重点大学计算机教材

数 值 方 法

浙 江 大 学

金一庆 陈 越 编著

机械工业出版社
China Machine Press

本书采用较少的篇幅，深入浅出地讲述了一些重要数值方法的来龙去脉。全书详细介绍了误差概念、非线性方程求根方法、矩阵的特征值与特征向量的计算、插值、曲线拟合与函数逼近、数值积分的方法、常微分方程求解、偏微分方程求解等问题。希望通过此书的介绍，使读者掌握这些方法的基本思想和基本技巧。

图书在版编目（CIP）数据

数值方法／金一庆，陈越编著. -北京：机械工业出版社，2000.2
（重点大学计算机教材）
ISBN 7-111-07578-1

Ⅰ.数… Ⅱ.①金… ②陈… Ⅲ.数值计算-计算方法 Ⅳ.0241

中国版本图书馆CIP数据核字（2000）第10352号

机械工业出版社（北京市西城区百万庄大街22号 邮政编码 100037）
责任编辑：艾玉娟
北京第二外国语学院印刷厂印刷 · 新华书店北京发行所发行
2000年2月第1版第1次印刷
787mm × 1092mm 1/16 · 17印张
印数：0 001-5 000册
定价：25.00元

第1章 误 差

1.1 误差的来源与分类

从实际工程问题出发，一直到算出问题结果，这其中的每个过程都会产生误差。如果用理论数学模型来描绘物理量，则产生的误差 $|u-u_1|$ 称为模型误差。如果将理论数学模型中的参数用有限位观测数据代入得到实际数学模型，则产生的误差 $|u_1-u_2|$ 称为观测误差或参量误差。

用如下的字母表示模型中的量和解：

u ——物理模型客观量

u_1 ——理论数学模型的解

u_2 ——实际数学模型的解

u_3 ——可解数学模型的解

u_4 ——有限位运算后得到的结果

如果用数值方法求近似解，又会出现方法误差。如：用泰勒展开求函数 $\sin x$ 的值：

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

从上式看，要求出准确无误的解必须经过无穷步运算来完成，而实际上这是办不到的，只能取式中的有限项，如当 x 趋向 0 时，用 x 近似代替 $\sin x$，这样造成的截断误差差不多是 $\frac{1}{6}x^3$，这时，截断误差就是一种方法误差。

由于计算机只能进行有限位运算，所以在运算过程中会不断地按某种规则进行舍入，这样会产生舍入误差，也称为凑整误差。同样在函数的运算中，如果自变量有误差，函数也会因此产生误差，这称为误差传播。如可以把观测值看成自变量，运算结果看成函数。如果观测值有初始误差，就会造成误差传播。而且，在任何中间步骤产生的误差（如舍入误差）也会引起误差传播。这些误差传播积累在一起，往往是十分可观的。

我们仅研究方法误差，舍入误差以及误差的传播积累，模型误差和观测误差不在该课程研究的范围内。本章将简要介绍舍入误差及误差的传播与积累。

舍入误差往往会被人们忽视，这必须引起警惕，因为一个正确无误的计算公式，会由于舍入误差的传播积累得出极其荒谬的结论。

例 1-1：已知积分 $I_n = e^{-1}\int_0^1 x^n e^x \mathrm{d}x$，要在 8 位机上求 I_{20}。

解：

$$I_n = e^{-1}\int_0^1 x^n \mathrm{d}e^x = e^{-1}x^n e^x\Big|_0^1 - ne^{-1}\int_0^1 e^x x^{n-1}\mathrm{d}x$$

$$= 1 - nI_{n-1}$$

$I_n = 1 - nI_{n-1}$ 是一个精确的递值公式。

$$I_o = e^{-1}\int_0^1 e^x \mathrm{d}x = e^{-1}e^x\Big|_0^1 = 1 - \frac{1}{e}$$

因为计算机只能接收有限位的数，且 $e \doteq 2.718\ 281\ 83$，所以 I_o 放入机器中成了 I_o^*，这就产生了

初始误差 E_o

$$E_o = | I_o - I_o^* | \leqslant 0.5 \times 10^{-8}$$

因为

$$\begin{aligned} | E_n | &= | I_n - I_n^* | = | (1 - nI_{n-1}) - (1 - nI_{n-1}^*) | \\ &= n | I_{n-1} - I_{n-1}^* | = n(n-1) | I_{n-2} - I_{n-2}^* | \\ &= \cdots = n! | I_o - I_o^* | = n! E_o \end{aligned}$$

$$E_{20} = | I_{2o} - I_{2o}^* | = 20! E_o$$

$$20! = \sqrt{20!^2} > (20^{20})^{\frac{1}{2}} = 20^{10}$$

不妨记 $E_o = 0.5 \times 10^{-8}$

$$\text{所以 } E_{20} > 20^{10} \times 0.5 \times 10^{-8} = 2^9 \times 10^2 = 51\,200$$

如果初始一点点误差，但经过某种方法的计算，造成相当大的结果误差，则称此方法(或公式)是不稳定的。显然，以上的递推公式是不稳定的公式。

以下我们粗略地估计例 1-1 中的 I_{20}：

$$I_{20} = e^{-1} \int_0^1 e^x x^{20} \mathrm{d}x$$

因为

$$\max_{0 \leqslant x \leqslant 1} e^x = e, \qquad \min_{0 \leqslant x \leqslant 1} e^x = e^0 = 1$$

$$I_{20} < e^0 \int_0^1 x^{20} \mathrm{d}x = 0.047\,619\,04$$

$$I_{20} > e^{-1} \int_0^1 x^{20} \mathrm{d}x = 0.017\,518\,068$$

若取平均值，$I_{20} = e^{-1}\left(\frac{1+e}{2}\right) \int_0^1 x^{20} \mathrm{d}x = 0.032\,568\,557$

$$\text{所以 } | E_{20} | < 0.015\,050\,49$$

这种方法比用计算机递推计算造成的误差还小。

如果用估计得到的 I_{20}，利用 $I_{n-1} = \frac{1}{n}(1 - I_n)$ 公式反向递推来计算 I_{10}：

因为

$$| E_{n-1} | = | I_{n-1} - I_{n-1}^* | = \left| \frac{1}{n}(1 - I_n) - \frac{1}{n}(1 - I_n^*) \right| = \frac{1}{n} | E_n |$$

$$| E_{n-2} | = \frac{1}{n \cdot n - 1} | E_n |$$

$$\cdots$$

所以

$$| E_{10} | = \frac{1}{20 \cdot 19 \cdots \cdot 11} | E_{20} | = \frac{10!}{20!} | E_{20} | < 0.225 \times 10^{-13}$$

尽管初始误差较大，但经过递推求得的 I_{10} 精度却相当高，远比由 I_0 推出 I_{10} 要好得多，反向递推公式是稳定公式。由此可见，方法的好坏，对误差传播积累很有关系，不能掉以轻心。

又如著名的“蝴蝶效应”问题，通俗的说法是“纽约的一只蝴蝶翅膀一拍，风和日丽的北京就刮起台风来了”。实际问题是由大气运动所遵从的一系列物理学定律引发出的，并可抽象出一个庞大的复杂方程组。1921 年，气象学家 Richardson 曾试图用这样的方程组作“数值天气预报”，即由初始时刻的风速、压力、温度等数值出发，一步一步地推出以后的天气情况。若用 x_0 表示初始天气变量，则 n 步以后的天气为 $x_{n+1} = f(x_n)$，$(n = 0,1,2,\cdots)$，其中 f 为非线性函数。在初始时刻，蝴蝶翅膀一拍造成空气的风速或温度在第 1000 位小数上只差一点，而这两个只差一点的初始值推算了 1000 步后，使得初始值中的有效信息全部丧失。这就是计算不稳定性带来的麻

烦,可见误差的传播与积累是个不可忽视的问题。

1.2 误差与有效数字

为了比较方法的好坏及衡量一个结果的精确程度,给出了绝对误差、相对误差和有效数字的概念。

某一个量的准确值,称为真值,记作 x,其近似值记为 x^*。

1)绝对误差:误差本身的大小。

记为 $$e^* = x^* - x$$

e^* 称为近似数 x^* 的绝对误差。

由于 x 的准确值一般无法知道,所以 e^* 实际上也就无法求得,不过 e^* 的大小范围一般可以估计,若

$$|e^*| = |x^* - x| \leqslant \varepsilon^*$$

则 ε^* 是 e^* 的绝对值的上界,称为绝对误差限。

e^* 可以是正的,也可以是负的,但误差限 ε^* 只能是正的。真正的绝对误差 e^* 是唯一的,而 ε^* 并不唯一。

由 $|x^* - x| \leqslant \varepsilon^*$ 可得 $x^* - \varepsilon^* \leqslant x \leqslant x^* + \varepsilon^*$,工程上常这样表示一个数 $x = x^* \pm \varepsilon^*$,明确表示了近似数及绝对误差限。

例 1-2:测量会议室的长和宽,测得会议室长为 30m,宽为 10m,长的误差不超过 5cm,宽的误差不超过 2cm,应写成: $y(\text{长}) = 30 \pm 0.05$, $x(\text{宽}) = 10 \pm 0.02$

绝对误差及绝对误差限都是有量纲的,且必须与近似数有统一的量纲。

例 1-2 中的长与宽哪一个精度高呢?看上去 $0.02 < 0.05$,似乎宽的精度高,其实不然,为了解释该问题,引进了相对误差的概念。

2)相对误差:绝对误差与真值之比。

记为 $$e_r^* = \frac{e^*}{x} = \frac{x^* - x}{x}$$

由于真值 x 无法知道,常取 $e_r^* = \frac{e^*}{x^*}$ 作为定义,这是因为

$$\frac{e^*}{x} - \frac{e^*}{x^*} = \frac{e^*(x^* - x)}{xx^*} = \frac{e^{*2}}{xx^*} = \frac{e^{*2}}{(x^* - e^*)x^*} = \frac{\left(\frac{e^*}{x^*}\right)^2}{1 - \left(\frac{e^*}{x^*}\right)}$$

当绝对误差相对于真值来说是相当小的时候,这二种定义之差是关于 $\frac{e^*}{x^*}$ 的高阶无穷小,所以用 $e_r^* = \frac{e^*}{x^*}$ 作为定义还是合理的。

同样 $e_r^* = \frac{x^* - x}{x^*}$ 中还含有 x,也无法确切计算,由 $|e^*| \leqslant \varepsilon^*$,可得 $\varepsilon_r^* = \left|\frac{\varepsilon^*}{x^*}\right|$ 为相对误差限。

例 1-2 中: $$\varepsilon_r^*(x) = \left|\frac{\varepsilon^*(x)}{x^*}\right| = \frac{0.02}{10} = 0.002$$

$$\varepsilon_r^*(y) = \left|\frac{\varepsilon^*(y)}{y^*}\right| = \frac{0.05}{30} = 0.0016 < 0.002$$

注:$\varepsilon_r^*(x)$表示宽的相对误差限,$\varepsilon_r^*(y)$表示长的相对误差限。

因此,测得的长度精度比测得的宽度精度要高一些。

通常,是用相对误差来衡量一个近似数的精度的。但相对误差本身没有量纲,只是绝对误差与近似数的比值。

3)四舍五入规则。

i)舍入后使绝对误差限不超过其末位数的半个单位。

ii)若需舍入的部分刚好是末位的半个单位时,要使末位凑成偶数。

例 1-3:对 0.7135,0.7265,0.73251 分别取三位小数,应为:0.714,0.726,0.733。

4)有效数字,有效数。

如果近似值 x^* 的绝对误差限不超过某一位的半个单位,从该位向左数到 x^* 的第一个非零的数字,共有 n 位,则称这 n 位数字为有效数字,并说 x^* 具有 n 位有效数字。

如果表示一个数的数字全是有效数字,称此数为有效数。

例 1-4:0.2300 表示有四位有效数字,0.023 只有二位有效数字。

如果整数并非全是有效数字,可用浮点数表示,把 71600 表示成只有 4 位有效数字,可写为 $0.716\ 0\times10^5$。

5)数的浮点表示。

任一个有效数 x^* 均能表示成一个 0.1～1 之间的数与 $\pm10^m$ 的乘积(m 为任何整数)的形式。

$$x^* = \pm 10^m(\alpha_1\times10^{-1} + \alpha_2\times10^{-2} + \cdots + \alpha_n\times10^{-n}) \qquad (\alpha_1 \neq 0)$$

其中 α_1 是第一位不是零的数字,x^* 有 n 位有效数字分别为:$\alpha_1,\alpha_2,\cdots\alpha_n$。

6)有效数字与相对误差的关系。

近似数的相对误差与有效数字的多少有着密切的关系。

因为
$$x^* = \pm 10^m(\alpha_1\times10^{-1} + \alpha_2\times10^{-2} + \cdots + \alpha_n\times10^{-n}) \qquad (\alpha_1 \neq 0)$$

$$\varepsilon^*(x) = 0.5\times10^{m-n}$$

所以
$$\alpha_1\times10^{m-1} \leqslant |x^*| < (\alpha_1+1)\times10^{m-1}$$

$$\varepsilon_r^*(x) = \frac{\varepsilon^*(x)}{|x^*|} \leqslant \frac{0.5\times10^{m-n}}{\alpha_1\times10^{m-1}} = \frac{1}{2\alpha_1}\times10^{-n+1}$$

以上结果可归纳成定理:

定理 1-1:若 x^* 具有 n 位有效数字,则其相对误差限为

$$\varepsilon_r^*(x) = \frac{1}{2\alpha_1}\times10^{-(n-1)}$$

其中 α_1 为 x^* 的第一个有效数字。

反之也可归纳成定理:

定理 1-2:若 x^* 的相对误差为 $|e^*(x)| \leqslant \frac{1}{2(\alpha_1+1)}\times10^{-(n-1)}$,则 x^* 具有 n 位有效数字。

因为
$$\varepsilon_r^*(x) = \frac{1}{2(\alpha_1+1)}\times10^{-(n-1)}$$

$$\varepsilon^*(x) = |\varepsilon_r^*(x)\cdot x^*| < \frac{1}{2(\alpha_1+1)}\times10^{-(n-1)}\times(\alpha_1+1)\times10^{m-1}$$

所以
$$= 0.5\times10^{m-n}$$

x^* 至少有 n 位有效数字。

由上所述，有效数字的多少也反映了一个数的精度。如：0.23×10^4，0.230×10^4，2300，从数值上来说几乎一样大，但精度却不同，有效数字多的精度高。0.02300 与 2300 绝对误差限不同，但由于有效数字一样多，相对误差一样，精度却是相同的。

1.3　函数的误差估计

对于函数 $y=f(x)\quad x\in[a,b]$，如果自变量 x 被近似值 x^* 代替，那么函数值 $f(x)$ 被 $f(x^*)$ 代替，将如何估计 $f(x^*)$ 的精度呢？由微分中值定理知：

$$e^*(f(x))=f(x^*)-f(x)=f'(\xi)e^*(x)\qquad(\text{其中 }\xi\text{ 在 }x\text{ 与 }x^*\text{ 之间})$$

$$|e^*(f(x))|=|f'(\xi)||e^*(x)|\approx|f'(x^*)||e^*(x)|$$

$|f'(x^*)|$ 可看成是函数的绝对误差关于自变量的绝对误差的放大数，称为绝对误差条件数。如果条件数较小，函数的绝对误差可以被控制在某个范围内，则称函数 $f(x)$ 为好条件的，如果 $|f'(x^*)|$ 在某点 x_0 的值很大，则称函数 $f(x)$ 在 x_0 这点在绝对误差意义下是坏条件的。

因为
$$|e_r^*(f(x))|=\left|\frac{e^*(f(x))}{f(x^*)}\right|=\left|\frac{f(x^*)-f(x)}{f(x^*)}\right|$$

$$=\left|\frac{f(x^*)-f(x)}{x^*-x}\right|\left|\frac{x^*-x}{x^*}\right|\left|\frac{x^*}{f(x^*)}\right|$$

$$\approx|f'(x^*)||e_r^*(x)|\left|\frac{x^*}{f(x^*)}\right|=\left|\frac{x^*f'(x^*)}{f(x^*)}\right||e_r^*(x)|$$

所以 $\left|\dfrac{x^*f'(x^*)}{f(x^*)}\right|$ 被称为相对误差条件数。

对于多元函数 $y=f(x_1,x_2,\cdots,x_n)$，有全微分公式

$$\mathrm{d}y=\frac{\partial f}{\partial x_1}\mathrm{d}x_1+\frac{\partial f}{\partial x_2}\mathrm{d}x_2+\cdots+\frac{\partial f}{\partial x_n}\mathrm{d}x_n$$

如以增量代替微分，即 $\Delta y\approx\mathrm{d}y$，$\Delta x_i\approx\mathrm{d}x_i$，此处 $\Delta y=y^*-y=e^*(y)$，$\Delta x_i=x_i^*-x_i=e^*(x_i)$，可得近似公式：

因为
$$e^*(y)\approx\frac{\partial f}{\partial x_1}\Big|_{x_1^*,x_2^*,\cdots,x_n^*}e^*(x_1)+\frac{\partial f}{\partial x_2}\Big|_{(x_1^*,x_2^*,\cdots,x_n^*)}e^*(x_2)+\cdots$$

$$+\frac{\partial f}{\partial x_n}\Big|_{(x_1^*,x_2^*,\cdots,x_n^*)}e^*(x_n)$$

$$|e^*(y)|\approx|f_{x_1}e^*(x_1)+f_{x_2}e^*(x_2)+\cdots+f_{x_n}e^*(x_n)|$$

由柯西不等式 $|(a,b)|\leqslant|a||b|$，其中 a,b 为向量得：

$$\leqslant\sqrt{f_{x_1}^2+f_{x_2}^2+\cdots+f_{x_n}^2}\cdot\sqrt{e_{(x_1)}^{*2}+e_{(x_2)}^{*2}+\cdots+e_{(x_n)}^{*2}}$$

所以称 $\sqrt{f_{x_1}^2+f_{x_2}^2+\cdots+f_{x_n}^2}$ 为绝对误差条件数，其中 f_{x_i} 表示函数 $f(x)$ 关于 x_i 的在 $(x_1,x_2,\cdots x_n)$ 点上的偏导数值。

同样，

因为
$$|e_r^*(y)|=\left|\frac{e^*(y)}{f^*}\right|=\left|\frac{f_{x_1}e^*(x_1)}{f^*}+\frac{f_{x_2}e^*(x_2)}{f^*}+\cdots+\frac{f_{x_n}e^*(x_n)}{f^*}\right|$$

$$=\left|\frac{x_1^*f_{x_1}}{f^*}\cdot\frac{e^*(x_1)}{x_1^*}+\frac{x_2^*f_{x_2}}{f^*}\cdot\frac{e^*(x_2)}{x_2^*}+\cdots+\frac{x_n^*f_{x_n}}{f^*}\cdot\frac{e^*(x_n)}{x_n^*}\right|$$

$$\leqslant\sqrt{\left(\frac{x_1^* f_{x_1}}{f^*}\right)^2+\left(\frac{x_2^* f_{x_2}}{f^*}\right)^2+\cdots+\left(\frac{x_n^* f_{x_n}}{f^*}\right)^2}\cdot\sqrt{e_r^{*2}(x_1)+e_r^{*2}(x_2)+\cdots+e_r^{*2}(x_n)}$$

所以称$\sqrt{\left(\frac{x_1^* f_{x_1}}{f^*}\right)^2+\left(\frac{x_2^* f_{x_2}}{f^*}\right)^2+\cdots+\left(\frac{x_n^* f_{x_n}}{f^*}\right)^2}$为相对误差条件数，其中$f^*$表示函数$f(x_1^*\ x_2^*\cdots x_n^*)$。当自变量为一个时，正是$y=f(x)$时的情况。

例 1-5：要使$\sqrt{20}$的近似值的相对误差限小于0.1%，要取几位有效数字？若求$\sqrt{B}$，$B\approx 20$，B需取几位有效数字，才能保证使$\sqrt{B}$的相对误差小于0.1%？

解：1)$\sqrt{20}\doteq 4.472$　　$\alpha_1=4$

由定理 1-1 知，$\varepsilon_r^*=\frac{1}{2\times 4}\times 10^{-(n-1)}<0.001=10^{-3}$

n 取 4 能满足上式，应取 4 位有效数字。

2)设 $y=\sqrt{B}$

方法一：

$$y'=\frac{1}{2\sqrt{B}},\qquad y'|_{B=20}=\frac{1}{2\sqrt{20}}$$

$$\text{相对误差条件数}=\frac{B^* y'|_{B=20}}{y^*}=\frac{20\dfrac{1}{2\sqrt{20}}}{\sqrt{20}}=\frac{1}{2}$$

$$\varepsilon_r^*(y)\doteq\frac{1}{2}\varepsilon_r^*(B)$$

$$\varepsilon_r^*(B)=2\varepsilon_r^*(y)=2\times 10^{-3}$$

注意到 B 的 α_1 有可能是 1，也可能是 2，应考虑最坏的情况，取 $\alpha_1=1$。要寻找一个 n，使

$$\varepsilon_r^*(B)=\frac{1}{2\times 1}\times 10^{-(n-1)}<2\times 10^{-3}\qquad \text{成立。}$$

可知当 n 取 4 时能满足上式，所以 B 应取 4 位有效数字。

方法二：可以利用绝对误差条件数来解答。欲使$\sqrt{B}$相对误差限小于 10^{-3}

则应使　　$|e^*(y)|=|y^* e_r^*(y)|\leqslant|y^*||e_r^*(y)|=\sqrt{20}\times 10^{-3}$

且　　$|e^*(y)|\doteq y'|_{B=20}|e^*(B)|$

为了满足题设，希望$|e^*(B)|\leqslant\frac{1}{y'|_{B=20}}|e^*(y)|=2\sqrt{20}\cdot\sqrt{20}\times 10^{-3}=40\times 10^{-3}$。可见若 B 取 4 位有效数字($B\doteq 20$)，则 $\varepsilon^*(B)=0.5\times 10^{-2}<0.4\times 10^{-1}$能保证$\sqrt{B}$的相对误差限小于 10^{-3}。

1.4　近似数的四则运算及数值计算中需注意的几个问题

近似数的四则运算可以看成是多元函数的特殊情况。

1)加、减运算。

$$y=x_1+x_2=f(x_1,x_2)$$

$$f_{x_1}=\frac{\partial f}{\partial x_1}=1,\qquad f_{x_2}=\frac{\partial f}{\partial x_2}=1$$

$$\mathrm{d}y=\mathrm{d}x_1+\mathrm{d}x_2$$

$$|e^*(y)|=|e^*(x_1)+e^*(x_2)|\leqslant|e^*(x_1)|+|e^*(x_2)|$$

$$\leqslant \varepsilon^*(x_1)+\varepsilon^*(x_2)$$

这个结果表明和差的绝对误差不超过二数的绝对误差限之和。且

$$|e^*(y)| \leqslant \sqrt{1^2+1^2}\sqrt{e^{*2}(x_1)+e^{*2}(x_2)} = \sqrt{2}\sqrt{e^{*2}(x_1)+e^{*2}(x_2)}$$

即绝对误差条件数是$\sqrt{2}$,而相对误差条件数为

$$\sqrt{\left(\frac{x_1^*}{x_1^*+x_2^*}\right)^2+\left(\frac{x_2^*}{x_1^*+x_2^*}\right)^2} = \frac{\sqrt{x_1^{*2}+x_2^{*2}}}{x_1^*+x_2^*}$$

当 $x_1^*+x_2^*=0$ 时,条件数为∞,达到相对误差坏条件,因此,要避免二个绝对值差不多的数相互抵消,否则对运算结果的精度影响很大。

2)乘、除运算。$y=x_1\cdot x_2, y=x_1/x_2$

二边取对数　$\ln y=\ln x_1 \pm \ln x_2$

二边取微分　$\mathrm{d}\ln y=\mathrm{d}\ln x_1 \pm \mathrm{d}\ln x_2$

因为　$\mathrm{d}\ln x=\frac{1}{x}\mathrm{d}x \doteq \frac{e^*(x)}{x}=e_r^*(x)$

$$e_r^*(y)=e_r^*(x_1)\pm e_r^*(x_2)$$

所以　$|e_r^*(y)|=|e_r^*(x_1)|+|e_r^*(x_2)| \leqslant \varepsilon_r^*(x_1)+\varepsilon_r^*(x_2)$

即乘除法中,积(或商)的相对误差不超过二数相对误差限之和。且注意到

$$y=\frac{x_1}{x_2}$$

$$\mathrm{d}y=\frac{x_2\mathrm{d}x_1-x_1\mathrm{d}x_2}{x_2^2}=\frac{\mathrm{d}x_1}{x_2}-\frac{x_1}{x_2^2}\mathrm{d}x_2$$

$$|e^*(y)| = \left|\frac{1}{x_2}e^*(x_1)-\frac{x_1}{x_2^2}e^*(x_2)\right|$$

$$\leqslant \frac{1}{|x_2|}|e^*(x_1)|+\left|\frac{x_1}{x_2^2}\right||e^*(x_2)|$$

若 x_2 绝对值很小,有可能导致绝对误差很大,因此,要避免除数接近于 0。

在加减法运算中,还要注意不要把绝对值差异很大的数做加减法,看一个例子:

例 1-6:求 $x^2+(\alpha+\beta)x+10^9=0$ 的根,其中 $\alpha=-10^9, \beta=-1$

解:用二次方程求根公式:$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$

$a=1$,　$b=\alpha+\beta=-0.1\times10^{10}-0.000\ 000\ 000\ 1\times10^{10}$

$=-0.1\times10^{10}$(设用八位机运算)

$c=10^9$

$\Rightarrow b^2-4ac=10^{18}-4\times1\times10^9$

$=0.1\times10^{19}-0.000\ 000\ 000\ 4\times10^{19}$

$=0.1\times10^{19}=10^{18}$

$$\Rightarrow x_{1,2}=\frac{10^9\pm10^9}{2}=\begin{cases}10^9\\0\end{cases}$$

而此方程的根应为 $x_1=10^9, x_2=1$。计算机所得结果中的一个根结果很好,另一个根结果很糟,原因是进行加减法时要对阶,大数吃掉了小数,而且又出现了二个大小相同的数相减,丧失了大量有效数字,因此,x_2 的精度很差。但若换一种方法,利用韦达定理:

$$\begin{cases} x_1 + x_2 = -\dfrac{b}{a} \\ x_1 x_2 = \dfrac{c}{a} \end{cases}$$

$$\Rightarrow \begin{cases} x_1 = \dfrac{-b - \mathrm{sign}(b)\sqrt{b^2 - 4ac}}{2a} = 10^9 \\ x_2 = \dfrac{c}{ax_1} = \dfrac{10^9}{10^9} = 1 \end{cases}$$

这个结果就很好了。此方程避免了二个差不多大的数相抵消，并没避免大数吃小数，但这种方法大大减小了大数吃小数产生的影响。

从舍入误差的积累方面考虑，减少运算次数，简化步骤，也会减少误差的积累并提高速度。如：

$$x^{255} = x \cdot x^2 \cdot x^4 \cdot x^8 \cdot x^{16} \cdot x^{32} \cdot x^{64} \cdot x^{128}$$

把 254 次乘法化为 14 次乘法。

归纳以上的内容，数值计算中需要注意以下几个问题：

1)避免二个相近的数相减(代数和接近于 0)。

2)避免除数绝对值太小。

3)避免绝对值相差很大的二个数进行加减运算。

4)尽可能减少运算步骤。

5)算法或公式要稳定。

习　题

1. 序列$\{y_n\}$满足递推关系 $y_{n+1} = 100.01y_n - y_{n-1}$，取 $y_0 = 1, y_1 = 0.01$ 及 $\overline{y_0} = 1 + 10^{-6}, \overline{y_1} = 0.01$，试分别计算 $y_5, \overline{y_5}$，从而说明递推公式对于计算是不稳定的。

2. 取 $y_0 = 28$，按递推公式 $y_n = y_{n-1} - \dfrac{1}{100}\sqrt{783}$，去计算 y_{100}，若取$\sqrt{783} \approx 27.982$(五位有效数字)，试问计算 y_{100}将有多大误差？y_{100}中尚留有几位有效数字？

3. 函数 $f(x) = \ln(x - \sqrt{x^2 - 1})$，求 $f(30)$的值。若开平方用六位函数表，问求对数时误差有多大？若改用另一等价公式 $\ln(x - \sqrt{x^2 - 1}) = -\ln(x + \sqrt{x^2 - 1})$计算，求对数时误差有多大？

4. 下列各数都按有效数字给出，试估计 f 的绝对误差限和相对误差限。

1)$f = \sin[(3.14)(2.685)]$

2)$f = (1.56)^{3.414}$

5. 计算 $f = (\sqrt{2} - 1)^6$，取$\sqrt{2} \doteq 1.4$，利用下列等式计算，哪一个得到的结果最好，为什么？

1)$\dfrac{1}{(\sqrt{2}+1)^6}$，　2)$(3 - 2\sqrt{2})^3$，　3)$\dfrac{1}{(3+2\sqrt{2})^3}$，　4)$99 - 70\sqrt{2}$

6. 下列各式怎样计算才可减小误差？

1)$\sqrt{1+x} - \sqrt{x}$

2)$\ln x_1 - \ln x_2$　　$(x_1 \approx x_2)$

3)$\dfrac{1 - \cos x}{\sin x}$　　(x 在 0 附近)

4)$\sin(x+\varepsilon)-\sin x$

5)$\int_N^{N+1}\frac{dx}{1+x^2}$

7. 求方程 $x^2-56x+1=0$ 的两个根,问要使它们具有四位有效数字,$\Delta=\sqrt{b^2-4ac}$ 至少要取几位有效数字? 如果利用韦达定理,Δ 又该取多少位有效数字呢?

8. 求近似数 x^* 的幂 x^{*n} 的相对误差公式。

9. 设 $y=\lg x$,证明 x^* 的常用对数的绝对误差不超过 x^* 的相对误差的一半,反之,y^* 的反对数的相对误差不超过 y^* 的绝对误差的 2.5 倍。

10. 简化求多项式 $y=a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0$ 的值的运算式。

第2章　非线性方程求根

在人口增长模型中假设某一地区人口的数量随时间连续增长，其增长率与人口总数成正比，即$\frac{dN(t)}{dt}=\lambda N(t)$，其中 $N(t)$表示该地区在 t 时刻的人口总数，λ 为人口增长率常数。该微分方程的解为 $N(t)=N_0e^{\lambda t}$，其中 N_0 为该地区的初始人口总数。但上述模型没有考虑到有外部移民迁入的情况，若允许移民以某常速率 v 进入该地区，则微分方程是$\frac{dN(t)}{dt}=\lambda N(t)+v$，其解为 $N(t)=N_0e^{\lambda t}+\frac{v}{\lambda}(e^{\lambda t}-1)$。

现假设某地区有 100 万初始人口，第一年内有 43.5 万移民迁入，第一年末总计人口达 156.4 万。根据上述数据推算该地区的增长率常数 λ，即要求解方程：

$$156.4=100e^{\lambda}+\frac{43.5}{\lambda}(e^{\lambda}-1)$$

直接通过代数方法求解以上方程显然是不可能的。在这一章里，就将讨论如何计算这一类问题的近似解的方法。（上述问题的具体解法作为习题 2-4 的题 11，留给读者去完成）。

首先复习几个定义和重要的结论：

定义 2-1：如果有一个数 x^*，使 $f(x^*)=0$，则称 x^* 为方程 $f(x)=0$ 的根，或称 x^* 为函数 $f(x)$的零点。

定义 2-2：如果 $f(x)=(x-x^*)^mg(x)$，其中 m 为正整数，$g(x)$的分母不含因子$(x-x^*)$，且 $g(x^*)\neq0$，则称 x^* 为 $f(x)$的 m 重零点，x^* 为 $f(x)=0$ 的 m 重根。当 $m=1$ 时，称 x^* 为单根。

定义 2-3：若 $f(x)=a_0x^n+a_1x^{n-1}+\cdots+a_n(a_0\neq0)$，则称 $f(x)=0$ 为 n 次代数方程。当 $n=1$ 时，此方程称为线性方程。

当 $f(x)=0$ 不是代数方程时，就称为超越方程。

非线性方程包括 $n\neq1$ 的代数方程（高次方程）及超越方程。

下面列出几个重要的结论：

1）在$[a,b]$上连续的函数 $f(x)$，若 $f(a)f(b)<0$，则存在一实数 $x^*\in(a,b)$，满足 $f(x^*)=0$。

2）$f(x)=0$ 在(a,b)上有根，且 $f'(x)$在(a,b)中不变号，且不为 0，则 $f(x)=0$ 在(a,b)上只有唯一根。

3）n 次代数方程在复数域上恰有 n 个根（一个 r 重根算 r 个根），则实系数代数方程的复根会成对出现。

4）高于四次的代数方程没有求根公式。

2.1　多项式及代数方程根的界

2.1.1　多项式

$$P_n(x)=a_0x^n+a_1x^{n-1}+\cdots+a_n$$

若直接按此式计算 $P_n(x)$在 x_0 的值 $P_n(x_0)$,需要 $n+(n+1)+\cdots+2+1=\frac{(1+n)n}{2}$次乘法,$n$ 次加法。若改写成

$$P_n(x)=(\cdots((a_0x+a_1)x+a_2)x\cdots+a_{n-1})x+a_n$$

只需 n 次乘法和 n 次加法,节省了不少机器时间,这就是所谓豪纳(*Horner*)规则。事实上,我国古代数学家秦九韶早在 1247 年就用这程序来计算多项式的值了。

余数定理:$P_n(x_0)$等于 $P_n(x)$除以$(x-x_0)$所得之余数。

$$P_n(x)=(x-x_0)(b_0x^{n-1}+b_1x^{n-2}+\cdots+b_{n-2}x+b_{n-1})+R=(x-x_0)g(x)+R$$

$$P_n(x_0)=R$$

$$\begin{aligned}P_n(x)&=b_0x^n+(b_1-b_0x_0)x^{n-1}+(b_2-b_1x_0)x^{n-2}+\cdots\\&\quad+(b_{n-1}-b_{n-2}x_0)x+(R-b_{n-1}x_0)\\&=a_0x^n+a_1x^{n-1}+a_2x^{n-2}+\cdots+a_{n-1}x+a_n\end{aligned}$$

对照系数得:

$$\begin{aligned}&b_0=a_0 && b_0=a_0\\&b_1-b_0x_0=a_1 && b_1=a_1+b_0x_0\\&b_2-b_1x_0=a_2 && b_2=a_2+b_1x_0\\&\vdots && \vdots\\&b_{n-1}-b_{n-2}x_0=a_{n-1} && b_{n-1}=a_{n-1}+b_{n-2}x_0\\&R-b_{n-1}x_0=a_n && R=a_n+b_{n-1}x_0\xlongequal{\text{记成}}b_n\end{aligned}$$

这就是有名的秦九韶程序:

$$R=b_n=a_n+x_0(a_{n-1}+x_0(a_{n-2}+\cdots+x_0(a_1+x_0a_0)\cdots))$$

本质上与豪纳(*Horner*)规则是一致的。

下面介绍一种手算的格式来计算 $P_n(x_0)$:

a_0+	a_1+	a_2+	$\cdots$	$+a_{n-1}$	$+a_n$	x_0
	b_0x_0+	b_1x_0+	$\cdots$	$+b_{n-2}x_0$	$+b_{n-1}x_0$	
a_0+	$a_1+b_0x_0+$	$(a_2+b_1x_0)+$	$\cdots$	$+(a_{n-1}+b_{n-2}x_0)+$	$(a_n+b_{n-1}x_0)$	
$\|\|$	$\|\|$	$\|\|$		$\|\|$	$\|\|$	
b_0	b_1	b_2	$\cdots$	b_{n-1}	$R=P_n(x_0)$	

例 2-1:$P_3(x)=x^3-7x^2+7x+15$,求 $P_3(-1)$

解:用以上手算格式得:

1	−7	+7	+15	−1
	−1	+8	−15	
1	−8	+15	+0	

$P_3(-1)=0$,也说明 -1 恰是 $P_3(x)=0$ 的一个根,且横线下的数恰为 $g(x)$的系数,即可知:

$$x^3-7x^2+7x+15=(x+1)(x^2-8x+15)$$

可见此方法可用来验证方程的根,以及降阶再求其他的根。由于 $x^2-8x+15=0$ 的根是 3,5,可得:

$$x^3-7x^2+7x+15=(x+1)(x-3)(x-5)$$

从以上方法可总结出用计算机求多项式的值，可通过以下的函数过程：

$$\text{PoLy}(n, A, x_0)$$

n 为多项式的最高次项次数，$A[0:n]$为存放多项式系数下的数组，x_0 是 $P_n(x_0)$中的 x_0。

框图如下：

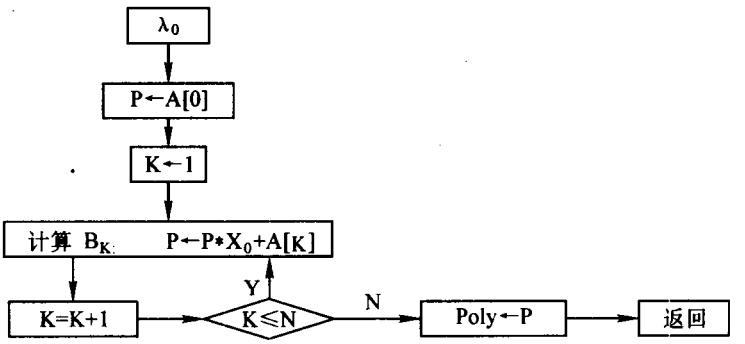

图　2-1

求多项式的值 $P_n(x_0)$，只要对 PoLy 的形参 n, A, X_0 赋值后调用这个函数过程即可。

2.1.2　代数方程根的界

定理 2-1：笛卡儿符号法则

实系数代数方程 $f(x)=0$ 的正实根个数或者恰为 $f(x)$中系数列的变号数，或者为比 $f(x)$系数列变号数少一个的正偶数。（证明略）

推论 1：$f(x)=0$ 的负实根个数，为 $f(-x)$的系数列变号数或比之少一个的正偶数。

推论 2：如果 $f(x)$没有负系数，则 $f(x)=0$ 没有正根。

例 2-2：确定 $f(x)=x^4+15x^3-2x^2-1=0$ 的正实根与负实根个数。

解：因为{1,15,−2,0,−1}只有一个变号，所以必有一个正根。

$f(-x)=x^4-15x^3-2x^2-1=0$

{1,−15,−2,0,−1}也只有一个变号，所以必有一个负根。由代数基本定理知四次方程共有 4 个复根，因此还有一对共轭虚根。

定理 2-2：设 $f(x)=a_0x^n+a_1x^{n-1}+\cdots+a_{n-1}x+a_n=0$ 是实系数代数方程，记 $A=\max\{|a_1|,|a_2|,\cdots|a_n|\}$，则 $f(x)=0$ 的一切根 $x_K(K=1,2,\cdots n)$满足

$$|x_K|<1+\frac{A}{|a_0|}=R$$

即 $f(x)=0$ 的所有复根都分布在复平面上的某一圆内，此圆以原点为圆心，$1+\frac{A}{|a_0|}$为半径，事实上，$1+\frac{A}{|a_0|}$是一个根模上界。（证明略）

推论 1：根模下界 $|x|>\dfrac{1}{1+\dfrac{B}{|a_n|}}=r$，其中

$|B|=\max\{|a_0|,|a_1|,\cdots,|a_{n-1}|\}$。

如果 $f(x)=0$ 有零根，$f(x)=xg(x)=x(a_0x^{n-1}+\cdots+a_{n-2}x+a_{n-1})$，则 a_n 只能是 0，容

易得出此时 $r=0$,正是 $a_n \to 0$ 的极限状态。

定理 2-3:设 $f(x)=a_0x^n+a_1x^{n-1}+\cdots+a_n=0$

1)$a_0>0$。

2)$a_K(K\geqslant 1)$是第一个负系数(K 是下标)。

3)$D=\{f(x)$的所有负系数的绝对值的最大值$\}$。

则 $f(x)=0$ 的根 $x<1+\sqrt[K]{\dfrac{D}{a_0}}$(正根上界)。

推论 1:$f(-x)=0$ 的正根上界可作为 $f(x)=0$ 的负根下界的相反数。

例 2-3:求 $f(x)=x^3-3x+1=0$ 的实根所在区间。

解 $a_0=1,K=2,D=3,f(x)=0$ 的正根上界为 $1+\sqrt{\dfrac{3}{1}}=2.732$

$f(-x)=-x^3+3x+1=0$,由于要求 $a_0>0$,改写

$f(-x)=0$ 为 $-g(x)=0,g(x)=x^3-3x-1=0$

针对上式:$a_0=1,K=2,D=3$,所以 $g(x)=0$ 的正根上界为 2.732,由推论 1 知 $f(x)=0$ 的负根下界为 -2.732。

由笛卡儿符号规则知,因为 $f(x)$的系数列变号个数为 2,所以 $f(x)$或有二个正根,或一个正根也没有;因为 $f(-x)$系数变号只有一个,则 $f(-x)$必有一个负根。

x	-2.732	-2	-1	0	1	2	2.732
$f(x)$	$-$	$-$	$+$	$+$	$-$	$+$	$+$

由重要结论 1 知,三个实根分别在$(-2,-1),(0,1),(1,2)$区间内,由于三次方程只有三个根,所以不存在别的复根了。

由于超越方程根的范围讨论起来很复杂,只能就具体情况具体分析。在解题中经常使用重要结论 1 来确定某些方程的某些求根区间。

习　题　2-1

1. 试寻找 $f(x)=x^3+6.6x^2-29.05x+22.64=0$ 的实根上、下界,及正根所在区间,区间长度取为 1。

2. 你能不利用多项式的求导公式,而借鉴于余数定理的思想,构造出求多项式:$P_n(x)=a_0x^n+a_1x^{n-1}+\cdots+a_{n-1}x+a_n$ 在 x_0 这点上的导数值 $P'_n(x_0)$的算法和框图吗?

2.2　二分法(分半法)

设 $f(x)\in C[a,b]$,且 $f(a)f(b)<0$,用二分法在$[a,b]$区间寻找实根。记$[a_1,b_1]=[a,b]$

取 $x_1=\dfrac{a_1+b_1}{2}$是$[a_1,b_1]$的中点,若 $f(x_1)=0$,则 x_1 是 $f(x)=0$ 的根,若 $f(x_1)f(a_1)>0$,则取 $a_2=x_1,b_2=b_1$,否

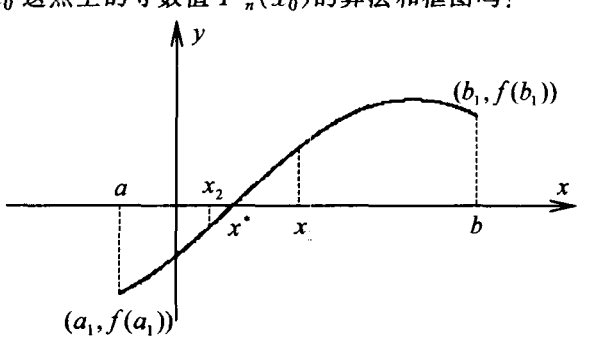

图　2-2

则取 $a_2=a_1, b_2=x_1$，得到$[a_2,b_2]$，满足：

1) $f(a_2)f(b_2)<0$。

2) $b_2-a_2=\frac{1}{2}(b_1-a_1)=\frac{1}{2}(b-a)$。

3) $a_2\geqslant a_1, b_2\leqslant b_1$。

以$[a_2,b_2]$取代$[a_1,b_1]$，继续以上过程，得$[a_3,b_3]$，…，直到某一 x_K 时，$x_K=\frac{a_K+b_K}{2}$，使 $f(x_K)=0$，则 x_K 是 $f(x)=0$ 的根，若不存在这样的 x_K，能得到一系列闭区间：

$[a_1,b_1],[a_2,b_2],\cdots[a_K,b_K],\cdots,K=1,2,\cdots$满足：

1) $f(a_K)f(b_K)<0$。

2) $b_K-a_K=\frac{1}{2^{K-1}}(b-a)$。

3) $a_1\leqslant a_2\leqslant\cdots\leqslant a_K\leqslant\cdots, b_1\geqslant b_2\geqslant\cdots\geqslant b_K\cdots$。

表明 $\exists x^*, f(x^*)=0, x^*\in(a_K,b_K)$，因此，$\{a_K\}$单调上升，有上界 x^*，$\{b_K\}$单调下降，有下界 x^*，且这二个序列均存在极限：

因为：$\lim\limits_{K\to\infty}a_K\leqslant x^*$，　$\lim\limits_{K\to\infty}b_K\geqslant x^*$

$$\lim_{K\to\infty}b_K-\lim_{K\to\infty}a_K=\lim_{K\to\infty}(b_K-a_K)=\lim_{K\to\infty}\frac{1}{2^{K-1}}(b-a)=0$$

$$\lim_{K\to\infty}b_K=\lim_{K\to\infty}a_K=x^*$$

$$\lim_{K\to\infty}x_K=\lim_{K\to\infty}\frac{a_K+b_K}{2}=\frac{1}{2}\lim_{K\to\infty}a_K=\frac{1}{2}x^*+\frac{1}{2}x^*=x^*$$

且　$|x_K-x^*|=\left|\frac{a_K+b_K}{2}-x^*\right|\leqslant\frac{1}{2}(b_K-a_K)=\frac{1}{2^K}(b-a)$

所以得到定理：

定理 2-4：$f(x)\in C[a,b]$，且 $f(a)f(b)<0$，则由二分法产生的序列$\{x_K\}$收敛于 $f(x)=0$ 的一个根 x^*，且

$$|x_K-x^*|\leqslant\frac{b-a}{2^K}$$

对于指定的精度 ε，可以估计二分法执行次数 K。

$$\frac{b-a}{2^K}<\varepsilon$$

两边取对数：　　$\ln(b-a)-K\ln 2<\ln\varepsilon$

$$K>\frac{1}{\ln 2}[\ln(b-a)-\ln\varepsilon]=\frac{\ln\frac{b-a}{\varepsilon}}{\ln 2}$$

取　$K=[\frac{\ln\frac{b-a}{\varepsilon}}{\ln 2}]+1$

为了节省计算机内存空间，用工作单元 a 存放 a_K，b 存放 b_K，x 存放 x_K。为了避免无休止地做下去，给出二个适当小的正数 $\varepsilon_1,\varepsilon_2$。当$|f(x)|<\varepsilon_1$ 时，x 就作为近似解，应当说明的是，这样的停机条件虽然比较实用，但不能保证 x 的精度。当 $f(x)$在根的附近变化非常缓慢时，有可能出现$|f(x)|$很小，而 x 距离精确解还有相当距离的情况。如图 2-3 所示，该图是 $y=f(x)=$

$(x-1)^{11}$的图象，取 $\varepsilon=0.001$，从图中可见 $|f(1.5)|<\varepsilon$，然而 $x=1.5$ 距离精确解 $x^*=1$ 还有较大误差。

当 $\left|\frac{b-a}{2}\right|<\varepsilon_2$ 时，$x=\frac{a+b}{2}$就作为近似解。也可给出一个适当大的正整数 N，若执行次数超过 N，还不能满足精度要求，就给出一个出口，显示不收敛信息。而二分法是必然收敛的，实际情况是收敛太慢，或 $\varepsilon_1,\varepsilon_2,N$ 给得不适当，这时可以适当调整这些参数，再用二分法求根，就有可能得到结果了。

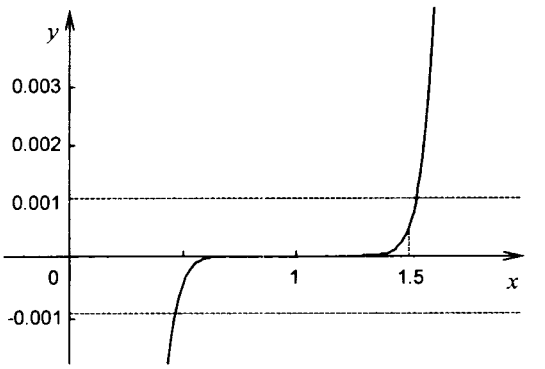

图　2-3

设 BM 为二分法的函数过程：

具体实现如下：

BM($a,b,f,eps1,eps2$)

FOR $K=1,2,\cdots,N$ DO

$$\begin{cases}1)\ x\leftarrow\dfrac{a+b}{2}\\ 2)\ xe\leftarrow\dfrac{b-a}{2}\\ 3)\ If\,|xe|<eps2\ or\ |f(x)|<eps1\ then\ output\ K,x,f(x)\ STOP\\ 4)\ If\quad f(a)\cdot f(x)<0\ then\ b\leftarrow x\ else\ a\leftarrow x\end{cases}$$

output "method failed after N iterations."

二分法优点：

1)程序简单。

2)对 $f(x)$要求不高，收敛性好。

二分法缺点：

1)收敛速度不快。

2)对偶重根无法求，也无法求复根。

3)调用一次 BM，只能求出一个实根。

当 $f(x)\in C[a,b]$，$f(a)\cdot f(b)<0$，则$[a,b]$内必有根，但可以不止一个根，如图 2-4 所示：当用二分法求根时，由于 $x_1,x_2,\cdots$分别取代了端点，得到的只是 x^* 一个根，其余根都失去了。要补救这个缺点是容易的，而且可以不必要求 $f(a)\cdot f(b)<0$。

可设二分法函数过程 BMZ：

BMZ($a,b,eps1,eps2,dx,K,x,f$)

其中 a,b 为实根所在区间，$eps1,eps2$ 是精度要求，dx 为步长，K 为实根个数，$x[1:n]$为存实根的数组，f 是方程 $f(x)=0$ 左端的函数过程名，具体实现如下：

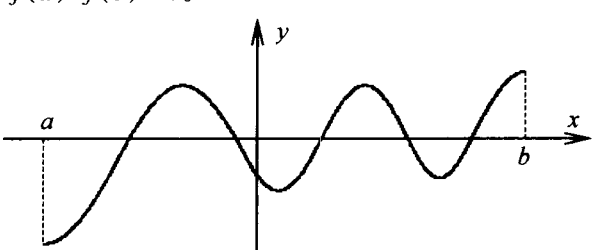

图　2-4

记 $x_i=a+i\,dx(i=0,1,2,\cdots)$，用循环语句实现条件语句：若 $f(x_{i+1})=0$，则 x_{i+1}是 $f(x)=0$ 的根，若 $f(x_i)f(x_{i+1})>0$，则不做，若 $f(x_i)f(x_{i+1})<0$，则以 x_i,x_{i+1}分别取代