电子计算机应用系列教材

数型计算机实用数值方法

郑慧娆　官士鸿 编著

科学出版社

(京)新登字 092 号

内 容 简 介

本书是电子计算机应用系列教材之一，主要介绍在计算机上解决各种应用课题的常用计算方法.

全书共八章，介绍了线性代数方程组的数值解法，插值计算和积分的数值计算，方程求根的常用算法和求一元函数极小的数值方法，以及线性规划问题的标准形式和常用的单纯形算法，还给出了常微分方程初值问题的数值解法和有限元方法的基本思想.

各章后均附有习题，书后还给出了有关误差分析内容的三个附录.

本书可作为计算机应用人员、工程技术人员及有关专业大学生的教材或参考书.

电子计算机应用系列教材

微型计算机实用数值方法

郑慧娆 宦士鸿 编著

责任编辑 刘晓融

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100707

天津市静一胶印厂印刷

新华书店北京发行所发行 各地新华书店经售

★

1992 年 12 月第 一 版 开本 787×1092 1/16

1992 年 12 月第一次印刷 印张 14

印数：0 001— 3,500 字数：314 000

ISBN 7-03 000906-1/TP・61

定价：8.80 元

目 录

第一章　数值计算与误差

1.1　计算方法与数值计算

计算机是解决在生产实践和科学研究中所提出的一些数学问题的重要计算工具. 例如，利用计算机可以进行人口普查、经济预测、天气预报、建筑设计等等.

通常，为使计算机处理这些实际问题，需要将它们表现为一个数学问题(提出数学模型)，并采用某种方法把这个数学问题化为一系列的加、减、乘、除四则运算，使计算机计算出结果. 这里，把**数学问题的求解过程，化为有限位数的有限次四则运算，就是计算方法的主要内容.**

如何把一个数学问题化为有限次的四则运算的计算呢?为此，我们介绍几种常用的办法.

把连续变量的问题化为离散问题

我们先看一个例子. 求物体的简谐振动中的位移量是由简谐振动方程

$$X = A\cos(\omega t + \varphi_0)$$

计算的. 对确定的角速度 ω，振幅 A 以及初相 φ_0，位移量 X 随着时间 t 的连续变化而连续改变. 我们可以利用列表的方法，对某些确定的时刻 $t_1,\cdots,t_k$(把时间变量离散化)，通过余弦函数表，计算出相应的位移量 $X_1,\cdots,X_k$. 列表就相当于把连续问题 $X = A\cos(\omega t+\varphi_0)$ 离散化了.

又例如，我们要求函数 $y = f(x)$ 在 $x = x_0$ 处的一阶导数 $f'(x_0)$，可以用差商代替微商，即

$$f'(x_0) = \frac{y(x_0 + \Delta x) - y(x_0)}{\Delta x}$$

同样，求解微分方程也可以通过计算它们的解在离散点处的近似结果来代替.

逼近

用简单的函数近似表示(代替)给定的数学问题中的函数，这是一种常用的处理方法. 这里所说的**简单函数**是指**可以用四则运算进行计算的函数**. 例如我们要计算

$$f(x) = e^{-x}$$

在 $x = 0.5$ 处的值，可以用一个二次多项式 $P_2(x)$ 近似代替 $f(x)$，计算 $P_2(0.5)$ 来近似代替 $f(0.5)$.

在计算方法中，这种近似代替又称为**逼近**. $f(x)$ 称为被逼近函数，简单函数 $P(x)$ 称为逼近函数，它们的差

$$E(x) = f(x) - P(x)$$

称为逼近的误差或余项. 按照不同的逼近方式又可分为插值、一致逼近、均方逼近(平方逼

近)等几种方法.

迭代

所谓**迭代**是指**把要求解的一个数学问题,化为一个极限过程**,而实现这个极限过程的每一步的结果,是把前一步所得的结果施行相同的运算得到的.例如,假设已知方程

$$9x^2 = \sin x + 1$$

在[0,1]内有一个根.首先,把方程化为等价的形式

$$x = \frac{1}{3}\sqrt{1 + \sin x}$$

并在区间[0,1]内任取初值,例如取 $x_0 = 0.4$,按迭代格式

$$x_{n+1} = \frac{1}{3}\sqrt{1 + \sin x_n}$$

从 $x_0 = 0.4$ 开始,求出 x_1,重复地计算,得到

$$x_0 = 0.4, x_1 = 0.3929, x_2 = 0.391985, x_3 = 0.391865,$$

$$x_4 = 0.391848, x_5 = 0.391847, x_6 = 0.391847$$

等等.如果我们每次都按6位小数进行计算,继续下去,可以得到数列 $x_0, x_1, x_2, \cdots$,这个数列的极限就是方程的一个根.在实际计算中,按照误差要求进行有限次迭代,就可以得到方程的一个近似根.

这种迭代方法常用于方程的求根、线性方程组、微分方程的求解等方面.

不论用什么方法,把求解一个数学问题的运算,化为有限次的四则运算时,都存在这样一个问题,即原数学问题的解与用某种方法进行运算的结果之间有多大差异的问题.显然,我们只对运算结果能较接近原数学问题的解(即精确结果)的计算方法感兴趣.因此,我们应从理论上对所选定的计算方法的运算结果的精确程度进行分析,以鉴别运算结果是否可靠,这是数值计算应考虑的内容.但是,我们在这里不准备从理论方面进行详细的讨论,而是着重介绍一些常用的计算方法和计算程序.

下面,我们通过一个简单的例子说明用某种计算方法进行计算所得的结果,与原数学问题的精确结果往往存有差异.

求二次方程

$$x^2 - 6.433x + 0.009474 = 0$$

的最小根.

如果按求根公式计算,其最小根为

$$x = \frac{6.433 - \sqrt{(6.433)^2 - 4 \times 0.009474}}{2}$$

$$= 1.4731 \times 10^{-3}$$

如果在一台只能表示4位数字的计算机上计算(注意,任何一种计算工具所能表示的位数都是有限的),按上面的计算公式,计算步骤为

(1) $\alpha = (6.433)^2 = 41.38$

(2) $\beta = 4 \times 0.0095 = 0.0380$

(3) $\gamma=\alpha-\beta=41.34$

(4) $\delta=\sqrt{\gamma}=6.430$

(5) $\sigma=6.433-\delta=0.003$

(6) $\tilde{x}=\sigma/2=0.0015=1.5\times10^{-3}$

这样算出的 $\tilde{x}$ 在第二位数字上与 x 就不一致了. 引起这个差异的主要原因是由于机器表示的位数有限,在上面运算过程中,使有些有效数字消失了.

但如果把计算式改为

$$x=\frac{2\times0.009474}{6.433+\sqrt{(6.433)^2-4\times0.009474}}$$

还是用同样的计算机计算,可算得

$$\tilde{\tilde{x}}=1.473\times10^{-3}$$

这是一个精确得多的结果.

可见,对一个数学问题,采用不同的计算方法,其计算结果与原数学问题的精确结果的差异是各不相同的. 由于计算工具表示数的位数是有限的,而计算的每一步对数都不是作精确的运算,有些计算方法,在大量的运算之后,就可能使运算结果与精确结果有较大的差异.

1.2 误 差

1.2.1 误差的种类

一个有实际意义的问题,从建立数学模型,选定计算方法,到在指定的计算机上进行计算,所得的计算结果与原数学问题的精确结果的差异,被我们称为误差. 从建立数学模型,到算出结果的整个过程,误差应包括以下几个方面:

模型误差. 对一个具体的实际问题,不论是经济方面的,社会方面的还是物理方面的,所建立的数学模型总是近似的. 这种误差称为模型误差.

观测误差. 在数学模型中,已知量都是有具体意义的,而且是通过观察、测量、实验等方式得到的. 因此,这些量也有误差,这种误差称为观测误差.

截断误差. 已建立起来的数学模型就是一个数学问题,理论上应有其精确的结果. 选定了某种行之有效的计算方法后,假设每一步的运算,都是对数作精确运算(即假设计算工具是理想化的),计算的结果也可以认为是精确的. 但这个计算的精确结果与理论上的精确结果也会有差异,这种误差称为截断误差.

例如,对于 $|x|\leqslant1$ 上的 $f(x)=\mathrm{arctg}x$,如果用 $\mathrm{arctg}x$ 在 $x=0$ 的泰勒 (Taylor)展开式

$$x-\frac{1}{3}x^3+\frac{1}{5}x^5-\frac{1}{7}x^7+\frac{1}{9}x^9\cdots$$

的前5项之和作为它的近似,即用一个多项式

$$g(x)=x-\frac{1}{3}x^3+\frac{1}{5}x^5-\frac{1}{7}x^7+\frac{1}{9}x^9$$

近似代替 arctg$x$. 假设运算的每一步都是精确的，但在 $|x| \leqslant 1$ 内，$f(x)$ 与 $g(x)$ 的值总是存在差异的.

舍入误差. 由于实际运算是按有限位数进行的，因此，在运算中常常要进行数的舍入. 这种由数的舍入带来的误差称为舍入误差.

所谓对**一个数 x 进行舍入，是指对它的十进制表示，保留从左起一定数目的数字，按某种规定确定最后一位的值，而把其余的数字都舍弃以后，得到数 x 的近似表示 $\tilde{x}$.**

例如，通常规定的舍入规则是对最后一位四舍五入. 那末，对 $x = 3.141592653\cdots$ 舍入为含有 3 个、4 个、5 个、6 个数字的近似数分别是：3.14，3.142，3.1416，3.14159 等等.

用计算机进行计算时，通常用二进制浮点数的形式表示数，因此，机器总是把十进制的原始数据翻译为在该机器上能表示的二进制数.

例如，一台阶码为 $-1 \leqslant J \leqslant 2$，尾数位数为 $t = 3$ 的计算机，它仅能表示33个二进制数①. 如果 $x = 2.5$，它可以表示为一个准确的二进制数 $x_{(2)} = 2^2 \times 0.101$；但对 $x = 2.25$，翻译为二进制数以后，在这台计算机上，只能近似地表示为 $\tilde{x}_{(2)} = 2^2 \times 0.101$ 或 $\tilde{x}_{(2)} = 2^2 \times 0.100$. 如果所使用的计算机的阶码仍为 $-1 \leqslant J \leqslant 2$，但尾数字长 $t = 4$，那末，$x = 2.25$ 就可以用一个准确的二进制数表示出来了(这时，$\tilde{x}_{(2)} = 2^2 \times 0.1001$).

另外，把原始数据翻译为二进制数输入到计算机进行运算，同样受到机器字长的限制，使每一步运算所得的结果都要进行舍入，每一步运算的结果都是近似的，从而使最后所得的运算结果是近似的.

例如，在 $-1 \leqslant J \leqslant 2, t = 3$ 这台计算机上，完成0.3125+2这一步运算时，虽然0.3125和2在这台计算机上都能准确地表示为一个二进制数，但其运算结果2.3125只能用一个近似的二进制数表示. 又例如在这台计算机上计算2.2+0.225时，由于只能近似地表示为2+0.25，且运算结果2.25也只能近似地表示为2.5.

一般地，两个浮点数运算时，按浮点数的舍入规则进行舍入引进的误差，称为浮点舍入误差②.

1.2.2 误差的几个基本概念

形式上，

$$\text{误差} = \text{精确值} - \text{近似值}$$

设 x 是一个精确数，x^* 表示 x 的近似数，我们给出以下定义.

定义 1.1 $\varepsilon^* = |x - x^*|$ 称为近似数 x^* 的绝对误差.

由于精确数 x 一般是未知的，因此，绝对误差 ε^* 也不能算出来. 但我们可以根据具体情况，得到估计式

$$|x - x^*| \leqslant \eta$$

从而可以知道 x 的范围：

$$x^* - \eta \leqslant x \leqslant x^* + \eta$$

即知道 x 落在区间 $[x^* - \eta, x^* + \eta]$ 内. 这里，η 称为绝对误差限.

---

① 这33个数在附录 C 表 C.1 中列出.
② 关于浮点数运算的舍入误差，可参阅附录 C.

绝对误差的大小，不能完全反映近似数的近似程度. 例如，$x_1 = 3.7$，它的近似数 $x_1^* = 3.6$，其绝对误差 $\varepsilon_1 = 0.1$；另一个数 $x_2 = 10000000$，它的近似数 $x_2^* = 10000005$，其绝对误差 $\varepsilon_2 = 5$. 形式上，5 比 0.1 要大得多，但按一千万之内相差 5 与十分之一内相差 1 的比值来看，后一个近似数更接近其精确值. 因此，我们除了要看绝对误差的大小外，还应顾及与这个准确数本身的比值. 下面，我们引进相对误差的定义.

定义1.2

$$k^* = \left|\frac{x - x^*}{x}\right|$$

称为近似数 x^* 的相对误差.

例如上面给出的例子中，对 x_1^*，其相对误差为 2.7%，对 x_2^*，其相对误差为 0.00005%.

类似地，我们可以找到 δ，满足

$$\left|\frac{x - x^*}{x}\right| \leqslant \delta$$

称 δ 为 x^* 的相对误差限.

在数值计算中，常常按四舍五入的原则进行舍入，得到一个精确数 x 的近似数 x^*. 四舍五入的原则是：4以下的数字舍，5以上的数字入. 若最后一位数字刚好是5，则作如下规定：5前面的一位数字如果是偶数，将5舍去；如果是奇数，5前面的一位数字加1. 也就是说，5前面的一位数字总是偶数. 例如，$\pi = 3.141592653\cdots$，如果要舍入为一个有5位数字的近似数，就是3.1416. 如果把3.1415舍入为有4位数字，舍入时因最后一位数字刚好是5，而它前面一位数字是奇数，故这个近似数就是3.142. 如果把3.14159265舍入为有8位数字，这时最后一位数字也刚好是5，但它前面一位数字是偶数，因此这个近似数为3.1415926.

实践证明，进行大量运算时，按上述原则舍入，整个运算的误差积累较小.

值得一提的是，当小数部分最后一位数是零时，仍表示一位数字. 例如，把8.00003舍入为有5位数字的近似数时，应写为8.0000，而不能写为8，前者表示这个近似数准确到小数后第4位.

在近似数的表示中，常用有效数字来描述. 一个近似数 x^* 具有 n 位有效数字是指，如果

$$|\Delta x| = |x - x^*| \leqslant \frac{1}{2}10^{-k}$$

那末，我们就说用 x^* 近似表示 x 时，准确到小数后第 k 位，又把以这个数字起直到最左的非零数字之间的一切数字都叫做有效数字.

例如，对于3.1415926…，3.1416是具有5位有效数字的近似数，对于8.00003，8.0000也是具有5位有效数字的近似数.

又例如，重力常数 g，如果以米/秒^2 为单位，$g \approx 9.80$ 米/秒^2，它是一个具有 3 位有效数字的近似数. 但如果以千米/秒^2 为单位，则有 $g \approx 0.00980$ 千米/秒^2，这时，它是一个有 5 位数字的近似数，但还是具有 3 位有效数字.

可见，有效数字的位数与小数后有多少位数字并没有什么直接关系.

第二章　线性方程组的数值解法

2.1　引　言

生产实践和科学实验中，许多问题常常直接或间接地归结为含多个未知量的线性方程组的求解问题.

如果未知量的个数为 n，而且关于这些未知量 $x_1, x_2, \cdots, x_n$ 的幂次都是一次的(线性的)，那末，n 个方程

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ \vdots \quad & \quad \vdots \quad \quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned}\right\} \tag{2.1}$$

构成一个含 n 个未知量的线性方程组，称为 n 阶线性方程组. 其中，系数 $a_{11}, \cdots, a_{1n}, a_{21}, \cdots, a_{2n}, \cdots, a_{n1}, \cdots, a_{nn}$ 是给定的常数；$b_1, \cdots, b_n$ 也是给定的常数，通常称为常数项，或称为方程组的右端.

方程组(2.1)也常用矩阵的形式表示，写为

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{2.2}$$

其中，A 是由系数按次序排列构成的一个 n 阶矩阵，即

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

称为方程组的系数矩阵. $\boldsymbol{x}$ 和 $\boldsymbol{b}$ 都是 n 维向量，其中

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

向量 $\boldsymbol{b}$ 也称为方程组的右端向量.

使方程组(2.1)中每一个方程的等号都成立的一组数 $x_1, x_2, \cdots, x_n$ 称为式(2.1)的解，把它记为向量的形式，称为解向量.

我们总是希望方程组有解，且有唯一解. 由线性代数的克莱姆(Cramer)规则可知，如果方程组(2.1)的系数矩阵 A 的行列式(一般记为 $D = |A|$)不等于零，那末，这个方程组有唯一解，而且它们可以表为

$$x_i = D_i/D \qquad (i = 1, \cdots, n)$$

这里，D_i 是指 D 中第 i 列元素用右端 $b_1, \cdots, b_n$ 代替构成的行列式.

如果方程组(2.1)有唯一解，我们按上面的等式求解，就必须计算 $n+1$ 个 n 阶行列式. 由行列式的定义，n 阶行列式包含有 $n!$ 项，每一项含有 n 个因子，计算一个 n 阶行列式就需要做 $(n-1)n!$ 次乘法. 而我们一共要计算 $n+1$ 个 n 阶行列式，共需做 $(n^2-1)n!$ 次乘法. 此外，还要做 n 次除法才能算出 $x_i(i=1,\cdots,n)$. 也就是说，用这个办法求解就要做

$$N=(n^2-1)n!+n$$

次乘除法运算，这个计算量是大得惊人的. 例如，当 $n=10$(即求解一个含 10 个未知量的方程组)，乘除法的运算次数共为 32,659,210 次；当 $n=40$，乘除法运算次数可达 3.18×10^{49} 次. 对于上百个未知量的方程组，运算次数就更大了. 因此，克莱姆规则在理论上尽管是完善的，但在实际计算中却没有什么实用价值. 这一章，我们将介绍求解线性方程组的有效的数值方法. 在这些数值方法中，一般可以分成两类：精确方法(也常常称为直接法)与迭代方法，分别在 2.2,2.3,2.4,2.5 和 2.6 节中介绍. 本章 2.7 节将介绍计算超定方程组的最小二乘解的数值方法.

2.2 求解线性方程组的直接法

求解线性方程组的直接法，实际上是一种精确方法. 即用有限次四则运算，并在假设每一步运算过程都不引进舍入误差的前提下，计算的结果(所求的解)应该是方程组的精确解. 但在实际计算中，由于计算工具的限制，例如，计算机表示数的位数只能是有限的等等，在运算过程中的每一步都会引进浮点舍入误差. 因此，计算的结果实际上是不精确的. 也就是说，用直接法求解方程组只能得到方程组的数值解.

消去法是求解线性方程组

$$\boldsymbol{Ax}=\boldsymbol{b} \tag{2.3}$$

和满秩矩阵的逆阵 A^{-1} 的一种直接方法. 尽管它比较古老，但它具有演算步骤、推算公式都系统化的特点(对其中选主元的消去法，还可以证明是稳定的). 因此，它至今仍然是求解方程组中的一种有效的数值的方法.

消去法包括几种计算方案，下面将分别介绍.

2.2.1 Gauss 消去法

首先讨论一个简单的例子.

求解一个 3 阶方程组

$$\left.\begin{aligned} 2x_1+2x_2+3x_3&=3\\ 4x_1+7x_2+7x_3&=1\\ -2x_1+4x_2+5x_3&=-7 \end{aligned}\right\} \tag{2.4}$$

我们希望通过等价变换，把它先化为一个系数矩阵呈三角形的方程组，以便求解. 通常采用的是消去的方法.

首先，用方程组(2.4)的第 1 个方程消去其余两个方程中第 1 个未知量 x_1(即使 x_1 的系数为 0). 只要用 -2 乘第 1 个方程加到第 2 个方程上，把第 1 个方程加到第 3 个方程上，就可以得到一个等价的新方程组

$$\left.\begin{aligned} 2x_1 + 2x_2 + 3x_3 &= 3 \\ 3x_2 + x_3 &= -5 \\ 6x_2 + 8x_3 &= -4 \end{aligned}\right\}$$

类似地，用第 2 个方程消去第 3 个方程中 x_2 的系数，即用 -2 乘第 2 个方程加到第 3 个方程上，得到三角形方程组

$$\left.\begin{aligned} 2x_1 + 2x_2 + 3x_3 &= 3 \\ 3x_2 + x_3 &= -5 \\ 6x_3 &= 6 \end{aligned}\right\}$$

它也与原方程组同解.

这时，直接从第 3 个方程可求出 $x_3 = 1$；把 x_3 代入到第 2 个方程中解出 $x_2 = -2$；最后，把 x_2, x_3 代入到第 1 个方程中，就可以得到 $x_1 = 2$.

即原方程组的解为 $x_1 = 2, x_2 = -2, x_3 = 1$.

综合上述求解过程，可以大致分为两个阶段：首先，把原方程组化为三角形方程组，我们把它称为“消去”过程；然后，用逆次序逐一求出三角形方程组（即原方程组）的解，并称之为“回代”过程.

这样的方法称为 Gauss（高斯）消去法. 下面，以 $n = 4$ 的情形，推导出用 Gauss 消去法求解一般的 n 阶线性方程组的计算公式.

考虑 4 阶方程组

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 &= b_3 \\ a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 &= b_4 \end{aligned}\right\} \tag{2.5}$$

假设 $a_{11} \neq 0$，我们保留第 1 个方程并用它消去其余 3 个方程的第一个未知量 x_1. 消去的办法是，令

$$l_{21} = \frac{a_{21}}{a_{11}}, \qquad l_{31} = \frac{a_{31}}{a_{11}}, \qquad l_{41} = \frac{a_{41}}{a_{11}}$$

然后分别用 $-l_{21}, -l_{31}, -l_{41}$ 乘第 1 个方程再分别加到第 2,3 和第 4 个方程上，得到一个等价的新方程组

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}a_4 &= b_1 \\ a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + a_{24}^{(1)}x_4 &= b_2^{(1)} \\ a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + a_{34}^{(1)}x_4 &= b_3^{(1)} \\ a_{42}^{(1)}x_2 + a_{43}^{(1)}x_3 + a_{44}^{(1)}x_4 &= b_4^{(1)} \end{aligned}\right\}$$

这里，

$$\begin{aligned} a_{ij}^{(1)} &= a_{2j} - l_{21} \cdot a_{1j} \qquad (j = 2,3,4) \\ a_{3j}^{(1)} &= a_{3j} - l_{31} \cdot a_{1j} \qquad (j = 2,3,4) \\ a_{4j}^{(1)} &= a_{4j} - l_{41} \cdot a_{1j} \qquad (j = 2,3,4) \\ b_i^{(1)} &= b_i - l_{i1} \cdot b_1 \qquad (i = 2,3,4) \end{aligned}$$

我们把这个计算过程称为第 1 步消去，记为 $k = 1$. 消去后新方程组中，除第 1 个方程

外，系数和右端分别记为 $a_{ij}^{(1)}$ 和 $b_i^{(1)}$. 这里，右上角的脚标数相应于消去的第 1 步.

假设 $a_{22}^{(1)} \neq 0$，保留第 1，2 两个方程，用第 2 个方程消去其余两个方程的 x_2. 令

$$l_{32} = \frac{a_{32}^{(1)}}{a_{22}^{(1)}}, \qquad l_{42} = \frac{a_{42}^{(1)}}{a_{22}^{(1)}}$$

分别用 $-l_{32}$，$-l_{42}$ 乘第 2 个方程加到第 3 和第 4 个方程上，就完成了消去的第 2 步($k=2$)，得到等价的方程组

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 &= b_1 \\ a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + a_{24}^{(1)}x_4 &= b_2^{(1)} \\ a_{33}^{(2)}x_3 + a_{34}^{(2)}x_4 &= b_3^{(2)} \\ a_{43}^{(2)}x_3 + a_{44}^{(2)}x_4 &= b_4^{(2)} \end{aligned}\right\}$$

其中，

$$\left.\begin{aligned} a_{ij}^{(2)} &= a_{ij}^{(1)} - l_{i2}\cdot a_{2j}^{(1)} \qquad (j=3,4) \\ b_i^{(2)} &= b_i^{(1)} - l_{i2}\cdot b_2^{(1)} \end{aligned}\right\}(i=3,4)$$

消去的第 3 步 ($k=3$)，用第 3 个方程消去 x_3. 假设 $a_{33}^{(2)} \neq 0$，令

$$l_{43} = \frac{a_{43}^{(2)}}{a_{33}^{(2)}}$$

并用 $-l_{43}$ 乘第 3 个方程加到第 4 个方程上，得到等价的三角形方程组

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 &= b_1 \\ a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + a_{24}^{(1)}x_4 &= b_2^{(1)} \\ a_{33}^{(2)}x_3 + a_{34}^{(2)}x_4 &= b_3^{(2)} \\ a_{44}^{(3)}x_4 &= b_4^{(3)} \end{aligned}\right\} \tag{2.6}$$

其中，

$$\begin{aligned} a_{44}^{(3)} &= a_{44}^{(2)} - l_{43}\cdot a_{34}^{(2)} \\ b_4^{(3)} &= b_4^{(2)} - l_{43}\cdot b_3^{(2)} \end{aligned}$$

这样，经过 3 步消去就完成了“消去”过程.

一般的 n 阶方程组，只要经过 $n-1$ 步消去，可以化为等价的三角方程组.

第 1 步，除第 1 个方程外，消去其余各方程的 x_1，假如 $a_{11} \neq 0$，令

$$l_{i1} = \frac{a_{i1}}{a_{11}} \qquad (i = 2,3,\cdots,n)$$

然后，计算

$$(-l_{i1}) \times \text{第 1 个方程} + \text{第 } i \text{ 个方程} \qquad (i=2,\cdots,n)$$

得到的新方程组中，除第 1 个方程外，其余各方程的系数和右端按下式计算：

$$\left.\begin{aligned} a_{ij}^{(1)} &= a_{ij} - l_{i1}\cdot a_{1j} \qquad (j=2,\cdots,n) \\ b_i^{(1)} &= b_i - l_{i1}\cdot b_1 \end{aligned}\right\} \quad (i=2,\cdots,n)$$

类似地，经过 $n-1$ 步消去，就可以得到等价的三角形方程组.

在消去过程的第 k 步，假定 $a_{kk}^{(k-1)} \neq 0$，令

$$l_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} \qquad (i = k+1,\cdots,n) \tag{2.7}$$

然后计算

$$(-l_{ik})\times \text{第}k\text{个方程}+\text{第}i\text{个方程}\qquad (i=k+1,\cdots,n)$$

就可以完成第 k 步的消去. 所得的新方程中，系数 $a_{ij}^{(k)}(j=k+1,\cdots,n;i=k+1,\cdots,n)$ 和右端 $b_i^{(k)}(i=k+1,\cdots,n)$ 可按下式计算：

$$\left.\begin{aligned} a_{ij}^{(k)} &= a_{ij}^{(k-1)} - l_{ik}\,a_{kj}^{(k-1)}(j=k+1,\cdots,n)\\ b_i^{(k)} &= b_i^{(k-1)} - l_{ik}b_k^{(k-1)} \end{aligned}\right\}\quad (i=k+1,\cdots,n) \tag{2.8}$$

只要让 k 从 1 变到 $n-1$，经过 $n-1$ 个消去步，完成“消去”过程的计算.

式(2.7)和式(2.8)是 Gauss 消去法“消去”过程的一般计算公式.

式(2.8)中，当 $k=1$ 时，$a_{ij}^{(k-1)}=a_{ij}^{(0)}$，$a_{ij}^{(0)}$ 是指原方程组的系数 a_{ij}. 同样 $b_i^{(k-1)}=b_i^{(0)}=b_i$.

下面再看“回代”过程. 还是先讨论 $n=4$ 的简单情形.

由方程组(2.6)看出，只要 $a_{44}^{(3)}\neq 0$，可以从第 4 个方程解出

$$x_4=b_4^{(3)}/a_{44}^{(3)}$$

用 x_4 的值代入第 3 个方程，得出

$$x_3=(b_3^{(2)}-a_{34}^{(2)}x_4)/a_{33}^{(2)}$$

同样，把 x_3,x_4 代入第 2 个方程中，得出

$$x_2=(b_2^{(2)}-a_{23}^{(1)}x_3-a_{24}^{(1)}x_4)/a_{22}^{(1)}$$

最后，由第 1 个方程和已算出的 x_2,x_3,x_4，可得到

$$x_1=(b_1-a_{12}x_2-a_{13}x_3-a_{14}x_4)/a_{11}$$

上面各式中的分母 $a_{33}^{(2)},a_{22}^{(1)}$ 和 a_{11} 在消去过程已假设不等于 0.

这就是求解的“回代”过程. 我们可以把它归纳为

$$\left.\begin{aligned} x_4 &= b_4^{(3)}/a_{44}^{(3)}\\ x_i &= (b_i^{(i-1)}-\sum_{j=i+1}^{4}a_{ij}^{(i-1)}x_j)/a_{ii}^{(i-1)}\quad (i=3,2,1) \end{aligned}\right\}$$

也就是说，可以用前面已求出的 $4-i$ 个未知量，代入到第 i 个方程中，计算出第 i 个未知量来. 注意到计算 x_i 的式子中，分子的项数是随着脚标 i 的减少而增加的. 例如，计算 x_3 时 $(i=3)$，分子有两项；而计算 x_2 时 $(i=2)$，分子有三项. 而且，回代过程是按脚标 i 从大到小的逆方向进行的. 特别，当 $i=n$ 时，只用一次除法就可以计算 x_n 了.

由此得出一般 n 阶方程组“回代”过程的计算式：

$$\left.\begin{aligned} x_n &= b_n^{(n-1)}/a_{nn}^{(n-1)}\\ x_i &= (b_i^{(i-1)}-\sum_{j=i+1}^{n}a_{ij}^{(i-1)}x_j)/a_{ii}^{(i-1)}\quad (i=n-1,\cdots,2,1) \end{aligned}\right\} \tag{2.9}$$

2.2.2 Gauss 消去法的计算步骤

给出详细的计算步骤之前，首先作两点说明.

1. 给出方程组 $Ax=b$，假设分别把系数阵 A 和右端向量 b 存放在一个二维数组 $A(N,N)$ 和一维数组 $B(N)$ 中. 在第一步消去时，除第一个方程外，其余各方程的系数和右端均按下式计算：

$$\left.\begin{aligned} a_{ij}^{(1)} &= a_{ij} - \frac{a_{i1}}{a_{11}}a_{1j} \quad (j = 2,\cdots,n) \\ b_i^{(1)} &= b_i - \frac{a_{i1}}{a_{11}}b_1 \end{aligned}\right\} \quad (i = 2,\cdots,n)$$

我们常把新的方程组的系数阵 $A^{(1)}$ 和右端向量 $b^{(1)}$ 仍存放回原来的数组 A,B 中. 这时，除了 A 的第一行各单元和 B 的第一个单元保持不变外，其余各单元内容均按上式计算. 因此，可把上式改写为

$$\left.\begin{aligned} a_{ij} - \frac{a_{i1}}{a_{11}}a_{1j} &\longrightarrow a_{ij} \quad (j = 2,\cdots,n) \\ b_i - \frac{a_{i1}}{a_{11}} \cdot b_1 &\longrightarrow b_i \end{aligned}\right\} \quad (i = 2,\cdots,n)$$

对于第 k 步消去的情形也类似地处理. 这样，完成“消去”过程的各步计算后，原来的数组 A 和 B，分别存放了三角形方程组的系数矩阵和右端. 因此，“消去”过程的式(2.7)和式(2.8)可以改写为

$$\frac{a_{ik}}{a_{kk}} \longrightarrow l_{ik} \qquad (i = k+1,\cdots,n) \tag{2.7}'$$

和

$$\left.\begin{aligned} a_{ij} - l_{ik} \cdot a_{kj} &\longrightarrow a_{ij} \quad (j = k+1,\cdots,n) \\ b_i - l_{ik} \cdot b_k &\longrightarrow b_i \end{aligned}\right\} \quad (i = k+1,\cdots,n) \tag{2.8}'$$

2. “回代”过程中，计算 x_i 的式(2.9)中，分子包含累加项 $\sum_{j=i+1}^{n} a_{ij}^{(i-1)}x_j$. 我们往往先计算这个累加项并存放在一工作单元 σ 中，然后再计算 $b_i^{(i-1)} - \sigma$. 在计算下一个 x_{i+1} 时，只要先将 σ 置0，再重新计算累加项. 这样处理，主要是为了便于程序中循环过程的实现，同时也节省了存储单元.

此外，考虑到“回代”过程中，计算出 x_i 后，相应右端 b_i 不需要参加运算了. 因此，x_i 可存放在 b_i 所在单元中. “回代”过程的计算完成后，数组 B 放所求的解向量 x.

同样，“回代”过程的式(2.9)可以改写

$$\left.\begin{aligned} &b_n/a_{nn} \longrightarrow b_n \\ &(b_i - \sum_{j=i+1}^{n} a_{ij}b_j)/a_{ii} \longrightarrow b_i \quad (i = n-1,\cdots,2,1) \end{aligned}\right\} \tag{2.9}'$$

下面分别写出“消去”和“回代”两个过程的计算步骤.

消去过程：

(1) K 循环($k = 1,\cdots,n-1$) 控制消去过程的第几步.

(1.1) I 循环($i = k+1,\cdots,n$) 消去第 i 个方程第 k 个未知量. i 依赖于循环参数 k.

(1.1.1) 计算 l_{ik}.

$$a_{ik}/a_{kk} \longrightarrow l_{ik}$$

(1.1.2) J 循环($j = k+1,\cdots,n$) 计算第 i 个方程中未知量 x_j 的系数 $a_{ij}^{(k)}$. j 依赖于循环参数 k.

(1.1.2.1) 计算 $a_{ij}^{(k)}$.

$$a_{ij} - l_{ik} \cdot a_{kj} \to a_{ij}$$

(1.1.3) J 循环结束.

(1.1.4) 计算右端 $\boldsymbol{b}$ 的第 i 个分量 $b_i^{(k)}$.

$$b_i - l_{ik} \cdot b_k \to b_i$$

(1.2) I 循环结束.

(2) K 循环结束.

回代过程：

(1) 计算 x_n(存放在 b_n 的单元中).

$$b_n / a_{nn} \to b_n$$

(2) I 循环($i = n-1, \cdots, 1$) 计算 x_i. i 的改变量 $\Delta = -1$.

(2.1) J 循环($j = i+1, \cdots, n$)，计算累加项 $\sum_{j=i+1}^{n} a_{ij}$. 累加项的项数为 $n-i$，j 依赖循环参数 i.

(2.1.1) 计算 σ.

$$\sum_{j=i+1}^{n} a_{ij} b_j \to \sigma$$

(2.2) J 循环结束.

(2.3) 计算 x_i(存放在 b_i 的单元中).

$$(b_i - \sigma) / a_{ii} \to b_i$$

(3) I 循环结束.

2.2.3 矩阵的三角分解(LU 分解)

上面已指出，Gauss 消去法的“消去”过程，实际上就是用若干次变换，把原方程组化为等价的三角形方程组. 由线性代数的基本知识可知，这些等价变换，又相当于对增广矩阵施行若干次行初等变换，即用若干个初等变换阵左乘方程组的增广矩阵.

下面，我们用初等变换矩阵的运算来表示这个变换过程. 为说明方便起见，我们还是先讨论 $n = 4$ 的简单情形.

设4阶方程组的增广矩阵为

$$\left[\begin{array}{cccc:c} a_{11} & a_{12} & a_{13} & a_{14} & b_1 \\ a_{21} & a_{22} & a_{23} & a_{24} & b_2 \\ a_{31} & a_{32} & a_{33} & a_{34} & b_3 \\ a_{41} & a_{42} & a_{43} & a_{44} & b_4 \end{array}\right]$$

简单的记为

$$(A \vdots \boldsymbol{b})$$

用 $-l_{i1} = a_{i1}/a_{11}$ 乘第一个方程加到第 i 个方程($i = 2,3,4$) 上，就相当于分别用初等变换矩阵

$$I_{12}(-l_{21}) = \begin{bmatrix} 1 & & & \\ -l_{21} & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$I_{13}(-l_{31})=\begin{bmatrix}1&&&\\0&1&&\\-l_{31}&0&1&\\0&0&0&1\end{bmatrix}$$

以及

$$I_{14}(-l_{41})=\begin{bmatrix}1&&&\\0&1&&\\0&0&1&\\-l_{41}&0&0&1\end{bmatrix}$$

左乘 $(A\ \vdots\ \boldsymbol{b})$. 其中形为 $I_{ij}(k)$ 的矩阵相当于用 k 乘单位矩阵 I 的第 i 行加到第 j 行上去所得的初等变换阵，即

$$I_{ij}(k)=\begin{bmatrix}1&&&&&\\&\ddots&&&&\\&&1&&&\\&&&\ddots&&\\&&k&\cdots&1&\\&&&&&\ddots\\&&&&&1\end{bmatrix}\begin{matrix}\\\\\text{第 }i\text{ 行}\\\\\text{第 }j\text{ 行}\\\\\end{matrix}$$

容易验证，

$$I_{14}(-l_{41})I_{13}(-l_{31})I_{12}(-l_{21})=\begin{bmatrix}1&&&\\-l_{21}&1&&\\-l_{31}&0&1&\\-l_{41}&0&0&1\end{bmatrix}$$

记为 $M^{(1)}$，那末，

$$M^{(1)}(A\ \vdots\ \boldsymbol{b})=\left[\begin{array}{cccc:c}a_{11}&a_{12}&a_{13}&a_{14}&b_1\\0&a_{22}^{(1)}&a_{23}^{(1)}&a_{24}^{(1)}&b_2^{(1)}\\0&a_{32}^{(1)}&a_{33}^{(1)}&a_{34}^{(1)}&b_3^{(1)}\\0&a_{42}^{(1)}&a_{43}^{(1)}&a_{44}^{(1)}&b_4^{(1)}\end{array}\right]$$

记为 $(A^{(1)}\ \vdots\ \boldsymbol{b}^{(1)})$.

第 2 步消去过程，就相当于初等变换矩阵 $I_{23}(-l_{32})$，$I_{24}(-l_{42})$ 左乘 $(A^{(1)}\ \vdots\ \boldsymbol{b}^{(1)})$. 若记

$$M^{(2)}=I_{23}(-l_{32})I_{24}(-l_{42})=\begin{bmatrix}1&&&\\0&1&&\\0&-l_{32}&1&\\0&-l_{42}&0&1\end{bmatrix}$$

那末，

$$M^{(2)}(A^{(1)}\ \vdots\ \boldsymbol{b}^{(1)})=M^{(2)}M^{(1)}(A\ \vdots\ \boldsymbol{b})=\left[\begin{array}{cccc:c}a_{11}&a_{12}&a_{13}&a_{14}&b_1\\0&a_{22}^{(1)}&a_{23}^{(1)}&a_{24}^{(1)}&b_2^{(1)}\\0&0&a_{33}^{(2)}&a_{34}^{(2)}&b_3^{(2)}\\0&0&a_{43}^{(2)}&a_{44}^{(2)}&b_4^{(2)}\end{array}\right]$$

记为 $(A^{(2)} \mid \boldsymbol{b}^{(2)})$.

最后,用初等变换矩阵

$$M^{(3)} = I_{34}(-l_{43}) = \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & -l_{43} & 1 \end{bmatrix}$$

左乘 $(A^{(2)} \mid \boldsymbol{b}^{(2)})$,得

$$M^{(3)}(A^{(2)} \mid \boldsymbol{b}^{(2)}) = M^{(3)}M^{(2)}M^{(1)}(A \mid \boldsymbol{b}) = \left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & a_{14} & b_1 \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & b_2^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & b_3^{(2)} \\ 0 & 0 & 0 & a_{44}^{(3)} & b_4^{(3)} \end{array}\right]$$

记为 $(A^{(3)} \mid \boldsymbol{b}^{(3)})$,且 $A^{(3)}$ 呈上三角形. 我们分别把 $A^{(3)}$ 和 $\boldsymbol{b}^{(3)}$ 记为

$$A^{(3)} = U = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ & & a_{33}^{(2)} & a_{34}^{(2)} \\ & & & a_{44}^{(3)} \end{bmatrix}, \quad \boldsymbol{b}^{(3)} = \boldsymbol{y} = \begin{bmatrix} b_1 \\ b_2^{(1)} \\ b_3^{(2)} \\ b_4^{(3)} \end{bmatrix}$$

并记 $M = M^{(3)}M^{(2)}M^{(1)}$,那么,$M$ 是一个单位下三角形矩阵(对角线元素都为 1 的下三角形矩阵),因此,M 是满秩的.

我们把"消去"过程记为

$$M(A \mid \boldsymbol{b}) = (MA \mid M\boldsymbol{b}) = (A^{(3)} \mid \boldsymbol{b}^{(3)}) = (U \mid \boldsymbol{y})$$

得到

$$\left.\begin{aligned} MA &= U \\ M\boldsymbol{b} &= \boldsymbol{y} \end{aligned}\right\} \tag{2.10}$$

记 $L = M^{-1} = (M^{(3)}M^{(2)}M^{(1)})^{-1} = M^{(1)^{-1}}M^{(2)^{-1}}M^{(3)^{-1}}$. 由初等变换矩阵的性质,可以得到

$$M^{(1)^{-1}} = \begin{bmatrix} 1 & & & \\ l_{21} & 1 & & \\ l_{31} & 0 & 1 & \\ l_{41} & 0 & 0 & 1 \end{bmatrix}, \quad M^{(2)^{-1}} = \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & l_{32} & 1 & \\ 0 & l_{42} & 0 & 1 \end{bmatrix}$$

以及

$$M^{(3)^{-1}} = \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & l_{43} & 1 \end{bmatrix}$$

从而有

$$L = \begin{bmatrix} 1 & & & \\ l_{21} & 1 & & \\ l_{31} & l_{32} & 1 & \\ l_{41} & l_{42} & l_{43} & 1 \end{bmatrix}$$