9－12

x_i	y_{1i}	y_{2i}	y_i	$y(x_i)$	$\lvert y(x_i) - y_i \rvert$
0	0	0	0	0	0
0.2	−0.002 407 991	0.204 007 989	0.201 600 005 3	0.201 600 00 0	0.53×10^{-8}
0.4	−0.006 655 031	0.432 255 024	0.425 600 008 0	0.425 600 00 0	0.80×10^{-8}
0.6	0.019 672 413	0.709 927 571	0.729 600 008 3	0.729 600 000	0.83×10^{-8}
0.8	0.145 529 585	1.064 070 385	1.209 600 005 8	1.209 600 000	0.58×10^{-8}
1.0	0.475 570 149	1.524 428 455	2.000 000 000	2.000 000 000	0

例 9.10　用打靶法求解非线性边值问题

$$\begin{cases} 4y'' + yy' = 2x^3 + 16, \\ y(2) = 8, \quad y(3) = \dfrac{35}{3}. \end{cases}$$

要求误差不超过 0.5×10^{-6}，其解析解是 $y(x) = x^2 + \dfrac{8}{x}$.

解　对应于(9.45)式的初值问题为

$$\begin{cases} y' = z, \\ z' = -\dfrac{yz}{4} + \dfrac{x^3}{2} + 4, \\ y(2) = 8, \quad z(2) = s_k. \end{cases}$$

对于每一个 s_k，取 $h = 0.02$，用经典 R-K 法求解. 初选 $s_0 = 1.5$，求得 $y(3, s_0) = 11.488\,9$，则有 $|y(3, s_0) - y(3)| = 0.177\,7 > 0.5 \times 10^{-6}$. 再选 $s_1 = 2.5$，求得 $y(3, s_1) = 11.842\,1$，则有 $|y(3, s_1) - y(3)| = 0.075\,5 > 0.5 \times 10^{-6}$. 以 s_0, s_1 作为割线法迭代初值，由割线法计算

$$s_2 = s_1 - \frac{s_1 - s_0}{y(3, s_1) - y(3, s_0)}(y(3, s_1) - y(3)) = 2.003\,224\,1.$$

由此得 $y(3, s_2) = 11.667\,8$，仍然不满足精度要求. 由 $s_1, s_2, y(3, s_1)$ 和 $y(3, s_2)$，用割线法得到 $s_3 = 1.999\,979$. 重复这个过程，直到 $s_4 = 2.000\,000$，再求解相应的初值问题，得到 $y(3, s_4) = 11.666\,666\,669$，有 $|y(3, s_4) - y(3)| < 0.5 \times 10^{-6}$. 于是得到边值问题的解 y_i，打靶过程和边值问题的计算解分别列于表 9－13 和表9－14.

表 9－13

s_k	1.5	2.5	2.003 224	1.999 979	2.000 000
$y(3, s_k)$	11.488 914	11.842 141	11.667 805	11.666 659	11.666 667

表 9－14

x_i	y_i	$y(x_i)$	$\lvert y(x_i) - y_i \rvert$
2.0	8	8	0
2.2	8.476 363 637 8	8.476 363 636 4	0.13×10^{-8}
2.4	9.093 333 352	9.093 333 333	0.18×10^{-8}
2.6	9.836 923 078 5	9.836 923 076 9	0.16×10^{-8}
2.8	10.697 142 656 2	10.697 142 857 1	0.10×10^{-8}
3.0	11.666 666 669	11.666 666 666 7	0.30×10^{-9}

计算结果表明打靶法的效果是很好的，计算精度取决于所选取的初值问题

数值方法的阶和所选取的步长 h 的大小. 不过打靶法过分依赖于经验,选取试射值,有一定的局限性.

9.6.2 差分法

差分法是解边值问题的一种基本方法,它利用差商代替导数,将微分方程离散化为线性或非线性方程组(即差分方程) 来求解.

先考虑线性边值问题(9.43) 的差分法. 将区间$[a,b]$ 分成 n 等分,子区间的长度 $h=\dfrac{b-a}{n}$,分点 $x_k=a+kh(k=0,1,\cdots,n)$. 由

$$y'(x_k)=\frac{y(x_{k+1})-y(x_{k-1})}{2h}+O(h^2),$$

$$y''(x_k)=\frac{y(x_{k+1})-2y(x_{k-1})+2(x_{k-1})}{h^2}+O(h^2),$$

忽略余项,将差商分别代替(9.43) 式中节点 x_k 处的一阶和二阶导数,实现离散化. 设 $p_k=p(x_k)$,$q_k=q(x_k)$,$f_k=f(x_k)$,用 y_k 近似表示 $y(x_k)$,建立差分方程

$$\frac{y_{k+1}-2y_k+y_{k-1}}{h^2}+p_k\frac{y_{k+1}-y_{k-1}}{2h}-q_ky_k=f_k,\quad k=1,2,\cdots,n-1.$$

整理后得到关于 y_k 的线性方程组

$$(2-hp_k)y_{k-1}+(2h^2q_k-4)y_k+(2+hp_k)y_{k+1}=2h^2f_k,\ k=1,2,\cdots,n-1 \tag{9.48}$$

利用边界条件 $y_0=\alpha$,$y_n=\beta$,将它们分别代入上面 $k=1$ 和 $k=n-1$ 的两个方程中,整理后得到关于 $y_1,y_2,\cdots,y_{n-1}$ 的方程组

$$\begin{cases}(2h^2q_1-4)y_1+(2+hp_1)y_2=2h^2f_1-(2-hp_1)\alpha,\\ (2-hp_k)y_{k-1}+(2h^2q_k-4)y_k+(2+hp_k)y_{k+1}=2h^2f_k,\\ \qquad k=2,3,\cdots,n-2,\\ (2-hp_{n-1})y_{n-2}+(2h^2q_{n-1}-4)y_{n-1}=2h^2f_{n-1}-(2+hp_{n-1})\beta.\end{cases} \tag{9.49}$$

这是一个三对角方程组.

若 $q(x)\leqslant 0$,$x\in[a,b]$,且步长满足 $|hp_k|<2$,则方程组(9.49) 的系数矩阵是严格对角占优的. 此时,方程组(9.49) 的解存在唯一,用追赶法求解此方程组时一定是数值稳定的,用 Jacobi 迭代法求解此方程组时一定是收敛的.

在应用上,有时边界条件按以下方式给出

$$y'(a)=\alpha_0y(a)+\beta_0,\quad y'(b)=\alpha_1y(b)+\beta_1.$$

这里 $\alpha_0,\beta_0,\alpha_1,\beta_1$ 均为已知常数. 这时,边界条件中所包含的导数也要替换成相应的差商

$$\frac{y_1-y_0}{h}=\alpha_0 y_0+\beta_0,\quad \frac{y_n-y_{n-1}}{h}=\alpha_1 y_n+\beta_1.$$

它们和差分方程(9.48)一起,仍然构成包含 $n+1$ 个未知数的线性方程组.

例 9.11　取不同的步长 h,用差分方法求解线性边值问题:

$$\begin{cases} y''-y'=-2\sin x, \\ y(0)=-1,\quad y\left(\dfrac{\pi}{2}\right)=1. \end{cases}$$

其解析解是 $y(x)=\sin x-\cos x$.

解　本题的 $p(x)=-1, q(x)=0$. 取 $n=4, h=\dfrac{\pi}{8}$,由方程组(9.49)得

$$\begin{pmatrix} -4 & 1.607\,3 & 0 \\ 2.392\,7 & -4 & 1.607\,3 \\ 0 & 2.392\,7 & -4 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}=-\begin{pmatrix} 2.156\,6 \\ -0.436\,2 \\ -2.177\,2 \end{pmatrix}.$$

解此方程组得 $y_1=-0.535\,1, y_2=0.010\,1, y_3=0.550\,3$. 而解析解是 $y(x_1)=-0.541\,2, y(x_2)=0, y(x_3)=0.541\,2$. 由于节点少,步长太大,所以计算精度差.

一般地,对应于方程组(9.49)有

$$\begin{pmatrix} -4 & 2-h & & & \\ 2+h & -4 & 2-h & & \\ & \ddots & \ddots & \ddots & \\ & & 2+h & -4 & 2-h \\ & & & 2+h & -4 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-2} \\ y_{n-1} \end{pmatrix}=\begin{pmatrix} 2+h-4h^2\sin x_1 \\ -4h^2\sin x_2 \\ \vdots \\ -4h^2\sin x_{n-2} \\ h-2-4h^2\sin x_{n-1} \end{pmatrix};$$

其中,$h=\dfrac{\pi}{2n}, x_k=kh, k=1,2,\cdots,n-1$. 当 $n=10$ 时,$h=\dfrac{\pi}{20}$,部分计算结果列于表 9-15. 当 $n=20$ 时,近似解误差的最大绝对值不超过 0.41×10^{-3}. 当 $n=500$ 时,误差的最大绝对值不超过 0.65×10^{-6}. 因此,随着节点数的增加,精度提高.

表 9-15

x_k	y_k	$y(x_k)$	$y(x_k)-y_k$
0	−1.000 0	−1.000 0	0
$2\dfrac{\pi}{20}$	−0.641 3	−0.642 0	-0.8×10^{-3}
$4\dfrac{\pi}{20}$	−0.219 8	−0.221 2	-0.14×10^{-2}
$6\dfrac{\pi}{20}$	0.222 9	0.221 2	-0.16×10^{-2}
$8\dfrac{\pi}{20}$	0.643 3	0.642 0	-0.13×10^{-2}
$\dfrac{\pi}{2}$	1.000 0	1.000 0	0

对于非线性两点边值问题，其离散化后所得到的方程组是非线性的. 下面说明用有限差分法求非线性边值问题(9.42)的数值解. 区间划分与离散化方法同线性情形，得到在 x_k 处的差分方程

$$\frac{y_{k+1}-2y_k+y_{k-1}}{h^2}=f\left(x_k,y_k,\frac{y_{k+1}-y_{k-1}}{2h}\right),\quad k=1,2,\cdots,n-1.$$

利用边界条件 $y_0=\alpha,y_n=\beta$，将它们分别代入 $k=1$ 和 $k=n-1$ 的两个方程中，并将已知量移到方程的右边，得到关于 $y_1,y_2,\cdots,y_{n-1}$ 的非线性方程组

$$\begin{cases}2y_1-y_2+h^2f\left(x_1,y_1\ \dfrac{y_2-\alpha}{2h}\right)=\alpha,\\ -y_{k-1}+2y_k-y_{k+1}+h^2f\left(x_k,y_k,\dfrac{y_{k+1}-y_{k-1}}{2h}\right)=0,\\ \qquad k=2,3,\cdots,n-2,\\ -y_{n-2}+2y_{n-1}+h^2f\left(x_{n-1},y_{n-1},\dfrac{\beta-y_{n-2}}{2h}\right)=\beta.\end{cases}$$

可以用非线性方程组迭代法解此方程组.

9.6.3 差分问题的收敛性

我们知道，通过自变量的适当变换可消除线性方程(9.43)中的一阶导数项. 因此，下面仅就缺一阶导数项的方程来讨论，即考察边值问题.

$$\begin{cases}y''+q(x)y=f(x),\\ y(a)=\alpha,y(b)=\beta.\end{cases}\tag{9.50}$$

这里假定 $q(x)\leqslant 0$，对应于(9.50)式的差分问题是

$$\begin{cases}\dfrac{y_{k+1}-2y_k+y_{k-1}}{h^2}+q_ky_k=f_k,\quad k=1,2,\cdots,n-1,\\ y_0=\alpha,\quad y_n=\beta.\end{cases}\tag{9.51}$$

为了研究差分问题的收敛性，我们先介绍下述极值原理.

定理 9.2 对于一组不全相等的数 $y_k,k=0,1,\cdots,n$，令

$$l(y_k)=\frac{y_{k+1}-2y_k+y_{k-1}}{h^2}+q_ky_k,$$

其中 $q_k\leqslant 0,k=1,2,\cdots,n-1$. 如果 $l(y_k)\geqslant 0(k=1,2,\cdots,n-1)$，则 y_k 的正的最大值只能是 y_0 或 y_n. 如果 $l(y_k)\leqslant 0(k=1,2,\cdots,n-1)$，则 y_k 的负的最小值只能是 y_0 或 y_n.

证 用反证法，考虑 $l(y_k)\geqslant 0$ 的情形. 假设 $y_m(0<m<n)$ 是正的最大值，即

$$y_m = \max_{0\leqslant k\leqslant n} y_k = M > 0,$$

且 y_{m-1} 和 y_{m+1} 中至少有一个小于 M，此时有

$$\begin{aligned} l(y_m) &= \frac{y_{m+1} - 2M + y_{m-1}}{h^2} + q_m M \\ &< \frac{M - 2M + M}{h^2} + q_m M = q_m M. \end{aligned}$$

由于 $q_m \leqslant 0, M > 0$，故由上式推出 $l(y_m) < 0$，此与原设矛盾. 此外，$l(y_k) \leqslant 0$ 的情形可类似地讨论. 定理得证.

由极值原理容易证明以下结论.

定理 9.3　差分问题(9.51)的解存在且是唯一的.

证　只要证明对应的齐次方程组

$$\begin{cases} l(y_k) = \dfrac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + q_k y_k = 0, \quad k = 1,2,\cdots,n-1, \\ y_0 = y_n = 0, \end{cases}$$

只有零解. 由于这里 $l(y_k) = 0 (k = 1,2,\cdots,n-1)$，由极值原理知，$y_k$ 的正的最大值和负的最小值只能是 y_0 和 y_n，按边界条件知 $y_k = 0 (k = 0,1,\cdots,n)$ 定理得证.

下面利用极值原理证明差分方法的收敛性定理.

定理 9.4　设 y_k 是差分问题(9.51)的解，而 $y(x_k)$ 是边值问题(9.50)的解 $y(x)$ 在节点 x_k 的值. 若 $y(x) \in c^4[a,b]$，则截断误差 $e_k = y(x_k) - y_k$ 有下列估计式

$$|e_k| \leqslant \frac{h^2}{24} M (x_k - a)(b - x_k), \quad k = 0,1,\cdots,n,$$

其中 $M = \max\limits_{a\leqslant x\leqslant b} |y^{(4)}(x)|$.

证　对 $k = 0, n$，结论显然成立. 对 $k = 1,2,\cdots,n-1$，由 Taylor 展开式易得

$$\frac{y(x_{k+1}) - 2y(x_k) + y(x_{k-1})}{h^2} + q_k y(x_k) = f(x_k) + \frac{h^2}{12} y^{(4)}(\xi_k), \quad (9.52)$$

其中 $x_{k-1} < \xi_k < x_{k+1}$. 将(9.52)式与(9.51)式相减得

$$\begin{cases} l(e_k) = \dfrac{e_{k+1} - 2e_k + e_{k-1}}{h^2} + q_k e_k = \dfrac{h^2}{12} y^{(4)}(\xi_k), \quad k = 1,2,\cdots,n-1, \\ e_0 = e_n = 0. \end{cases}$$

因为上式中的 ξ_k 未知，我们考虑以下差分问题

$$\begin{cases} l(\varepsilon_k) = \dfrac{\varepsilon_{k+1} - 2\varepsilon_k + \varepsilon_{k-1}}{h^2} + q_k \varepsilon_k = -\dfrac{h^2}{12} M, \quad k = 1,2,\cdots,n-1, \\ \varepsilon_0 = \varepsilon_n = 0, \end{cases} \quad (9.53)$$

由于

$$l(\varepsilon_k) = -\frac{h^2}{12}M \leqslant -\frac{h^2}{12}\,|\,y^{(4)}(\xi_k)\,| = -|\,l(e_k)\,|,$$

故有

$$l(\varepsilon_k - e_k) \leqslant 0, \quad l(\varepsilon_k + e_k) \leqslant 0.$$

又因 $\varepsilon_0 - e_0 = \varepsilon_n - e_n = 0$，$\varepsilon_0 + e_0 = \varepsilon_n + e_n = 0$，利用极值原理知 $\varepsilon_k - e_k \geqslant 0$，$\varepsilon_k + e_k \geqslant 0$，即得 $|\,e_k\,| \leqslant \varepsilon_k$，$k = 0,1,\cdots,n$.

差分问题(9.53)的解仍然难以求出，我们进一步考虑如下差分问题

$$\begin{cases} \bar{l}(\rho_k) = \dfrac{\rho_{k+1} - 2\rho_k + \rho_{k-1}}{h^2} = -\dfrac{h^2}{12}M, \quad k = 1,2,\cdots,n-1, \\ \rho_0 = \rho_n = 0, \end{cases} \tag{9.54}$$

这里由于

$$\bar{l}(\rho_k - \varepsilon_k) = q_k\varepsilon_k \leqslant 0, \quad \rho_0 - \varepsilon_0 = \rho_n - \varepsilon_n = 0,$$

由极值原理知 $\rho_k - \varepsilon_k \geqslant 0$，即 $\varepsilon_k \leqslant \rho_k$. 于是有

$$|\,e_k\,| \leqslant \varepsilon_k \leqslant \rho_k. \tag{9.55}$$

显然，差分问题(9.54)对应如下边值问题

$$\begin{cases} \rho'' = -\dfrac{h^2}{12}M, \\ \rho(x_0) = \rho(x_n) = 0, \end{cases}$$

其解为

$$\rho(x) = \frac{h^2}{24}M(x-a)(b-x).$$

由此与(9.55)式即得要证明的结论. 定理论证.

注意到 $\rho(x)$ 在 $x = \dfrac{a+b}{2}$ 处达到最大值，因此有估计式

$$|\,e_k\,| \leqslant \frac{M(b-a)^2}{96}h^2, \quad k = 0,1,\cdots,n.$$

误差估计式说明，当 $h \to 0$ 时，差分方程的解收敛到微分方程边值问题的解.

就差分方法而言，常微分方程的数值解法与偏微分方程的数值解法有着紧密的联系，有关内容可进一步参看偏微分方程数值解法的有关文献.

评　　注

本章介绍常微分方程初值问题和边值问题的基本数值解法. 初值问题的数值解法主要是单步法和线性多步法. 构造方法的主要途径是基于 Taylor 展开和

数值积分. 基于 Taylor 展开的方法灵活，具有一般性，它在构造差分公式的同时可以得到关于截断误差的估计.

4 阶 R-K 法是常用的算法，其优点是精度高，程序简单，计算过程稳定，并且容易调节步长. 但是，它要求斜率函数具有较高的光滑性，否则，它的精度还不如 Euler 方法或改进的 Euler 方法. 此外，4 阶 R-K 法的计算量较大，它计算斜率函数值的次数多. 如果斜率函数较复杂，宜用线性多步法和预估 — 校正方法，它们计算函数值次数少. 例如 Hamming 公式和 4 阶的 Admas 预估 — 校正公式，这种情况下还要用单步法提供所需的开始值.

步长的选取是很重要的问题，既要考虑节省计算量，步长不能太小，又要保证结果的精度，步长不能太大. 由于常微分方程初值问题的求解是一个逐步计算的过程，任何一步产生的误差都会对以后的计算产生影响，所以最好采用绝对稳定性较好的方法，并经常估计误差. 隐式方法求解麻烦，但绝对稳定性好，所以仍然常用，尤其是在刚性问题中常用.

边值问题是另一类常微分方程定解问题，有很多实际应用背景. 这类问题比初值问题复杂得多，通常要满足一定条件才存在唯一解. 本章只介绍了打靶法和差分方法. 打靶法将边值问题化为初值问题. 差分方法通过离散化将问题化为线性方程组的求解问题，这也是偏微分方程数值解的主要方法.

习　题　9

9.1　用 Euler 法计算积分

$$\int_0^x \mathrm{e}^{t^2}\,\mathrm{d}t$$

在点 $x=0.5,1.5,2$ 处的数值解.

9.2　用改进的 Euler 方法解初值问题

$$\begin{cases} y' = x+y, & 0<x\leqslant 1, \\ y(0)=1. \end{cases}$$

取步长 $h=0.1$ 计算，并与准确解 $y=2\mathrm{e}^x-x-1$ 相比较.

9.3　用改进的 Euler 方法解初值问题

$$\begin{cases} y' = x^2+x-y, \\ y(0)=0. \end{cases}$$

取步长 $h=0.1$，计算 $y(0.5)$，并与准确解 $y=x^2-x+1-\mathrm{e}^{-x}$ 相比较.

9.4　对初值问题

$$\begin{cases} y' = -y, \\ y(0)=1, \end{cases}$$

证明 Euler 公式和梯形公式求得的近似解分别为

$$y_n=(1-h)^n,\quad y_n=\left(\frac{2-h}{2+h}\right)^n.$$

并证明当 $h\to 0$ 时,它们都收敛于准确解 $y(x)=\mathrm{e}^{-x}$.

9.5 取 $h=0.2$,用经典 R-K 法求解下列初值问题

(1) $\begin{cases}y'=x+y, & 0<x\leqslant 1,\\ y(0)=1;\end{cases}$

(2) $\begin{cases}y'=\dfrac{3y}{1+x}, & 0<x\leqslant 1,\\ y(0)=1.\end{cases}$

9.6 证明对任意参数 t,下列 R-K 公式是 2 阶的

$$\begin{aligned}y_{n+1}&=y_n+\frac{h}{2}(K_2+K_3),\\ K_1&=f(x_n,y_n),\\ K_2&=f(x_n+th,y_n+thK_1),\\ K_3&=f(x_n+(1-t)h,y_n+(1-t)hK_1).\end{aligned}$$

9.7 对试验方程 $y'=\lambda y(\lambda<0)$,试证明如下方法给出的绝对稳定条件:

(1) 改进的 Euler 公式:$\left|1+\lambda h+\dfrac{(\lambda h)^2}{2}\right|\leqslant 1$;

(2) 经典 R-K 公式:$\left|1+\lambda h+\dfrac{(\lambda h)^2}{2}+\dfrac{(\lambda h)^3}{6}+\dfrac{(\lambda h)^4}{24}\right|\leqslant 1.$

9.8 分别用 2 阶显式 Adams 方法和 2 阶隐式 Admas 方法解下列初值问题

$$\begin{cases}y'=1-y,\\ y(0)=0.\end{cases}$$

取 $h=0.2$,$y_0=0$,$y_1=0.181$,计算 $y(1)$,并与准确解 $y(x)=1-\mathrm{e}^{-x}$ 相比较.

9.9 利用经典 R-K 法提供初始值,取 $h=0.1$,分别用 4 阶显式 Adams 公式和 Adams 预估 — 校正公式求解初值问题.

$$\begin{cases}y'=x^2-y^2, & -1\leqslant x\leqslant -0,4,\\ y(-1)=0.\end{cases}$$

9.10 证明求解初值问题 $y'=f(x,y),y(x_0)=y_0$ 的差分公式

$$y_{n+1}=\frac{1}{2}(y_n+y_{n-1})+\frac{h}{4}(4f_{n+1}-f_n+3f_{n-1})$$

是 2 阶的,并求出其局部截断误差的主项.

9.11 设有初值问题 $y'=f(x,y),y(x_0)=y_0$,用 Taylor 展开定理构造形如

$$y_{n+1}=\alpha(y_n+y_{n-1})+h(\beta_0 f_n+\beta_1 f_{n-1})$$

的两步法，试确定系数 α,β_0 和 β_1，使它具有二阶精度，并推导其局部截断误差的主项.

9.12　设有初值问题 $y'=f(x,y),y(x_0)=y_0$，用 Taylor 展开定理构造形如

$$y_{n+1}=\alpha_0 y_n+\alpha_1 y_{n-1}+h\beta_1 f_{n-1}$$

的差分公式，试确定系数 α_0,α_1 和 β_1，使它具有尽可能高的精度，并求出其局部截断误差.

9.13　取 $h=0.1$，试用经典 R-K 法求解方程组

$$\begin{cases}y_1'=3y_1+2y_2, & y_1(0)=0,\\ y_2'=4y_1+y_2, & y_2(0)=1, \quad 0<x\leqslant 0.2.\end{cases}$$

9.14　取 $h=0.1$，试用 Euler 法求解初值问题

$$\begin{cases}y''+4xyy'+2y^2=0, & 0<x\leqslant 0.2,\\ y(0)=1, & y'(0)=0.\end{cases}$$

9.15　将下列方程化为一阶方程组，并判断它们是否为刚性方程组：

(1) $y''+3y'+2y=\sin x$, $y(0)=\alpha$, $y'(0)=\beta$;

(2) $y''+16y'+15y=\sin(2x+1)$, $y(0)=\alpha$, $y'(0)=\beta$;

(3) $y'''+4y''+5y'+2y=0$, $y(0)=0$, $y'(0)=1$, $y''(0)=0$.

9.16　取 $h=0.5$，用差分法解边值问题

$$\begin{cases}y''=(1+x^2)y,\\ y(-1)=y(1)=1.\end{cases}$$

9.17　取 $h=0.2$，用差分法解边值问题

$$\begin{cases}(1+x^2)y''-xy'-3y=6x-3,\\ y(0)-y'(0)=1, \quad y(1)=2.\end{cases}$$

数值试验题 9

9.1　取 $h=0.01$，用你熟悉的数值解法求解下列初值问题

$$\begin{cases}y'=\sqrt{x^2+y^2}, & 0\leqslant x\leqslant 1.0,\\ y(0)=-1.\end{cases}$$

要求写出部分节点上的函数值.

9.2　设常微分方程初值问题

$$\begin{cases}y'=\alpha y-\alpha x+1, & 0<x\leqslant 1,\\ y(0)=1.\end{cases}$$

其中，$-50\leqslant\alpha\leqslant 50$，其准确解为 $y(x)=e^{\alpha x}+x$.

(1) 取步长 $h=0.01$，对参数 α 分别取 4 个不同的数值：一个大的正值，一个

小的正值，一个绝对值小的负值和一个绝对值大的负值，分别用经典 R-K 法计算，将计算结果画在同一张图上，比较说明相应初值问题的性态；

(2) 取 α 为一个绝对值不大的负值，对 h 取两个不同的数值：一个 h 在经典 R-K 法的稳定域内；另一个在稳定域外，分别用经典 R-K 法计算. 取全域等距的 10 个点上的计算值，列表说明.

9.3 常微分方程初值问题

$$\begin{cases} y' = -y + \cos 2x - 4\sin 2x + 2x\mathrm{e}^{-x}, & 0 < x \leqslant 2, \\ y(0) = 1, \end{cases}$$

有准确解 $y(x) = x^2\mathrm{e}^{-x} + \cos 2x$，选择一个步长 h，使四阶 Adams 预估—校正修正法和经典 R-K 法均稳定. 分别用这两种方法求解初值问题，以表格形式列出 10 个等距节点上的计算值和准确值，并比较计算精度. 运算时，取足以表示计算精度的有效值，多步法所需要的开始值由经典 R-K 法提供.

9.4 (Lorenz 问题与混沌)考虑著名的 Lorenz 方程

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = \alpha(y - x), \\ \dfrac{\mathrm{d}y}{\mathrm{d}t} = \beta x - y - xz, \\ \dfrac{\mathrm{d}z}{\mathrm{d}t} = xy - \gamma z, \end{cases}$$

其中 α, β, γ 为变化区域有一定限制的实参数. 该方程组形式简单，表面上看并无惊人之处，但由该方程揭示出的许多现象，促使“混沌”成为数学研究的崭新领域，在实际应用中也产生了巨大的影响.

选取适当的参数值 α, β, γ，再选取不同的初值，用你熟悉的数值解法求解 Lorenz 方程. 观察计算结果有什么特点？解的曲线是否有界？解的曲线是不是周期的或趋于某个固定的点？

9.5 考虑一个简单的边值问题

$$\begin{cases} y''(x) + y(x) = 6x + x^3, & 0 < x < \pi, \\ y(0) = 0, \quad y(\pi) = \pi^3. \end{cases}$$

(1) 验证上述边值问题的解为 $y(x) = \sin x + x^3$，并画出解析解的图形；

(2) 对给定的步长 h，用差分法离散化微分方程后，讨论如何选择求解差分方程的算法，并比较不同算法的效率；

(3) 选择不同的步长，求解差分方程便得到边值问题的近似解. 比较不同的步长所得数值解逼近原微分方程解的精确程度，分析解的精度与步长的关系.

习 题 答 案

习题 1

1.1 (1),(3) 2 位,0.67%; (2),(4) 4 位,0.010%.

1.2 (1) 57.563,1.8×10^{-5}; (2) 5.5×10^{-4},0.016.

1.3 $n\delta$.

1.4 δ.

1.5 0.333%.

1.6 $1-\cos2^\circ\approx6\times10^{-4}$, $\dfrac{\sin2^\circ}{1+\cos2^\circ}\approx6.093\times10^{-4}$,

$2\sin^2 1^\circ\approx6.090\times10^{-4}$,精确值 $1-\cos2^\circ=6.091\,7\times10^{-4}$.

1.7 0.5×10^{-3}.

1.8 $x_1=-28-\sqrt{783}\approx-55.982$, $x_2=x_1^{-1}=-0.017\,863$.

1.9 边长误差<0.005 cm.

1.10 0.5×10^{8},不稳定.

1.11 $I_0^*=0.632\,1$, $I_n^*=1-nI_{n-1}^*$, $E_n=I_n-I_n^*$, $E_n=(-1)^n n!\,E_0$.

1.12 $f(x)=\begin{cases}\ln(x+\sqrt{x^2+1}), & x\geqslant0,\\ -\ln(\sqrt{x^2+1}-x), & x<0.\end{cases}$

$f(30)=4.094\,62$, $f(-30)=-4.094\,62$.

1.13 (4) $\dfrac{e^{2x}-1}{2}=e^x\,\dfrac{e^x-e^{-x}}{2}$.

1.17 $\|\boldsymbol{A}\|_\infty=1.1$, $\|\boldsymbol{A}\|_1=0.8$, $\|\boldsymbol{A}\|_2=0.825$, $\|\boldsymbol{A}\|_F=0.842\,6$.

习题 2

2.1 $L_2(x)=\dfrac{5}{6}x^2+\dfrac{3}{2}x-\dfrac{7}{3}$.

2.2 $L_1(0.54)=-0.620\,219$, $L_2(0.54)=-0.618\,838$.

2.5 $L_2(x)=1.935\,7-9.291\,4x+10.074\,0x^2$, $|R_2(1.03)|\leqslant1.19\times10^{-4}$.

2.6 $L_3(x)=x^4-(x+1)x(x-1)(x-2)=2x^3+x^2-2x$.

2.7 $N_3(x)=x^3-x^2+0.25x+1$.

2.8 $f[2^\circ,2^1,\cdots,2^7]=1$, $f[2^\circ,2^1,\cdots,2^8]=0$.

2.11 $h\leqslant 0.006$.

2.12 $p(x)=\dfrac{x^2(x-3)^2}{4}$.

2.13 $H(x)=x^3-4x^2+4x$.

2.14 $|R_1(x)|\leqslant\dfrac{h^2}{4}$.

2.15 $|R_3(x)|\leqslant\dfrac{h^4}{16}$.

2.16 $S(x)=\begin{cases}\dfrac{1}{3}x^3-2x, & x\in[-1,0],\\ 3x^3-2x, & x\in[0,1].\end{cases}$

习题 3

3.1 $\varphi_0(x)=1$, $\varphi_1(x)=x-\dfrac{2}{5}$, $\varphi_2(x)=\left(x-\dfrac{4}{115}\right)\left(x-\dfrac{7}{5}\right)-\dfrac{46}{25}$.

3.2 $\varphi(x)=0.117\,187\,5+1.640\,625x^2-0.823\,125x^4$.

3.3 $\alpha=\dfrac{1}{\pi}\left(8-\dfrac{24}{\pi}\right)=0.114\,771$, $\beta=\dfrac{8}{\pi^2}\left(\dfrac{12}{\pi}-3\right)=0.664\,439$.

3.4 $\varphi(x)=\dfrac{14}{15}+\dfrac{12}{15}x$.

3.5 $\varphi_1(x)=1.266\,066+1.130\,318x$,

$\varphi_3(x)=0.994\,571+0.997\,308x+0.542\,99x^2+0.177\,347x^2$.

3.6 $p_1(x)=0.955+0.414\,2x$, $E=0.045$.

3.7 $S=22.253\,8t-7.855\,05$.

3.8 $\varphi(x)=0.999\,8x+4.000\,1\mathrm{e}^{-x}$, $\|\delta\|_2^2=6.824\,9\times10^{-6}$.

3.9 $\varphi(x)=\dfrac{t}{0.078\,9t+0.164\,9}$, $\|\delta\|_2^2=1.025\,2$.

习题 4

4.1 (1) $A_0=A_2=\dfrac{h}{3}$, $A_1=\dfrac{4h}{3}$, 有 3 次代数精度;

(2) $A_0=A_2=\dfrac{8h}{3}$, $A_1=-\dfrac{4h}{3}$, 有 3 次代数精度;

(3) $A_0=\dfrac{2}{3}$, $A_1=\dfrac{1}{3}$, $A_2=\dfrac{1}{6}$, 有 2 次代数精度.

4.3 $S=0.632\,33$, 误差限为 $0.000\,35$.

4.5 复化梯形公式需 516 个节点, 复化 Simpson 公式需 9 个节点.

4.7 0.713 27.

4.8 3.141 067 9.

4.9 $\frac{\pi}{2}$.

4.10 一阶导数值分别为-0.247,-0.217,-0.189.

4.11 $S'(101.5)=0.049\,629\,166$, $S''(101.5)=-0.000\,244\,478$.

习题 5

5.1 $\boldsymbol{x}=(0.678\,7,-0.642\,9,1.107\,1)^{\mathrm{T}}$,

$$\boldsymbol{L}\begin{pmatrix} 1 & & \\ 0.285\,7 & 1 & \\ -0.142\,9 & 0.307\,7 & 1 \end{pmatrix},\quad \boldsymbol{U}=\begin{pmatrix} 7 & 1 & -1 \\ & 3.714\,3 & 2.285\,8 \\ & & 2.153\,8 \end{pmatrix}.$$

5.2 用 Gauss 消去法得 $\widetilde{x}=(1.335,0,-5.003)^{\mathrm{T}}$.

用列主元素消去法得 $\widetilde{x}=(0.225\,2,0.279\,0,0.329\,5)^{\mathrm{T}}$.

5.6 $\boldsymbol{x}=(75,-46,-3)^{\mathrm{T}}$.

5.7 $\boldsymbol{A}$ 不可分,$\boldsymbol{B}$ 可分但不唯一,$\boldsymbol{C}$ 可分且唯一.

5.8 $\boldsymbol{L}$ 的次对角元:$-\frac{1}{2},-\frac{2}{3},-\frac{3}{4},-\frac{4}{5}$,

$\boldsymbol{U}$ 的对角元:$2,\frac{3}{2},\frac{4}{3},\frac{5}{4},\frac{6}{5}$,

$\boldsymbol{y}=\left(1,\frac{1}{2},\frac{1}{3},\frac{1}{4},\frac{1}{5}\right)^{\mathrm{T}}$, $\boldsymbol{x}=\left(\frac{5}{6},\frac{2}{3},\frac{1}{2},\frac{1}{3},\frac{1}{6}\right)^{\mathrm{T}}$.

5.9 $\boldsymbol{x}=(-0.047\,0,0.217\,4,0.212\,7)^{\mathrm{T}}$.

5.10 $\boldsymbol{x}=(1.111\,11,0.777\,78,2.555\,56)^{\mathrm{T}}$.

5.11 $\mathrm{cond}_2(\boldsymbol{A})=39\,601$, $\mathrm{cond}_\infty(\boldsymbol{A})=39\,206$.

5.14 $\frac{\|\delta\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty}\leqslant 600\%$.

5.15 $\frac{\|\delta\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty}\leqslant 67\%$.

习题 6

6.2 $\frac{|a_{12}a_{21}|}{|a_{11}a_{22}|}<1$.

6.3 精确解 $\boldsymbol{x}^*=(0.224\,9,0.305\,6,-493\,9)^{\mathrm{T}}$.

6.6 $\omega=0.9$ 时,迭代 8 次可达到精度要求,$\boldsymbol{x}^{(0)}=0$,

$\boldsymbol{x}^{(8)}=(-4.000\,027,0.299\,998\,9,0.200\,003)^{\mathrm{T}}$.

6.7 $\omega=1.03$ 时,迭代 5 次;$\omega=1$ 时,迭代 6 次;$\omega=1.1$ 时,迭代 6 次.

6.9 (1) $\rho(\boldsymbol{A}^{-1}\boldsymbol{B})<1$; (2) $\rho[(\boldsymbol{A}^{-1}\boldsymbol{B})^2]<1$.

6.10 精确解 $\boldsymbol{x}=(2,2,2,2)^{\mathrm{T}}$.

习题 7

7.1 0.905 46.

7.2 二分 14 次.

7.3 $x^*\approx 4.493\,42$.

7.4 (1) $\varphi(x)=\cos x$, $[a,b]=[0,1]$, $x^*=0.739\,1$;

(2) 在$[-1,0]$上取 $\varphi(x)=-\dfrac{\mathrm{e}^{\frac{x}{2}}}{\sqrt{3}}$, $x^*=0.045\,90$;

在$[0,1]$上取 $\varphi(x)=\dfrac{\mathrm{e}^{\frac{x}{2}}}{\sqrt{3}}$, $x^*=0.910\,0$,

在$[3,4]$上取 $\varphi(x)=\ln(3x^2)$, $x^*=3.733\,1$.

7.6 $x_{k+1}=\varphi^{-1}(x_k)=\pi+\arctan x_k$, $x_0=4.5$, $x_5=4.493\,4$.

7.9 (1) 不收敛; (2) 一阶收敛.

7.10 $x_0>0$ 时必收敛到 $x^*=\sqrt[3]{a}(>0)$.

7.12 $\Phi(\boldsymbol{x})=(0.7\sin x_1+0.2\cos x_2,0.7\cos x_1-0.2\sin x_2)^{\mathrm{T}}$,有 $\|\Phi'(x)\|_\infty\leqslant 0.9$. $\boldsymbol{x}^{(0)}=(0.5,0.5)^{\mathrm{T}}$, $\boldsymbol{x}^{(1)}=(0.511\,114,0.518\,423)^{\mathrm{T}}$, $\boldsymbol{x}^{(2)}=(0.516\,125,0.511\,438)^{\mathrm{T}}$,….

7.13 (1) $\boldsymbol{x}^{(0)}=(0.5,0.5)^{\mathrm{T}}$, $\boldsymbol{x}^{(1)}=(0.526\,824,0.508\,139)^{\mathrm{T}}$,…;

(2) $\boldsymbol{x}^{(0)}=(1.6,1.2)^{\mathrm{T}}$, $\boldsymbol{x}^{(1)}=(1.581\,25,1.225)^{\mathrm{T}}$, $\boldsymbol{x}^{(2)}=(1.581\,139,1.224\,745)^{\mathrm{T}}$.

7.14 (2) $\boldsymbol{x}^{(0)}=(1.6,1.2)^{\mathrm{T}}$, $\boldsymbol{x}^{(1)}=(1.581\,25,1.225)^{\mathrm{T}}$,….

习题 8

8.1 主特征值 $\lambda_1=7$,主特征向量 $\boldsymbol{x}_1=(-0.25,1,-0.25)^{\mathrm{T}}$.

8.2 迭代 57 次得 $\lambda_3\approx -3.599\,45$, $\boldsymbol{x}_3\approx(-0.861\,610,-0.671\,542,1)^{\mathrm{T}}$.

8.3 迭代 13 次得 $\lambda\approx 7.287\,99$, $\boldsymbol{x}\approx(1,0.522\,900,0.242\,192)^{\mathrm{T}}$.

8.4 对 A_2 有 $\lambda_1\approx 2.536\,5$, $\lambda_2\approx -0.016\,647$, $\lambda_3\approx 1.480\,2$.

$\boldsymbol{R}=\boldsymbol{P}_1^{\mathrm{T}}\boldsymbol{P}_2^{\mathrm{T}}\boldsymbol{P}_3^{\mathrm{T}}\boldsymbol{P}_4^{\mathrm{T}}\boldsymbol{P}_5^{\mathrm{T}}$ 的列向量为 $\boldsymbol{A}_2$ 的近似特征向量:

$$\boldsymbol{R}\approx\begin{pmatrix}0.531\,63 & -0.721\,10 & -0.444\,29\\ 0.461\,33 & 0.686\,45 & -0.562\,11\\ 0.710\,31 & 0.093\,87 & 0.697\,59\end{pmatrix}.$$

8.5 (1) 用平面旋转矩阵 $\boldsymbol{J}_1$,$\boldsymbol{J}_2$ 和 $\boldsymbol{J}_3$ 依次消去 $\boldsymbol{x}$ 的第 2,3,4 个元素,得

$$\boldsymbol{P}=\boldsymbol{J}_3\boldsymbol{J}_2\boldsymbol{J}_1=\begin{pmatrix}\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2}\\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0\\ -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & 0\\ -\frac{1}{2\sqrt{3}} & -\frac{1}{2\sqrt{3}} & -\frac{1}{2\sqrt{3}} & \frac{\sqrt{3}}{2}\end{pmatrix},\quad \boldsymbol{P}\boldsymbol{x}=\begin{pmatrix}2\\0\\0\\0\end{pmatrix};$$

(2) 用镜面反射变换得

$$\boldsymbol{P}=-\frac{1}{6}\begin{pmatrix}3 & 3 & 3 & 3\\ 3 & -5 & 1 & 1\\ 3 & 1 & -5 & 1\\ 3 & 1 & 1 & -5\end{pmatrix},\quad \boldsymbol{P}\boldsymbol{x}=\begin{pmatrix}-2\\0\\0\\0\end{pmatrix}.$$

8.6 (2) $$\boldsymbol{P}=\frac{1}{3}\begin{pmatrix}2 & 1 & 2\\ 1 & 2 & -2\\ 2 & -2 & -1\end{pmatrix},\quad \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{\mathrm{T}}=\begin{pmatrix}9 & & \\ & 18 & 0\\ & 0 & -9\end{pmatrix}.$$

8.8 构造镜面反射阵 $\boldsymbol{P}_1$ 和 $\boldsymbol{P}_2$,有

$$\boldsymbol{A}=\boldsymbol{Q}\boldsymbol{R},\quad \boldsymbol{Q}=\boldsymbol{P}_1\boldsymbol{P}_2=\frac{1}{3}\begin{pmatrix}-1 & 2 & 2\\ -2 & 1 & -2\\ -2 & -2 & 1\end{pmatrix},\quad \boldsymbol{R}=\begin{pmatrix}-3 & 3 & -3\\ & 3 & -3\\ & & 3\end{pmatrix}.$$

8.9 (1) $$\boldsymbol{Q}=\begin{pmatrix}0.000\,0 & 0.894\,4 & 0.447\,2\\ -1.000\,0 & 0.000\,0 & 0.000\,0\\ & -0.447\,2 & 0.894\,4\end{pmatrix},$$

$$\boldsymbol{B}=\begin{pmatrix}2.000\,0 & 0.000\,0 & 2.236\,1\\ -2.236\,1 & 1.000\,0 & -2.000\,0\\ & 0.000\,0 & 0.000\,0\end{pmatrix}.$$

(2) $$\boldsymbol{Q}=\begin{pmatrix}0.948\,7 & -0.274\,1 & 0.157\,6\\ 0.316\,2 & 0.822\,4 & -0.472\,9\\ & 0.498\,4 & 0.866\,9\end{pmatrix},$$

$$\boldsymbol{B}=\begin{pmatrix} 3.700\,0 & 1.268\,9 & \\ 1.268\,9 & 4.368\,3 & -0.039\,3 \\ & -0.039\,3 & -0.068\,3 \end{pmatrix}.$$

习题 9

9.1 0.500 00， 1.142 01， 2.501 15， 7.245 02.

9.2 1.11， 1.242 05， 1.398 47， 1.581 81， 1.794 90，
2.040 86， 2.323 15， 2.645 58， 3.012 37， 3.428 17.

9.3 0.145.

9.5 (1) 1.242 8， 1.583 6， 2.044 2， 2.651 0， 3.436 5；
(2) 1.727 6， 2.743 0， 4.094 2， 5.829 2， 7.996 0.

9.8 显式：0.626，隐式：0.633，准确值：0.632 1.

9.9 按经典 R-K 法：$y_1=0.090\,1$， $y_2=0.160\,8$， $y_3=0.213\,6$.
按 Adams 公式：$y_4=0.250\,6$， $y_5=0.274\,1$， $y_6=0.286\,6$.
按预估—校正公式：$y_4=0.250\,5$， $y_5=0.273\,9$， $y_6=0.286\,3$.

9.10 主项为 $-\frac{5}{8}h^3y'''(x_n)$.

9.11 $\alpha=\frac{1}{2}$， $\beta_0=\frac{7}{4}$， $\beta_1=-\frac{1}{4}$， 主项为 $\frac{3}{8}h^3y'''(x_n)$.

9.12 $\alpha_0=4$， $\alpha_1=-3$， $\beta_1=-2$， 局部截断误差为 $\frac{2}{3}h^3y'''(x_n)+O(h^4)$.

9.13 $y_1(x_1)\approx 0.247\,866\,666$， $y_1(x_2)\approx 0.632\,872\,209$，
$y_2(x_1)\approx 1.152\,704\,167$， $y_2(x_2)\approx 1.451\,602\,755$.

9.14 $y(0.2)\approx 0.98$.

9.15 (1) 非刚性方程组；(2) 刚性方程组；(3) 非刚性方程组.

9.16 $y_1=0.695\,3$， $y_2=0.608\,0$， $y_3=0.672\,6$.

9.17 1.014 87， 1.017 85， 1.070 10， 1.210 30， 1.513 29.

参 考 文 献

[1] Burden R L and Faires J D. *Numerical Analysis* (Fourth Edition). Prindle, Boston: Weder & Schmidt, 1989

[2] Laurene V F. *Applied Numerical Analysis Using MATLAB*. New Jersey: Prentice-Hall, 1999

[3] Stoer J, Bulirsch R. *Introduction to Numerical Analysis*. New Youk: Springer-Verlag, 1980

[4] 曹志浩,张玉德,李瑞遐. 矩阵计算与方程求根. 北京:高等教育出版社,1984

[5] 邓建中,刘之行. 计算方法(第 2 版). 西安:西安交通大学出版社,2001

[6] 关治,陆金甫. 数值分析基础. 北京:高等教育出版社,1998

[7] 韩旭里. 数值分析. 长沙:中南大学出版社,2003

[8] 黄友谦,程诗杰,陈译鹏. 数值试验. 北京:高等教育出版社,1989

[9] 黄友谦,李岳生. 数值逼近. 北京:高等教育出版社,1987

[10] 李荣华,冯果忱. 微分方程数值解法. 北京:人民教育出版社,1980

[11] 李庆扬,关治,白峰杉. 数值计算原理. 北京:清华大学出版社,2000

[12] 李庆扬,王能超,易大义. 数值分析(第三版). 武汉:华中理工大学出版社,1986

[13] 李庆扬. 常微分方程数值解法(刚性问题与边值问题). 北京:高等教育出版社,1990

[14] 李庆扬,莫孜中,祁力群. 非线性方程组的解法. 北京:科学出版社,1987

[15] 李庆扬,王能超,易大义. 现代数值分析. 北京:高等教育出版社,1995

[16] 施妙根,顾丽珍. 科学和工程计算基础. 北京:清华大学出版社,1999

[17] 王德人,杨忠华. 数值逼近引论. 北京:高等教育出版社,1990

[18] 魏毅强,张建国,张洪斌等. 数值计算方法. 北京:科学出版社,2004