第1章　预备知识

假设函数$f(x)=\cos(x)$,则它的导数$f'(x)=-\sin(x)$,不定积分为$F(x)=\sin(x)+C$。在微积分学中可以学到这些公式,前者确定函数曲线$y=f(x)$在点$(x_0,f(x_0))$的斜率$m=f'(x_0)$,后者可计算出函数曲线在$a\leqslant x\leqslant b$范围下的面积。

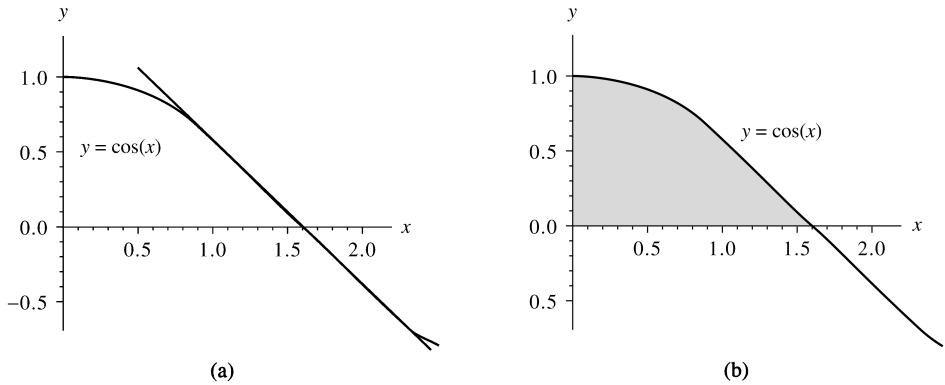
曲线y在点$(\pi/2,0)$的斜率$m=f'(\pi/2)=-1$,通过它可找到在这一点的切线[见图1.1(a)]:

$$y_{\tan}=m\left(x-\frac{\pi}{2}\right)+0=f'\left(\frac{\pi}{2}\right)\left(x-\frac{\pi}{2}\right)=-x+\frac{\pi}{2}$$

通过积分方法可以计算曲线y在$0\leqslant x\leqslant\pi/2$范围下的面积为[见图1.1(b)]:

$$\text{面积}=\int_0^{\pi/2}\cos(x)\,dx=F\left(\frac{\pi}{2}\right)-F(0)=\sin\left(\frac{\pi}{2}\right)-0=1$$

这是在微积分学中需要用到的一些结果。



(a) 函数曲线$y=\cos(x)$在点$(\pi/2,0)$的切线;(b) 函数曲线$y=\cos(x)$在区间$[0,\pi/2]$下的区域

1.1　微积分回顾

本书假定读者具有大学本科的微积分知识,即熟悉极限、连续性、求导、积分、序列和级数等微积分知识。下面回顾了本书中引用的微积分知识。

1.1.1　极限和连续性

定义1.1　设$f(x)$为定义在包含$x=x_0$的开区间上的函数,可能在$x=x_0$处无定义。如果对于任意给定的$\epsilon>0$,总存在$\delta>0$,使得当$0<|x-x_0|<\delta$时有$|f(x)-L|<\epsilon$,则称函数f在$x=x_0$处具有**极限**L,表示为

$$\lim_{x\to x_0}f(x)=L \tag{1}$$

当采用h增量表达式$x=x_0+h$时,上式可以表示为

$$\lim_{h\to 0}f(x_0+h)=L \tag{2}$$

▲

定义 1.2 设 $f(x)$ 为定义在包含 $x=x_0$ 的开区间上的函数,如果

$$\lim_{x\to x_0} f(x) = f(x_0) \tag{3}$$

则称函数 f **在点 $x=x_0$ 处连续**。

如果函数 f 在所有 $x\in S$ 上连续,则称函数 f 在集合 S 上连续。符号 $C^n(S)$ 表示函数 f 自身和它的前 n 阶导数在集合 S 上连续的所有函数 f 的集合。当 S 为区间 $[a,b]$ 时,则可以用符号 $C^n[a,b]$ 来表示。例如,若有函数 $f(x)=x^{4/3}$,其定义域为 $[-1,1]$,则显然 $f(x)$ 和 $f'(x)=(4/3)x^{1/3}$ 在区间 $[-1,1]$ 上连续,但 $f''(x)=(4/9)x^{-2/3}$ 在 $x=0$ 处不连续。 ▲

定义 1.3 设 $\{x_n\}_{n=1}^{\infty}$ 为一个无限序列。如果对于任意给定的 $\epsilon>0$,总存在一个正整数 $N=N(\epsilon)$,使得当 $n>N$ 时有 $|x_n-L|<\epsilon$,则称序列 $\{x_n\}_{n=1}^{\infty}$ 具有**极限 L**,记为

$$\lim_{n\to\infty} x_n = L \tag{4}$$

▲

当序列有极限时,则称其为**收敛序列**。另一个通常使用的表示形式为"当 $n\to\infty$ 时有 $x_n\to L$"。上式等价于

$$\lim_{n\to\infty} (x_n - L) = 0 \tag{5}$$

这样,可将序列 $\{\epsilon_n\}_{n=1}^{\infty}=\{x_n-L\}_{n=1}^{\infty}$ 看成一个**误差序列**。下列定理与连续性和收敛序列有关。

定理 1.1 设 $f(x)$ 为定义在实数集合 S 上的函数,且 $x_0\in S$,则下列命题是等价的:

(a) 函数 f 在 x_0 处连续。

(b) 如果 $\lim_{n\to\infty} x_n = x_0$,则 $\lim_{n\to\infty} f(x_n) = f(x_0)$。 (6)

定理 1.2(中值定理) 若 $f\in C[a,b]$,且 L 为 $f(a)$ 与 $f(b)$ 之间的任意值,则存在 $c\in(a,b)$,使得 $f(c)=L$。

例 1.1 函数 $f(x)=\cos(x-1)$ 在区间 $[0,1]$ 内连续,且常量 $L=0.8\in(\cos(0),\cos(1))$。函数 $f(x)=0.8$ 在区间 $[0,1]$ 的解为 $c_1=0.356499$。同样,函数 $f(x)$ 在区间 $[1,2.5]$ 内连续,且 $L=0.8\in(\cos(2.5),\cos(1))$。函数 $f(x)=0.8$ 在区间 $[1,2.5]$ 的解为 $c_2=1.643502$。这两种情况如图 1.2 所示。 ■

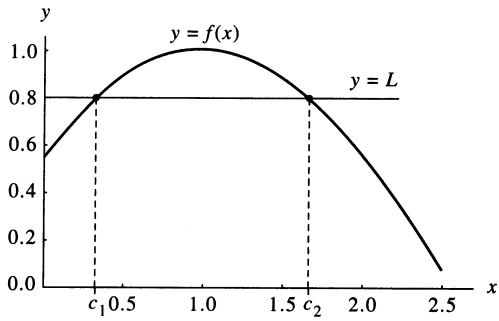


图 1.2 将中值定理运用在函数 $f(x)=\cos(x-1)$ 上,区间分别为 $[0,1]$ 和 $[1,2.5]$

定理 1.3(连续函数的极值定理) 设 $f\in C[a,b]$,则存在下界 M_1 和上界 M_2,以及 $x_1,x_2\in[a,b]$,满足

$$\text{当 } x\in[a,b] \text{ 时},M_1=f(x_1)\leqslant f(x)\leqslant f(x_2)=M_2 \tag{7}$$

有时也可以将其表示为

$$M_1=f(x_1)=\min_{a\leqslant x\leqslant b}\{f(x)\} \quad \text{和} \quad M_2=f(x_2)=\max_{a\leqslant x\leqslant b}\{f(x)\} \tag{8}$$

1.1.2 可微函数

定义 1.4 设 $f(x)$ 在一个包含 x_0 的开区间内有定义。如果

$$\lim_{x\to x_0}\frac{f(x)-f(x_0)}{x-x_0} \tag{9}$$

存在,则称函数 f 在点 x_0 处**可微**。如果此极限存在,则记为 $f'(x_0)$,并称之为 f 在点 x_0 处的**导数**。也可以采用 h 增量表达式来表示此极限:

$$\lim_{h\to 0}\frac{f(x_0+h)-f(x_0)}{h}=f'(x_0) \tag{10}$$

如果函数在集合 S 上的每一点都存在导数,则称函数在集合 S 上**可微**。要注意的是,数 $m=f'(x_0)$ 是函数曲线 $y=f(x)$ 在点 $(x_0,f(x_0))$ 的切线斜率。 ▲

定理 1.4　如果函数 $f(x)$ 在点 $x=x_0$ 处可微,则 $f(x)$ 在点 $x=x_0$ 处连续。

根据定理 1.3,如果函数 f 在闭区间 $[a,b]$ 内可微,则函数 f 的极值在闭区间的端点,或在开区间 (a,b) 的临界点,即在 $f'(x)=0$ 的解处取得。

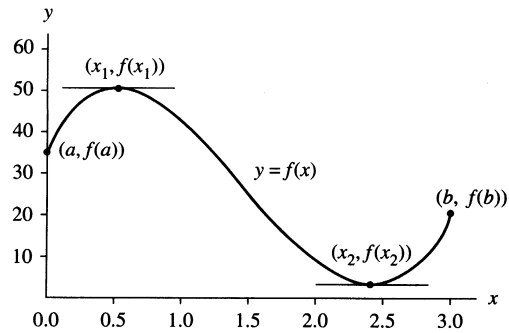


图 1.3　将极值定理运用在函数 $f(x)=35+59.5x-66.5x^2+15x^3$ 上,区间为 $[0,3]$

例 1.2　函数 $f(x)=15x^3-66.5x^2+59.5x+35$ 在区间 $[0,3]$ 内可微。$f'(x)=45x^2-123x+59.5=0$ 的解为 $x_1=0.54955$ 和 $x_2=2.40601$。如图 1.3 所示,函数 f 在区间 $[0,3]$ 内的极小值和极大值分别为

$$\min\{f(0),f(3),f(x_1),f(x_2)\}=\min\{35,20,50.10438,2.11850\}=2.11850$$

和

$$\max\{f(0),f(3),f(x_1),f(x_2)\}=\max\{35,20,50.10438,2.11850\}=50.10438$$ ■

定理 1.5[罗尔(Rolle)定理]　设 $f\in C[a,b]$,且对于所有 $x\in(a,b)$,存在导数 $f'(x)$,如果 $f(a)=f(b)=0$,则至少存在一点 $c\in(a,b)$,满足 $f'(c)=0$。

定理 1.6(均值定理)　如果 $f\in C[a,b]$,且对于所有 $x\in(a,b)$,存在导数 $f'(x)$,则至少存在一点 $c\in(a,b)$,使得下式成立:

$$f'(c)=\frac{f(b)-f(a)}{b-a} \tag{11}$$

均值定理的几何意义是:函数曲线 $y=f(x)$ 至少存在一点 $(c,f(c))$,其中 $c\in(a,b)$,该点上的切线斜率等于过点 $(a,f(a))$ 和点 $(b,f(b))$ 的割线的斜率。

例 1.3　函数 $f(x)=\sin(x)$ 在闭区间 $[0.1,2.1]$ 内连续,且在开区间 $(0.1,2.1)$ 内可微,则根据均值定理,至少存在 c,满足

$$f'(c)=\frac{f(2.1)-f(0.1)}{2.1-0.1}=\frac{0.863209-0.099833}{2.1-0.1}=0.381688$$

$f'(c)=\cos(c)=0.381688$ 在区间 $(0.1,2.1)$ 上的解为 $c=1.179174$。函数曲线 $f(x)$、割线 $y=0.381688x+0.099833$ 和切线 $y=0.381688x+0.474215$ 如图 1.4 所示。 ■

定理 1.7(广义罗尔定理)　设 $f\in C[a,b]$,$f'(x)$,$f''(x)$,$\cdots$,$f^{(n)}(x)$ 在开区间 (a,b) 内存在,且 x_0,x_1,$\cdots$,$x_n\in[a,b]$。如果当 $j=0,1,\cdots,n$ 时有 $f(x_j)=0$,则至少存在一点 $c\in(a,b)$,满足 $f^{(n)}(c)=0$。

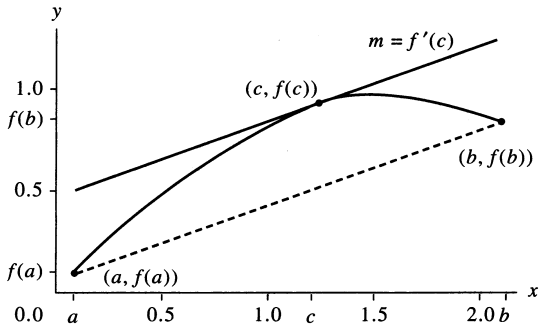


图 1.4 将均值定理运用于函数 $f(x)=\sin(x)$ 上,区间为[0.1,2.1]

1.1.3 积分

定理 1.8(第一基本定理) 如果函数 f 在区间 $[a,b]$ 内连续,且函数 F 是 f 在区间 $[a,b]$ 内的任一积分函数,则

$$\int_a^b f(x)\,\mathrm{d}x = F(b)-F(a), \qquad \text{其中 } F'(x)=f(x) \tag{12}$$

定理 1.9(第二基本定理) 如果函数 f 在区间 $[a,b]$ 内连续,且 $x\in(a,b)$,则

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_a^x f(t)\,\mathrm{d}t = f(x) \tag{13}$$

例 1.4 函数 $f(x)=\cos(x)$ 在区间 $[0,\pi/2]$ 内满足定理 1.9 的假设,则根据定理的结论可得

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_0^{x^2}\cos(t)\,\mathrm{d}t = \cos(x^2)(x^2)' = 2x\cos(x^2)$$ ■

定理 1.10(积分均值定理) 若 $f\in C[a,b]$,则至少存在一点 $c\in(a,b)$,满足

$$\frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x = f(c)$$

$f(c)$ 是函数 f 在区间 $[a,b]$ 内的平均值。

例 1.5 函数 $f(x)=\sin(x)+\frac{1}{3}\sin(3x)$ 在区间 $[0,2.5]$ 内满足定理 1.10 的假设。函数 $f(x)$ 的一个积分函数为 $F(x)=-\cos(x)-\frac{1}{9}\cos(3x)$,则函数 $f(x)$ 在区间 $[0,2.5]$ 内的平均值为

$$\frac{1}{2.5-0}\int_0^{2.5} f(x)\,\mathrm{d}x = \frac{F(2.5)-F(0)}{2.5} = \frac{0.762629-(-1.111111)}{2.5}$$
$$= \frac{1.873740}{2.5} = 0.749496$$

方程 $f(c)=0.749496$ 在区间 $[0,2.5]$ 内有 3 个解:$c_1=0.440566$,$c_2=1.268010$ 和 $c_3=1.873583$。长为 $b-a=2.5$,高为 $f(c_j)=0.749496$ 的矩形区域面积为 $f(c_j)(b-a)=1.873740$。此矩形区域面积值与函数 $f(x)$ 在区间 $[0,2.5]$ 内的积分值相同。两者区域的比较如图 1.5 所示。 ■

定理 1.11(加权积分均值定理) 若 $f,g\in C[a,b]$,且当 $x\in[a,b]$ 时有 $g(x)\geqslant 0$,则至少存在一点 $c\in(a,b)$,满足

$$\int_a^b f(x)g(x)\,\mathrm{d}x = f(c)\int_a^b g(x)\,\mathrm{d}x \quad (14)$$

例 1.6　函数 $f(x)=\sin(x)$ 和函数 $g(x)=x^2$ 在区间 $[0,\pi/2]$ 内满足定理 1.11 的假设，则至少存在一点 c，满足

$$\sin(c)=\frac{\int_0^{\pi/2}x^2\sin(x)\,\mathrm{d}x}{\int_0^{\pi/2}x^2\,\mathrm{d}x}=\frac{1.14159}{1.29193}=0.883631$$

或 $c=\arcsin(0.883631)=1.08356$。 ■

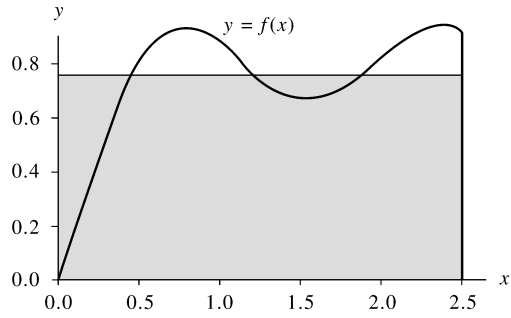


图 1.5　将积分均值定理运用在 $f(x)=\sin(x)+\frac{1}{3}\sin(3x)$ 上，区间为 $[0,2.5]$

1.1.4　级数

定义 1.5　设有序列 $\{a_n\}_{n=1}^{\infty}$，则 $\sum_{n=1}^{\infty}a_n$ 为一个无穷级数，且第 n 个部分和为 $S_n=\sum_{k=1}^{n}a_k$。无穷级数**收敛**当且仅当序列 $\{S_n\}_{n=1}^{\infty}$ 收敛于极限 S，可表示为

$$\lim_{n\to\infty}S_n=\lim_{n\to\infty}\sum_{k=1}^{n}a_k=S \quad (15)$$

如果级数不收敛，则称之为级数**发散**。 ▲

例 1.7　已知无穷序列 $\{a_n\}_{n=1}^{\infty}=\left\{\frac{1}{n(n+1)}\right\}_{n=1}^{\infty}$，则第 n 个部分和为

$$S_n=\sum_{k=1}^{n}\frac{1}{k(k+1)}=\sum_{k=1}^{n}\left(\frac{1}{k}-\frac{1}{k+1}\right)=1-\frac{1}{n+1}$$

因此，无穷级数的**和**为

$$S=\lim_{n\to\infty}S_n=\lim_{n\to\infty}\left(1-\frac{1}{n+1}\right)=1$$ ■

定理 1.12（泰勒定理）　若函数 $f\in C^{n+1}[a,b]$，且 $x_0\in[a,b]$，则对任意 $x\in(a,b)$，都存在 $c=c(x)$（c 依赖于 x）位于 x_0 和 x 之间，满足

$$f(x)=P_n(x)+R_n(x) \quad (16)$$

其中

$$P_n(x)=\sum_{k=0}^{n}\frac{f^{(k)}(x_0)}{k!}(x-x_0)^k \quad (17)$$

且

$$R_n(x)=\frac{f^{(n+1)}(c)}{(n+1)!}(x-x_0)^{n+1} \quad (18)$$

例 1.8　函数 $f(x)=\sin(x)$ 满足定理 1.12 的假设，则通过将函数 $f(x)$ 在 $x=0$ 处的各阶导数值代入式(17)，可得到在 $x_0=0$ 处 $n=9$ 的 n 阶泰勒多项式 $P_n(x)$：

$$\begin{aligned} f(x)&=\sin(x), & f(0)&=0,\\ f'(x)&=\cos(x), & f'(0)&=1,\\ f''(x)&=-\sin(x), & f''(0)&=0, \end{aligned}$$

$$f^{(3)}(x) = -\cos(x), \qquad f^{(3)}(0) = -1,$$
$$\vdots \qquad\qquad \vdots$$
$$f^{(9)}(x) = \cos(x), \qquad f^{(9)}(0) = 1,$$

$$P_9(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!}$$

函数 f 和 P_9 在区间 $[0,2\pi]$ 内的图形表示如图 1.6 所示。 ■

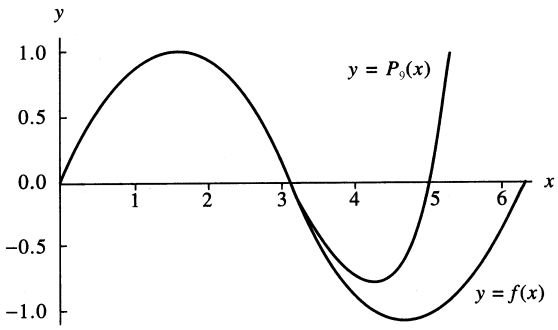


图 1.6 函数 $f(x) = \sin(x)$ 和泰勒多项式 $P_9(x) = x - x^3/3! + x^5/5! - x^7/7! + x^9/9!$ 的图形表示

推论 1.1 如果 $P_n(x)$ 是定理 1.12 中的 n 阶泰勒多项式,则

$$P_n^{(k)}(x_0) = f^{(k)}(x_0), \quad k = 0, 1, \cdots, n \tag{19}$$

1.1.5 多项式求值

设 n 阶多项式 $P(x)$ 有如下形式:

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0 \tag{20}$$

霍纳(Horner)方法[①]或**综合除法**是计算多项式的一种方法,可将其看成一种嵌套乘法。例如,一个 5 阶多项式可改写为如下嵌套乘法的形式:

$$P_5(x) = ((((a_5x + a_4)x + a_3)x + a_2)x + a_1)x + a_0$$

定理 1.13(用于多项式计算的霍纳方法) 设 $P(x)$ 是按式(20)给出的多项式,且 $x = c$ 是用于计算 $P(c)$ 的数。

设 $b_n = a_n$,并计算

$$b_k = a_k + cb_{k+1}, \quad k = n-1, n-2, \cdots, 1, 0 \tag{21}$$

则 $b_0 = P(c)$。进一步考虑,如果

$$Q_0(x) = b_n x^{n-1} + b_{n-1}x^{n-2} + \cdots + b_3 x^2 + b_2 x + b_1 \tag{22}$$

则

$$P(x) = (x - c)Q_0(x) + R_0 \tag{23}$$

其中 $Q_0(x)$ 是 $n-1$ 阶多项式的商,$R_0 = b_0 = P(c)$ 是余数。

① 秦九韶于 1247 年提出此方法,而霍纳于 1819 年提出同样的方法。——译者注

证明:在式(23)中,用式(22)的右边替换 $Q_0(x)$,用 b_0 替换 R_0,则可得到

$$\begin{aligned} P(x) &= (x-c)(b_n x^{n-1} + b_{n-1}x^{n-2} + \cdots + b_3 x^2 + b_2 x + b_1) + b_0 \\ &= b_n x^n + (b_{n-1} - cb_n)x^{n-1} + \cdots + (b_2 - cb_3)x^2 \\ &\quad + (b_1 - cb_2)x + (b_0 - cb_1) \end{aligned} \tag{24}$$

通过比较式(20)和式(24)中 x^k 的系数,可以确定 b_k 的值,如表 1.1 所示。

表 1.1　用于霍纳方法的系数 b_k

x^k	对比式 (20) 和式 (24)	求解 b_k
x^n	$a_n = b_n$	$b_n = a_n$
x^{n-1}	$a_{n-1} = b_{n-1} - cb_n$	$b_{n-1} = a_{n-1} + cb_n$
⋮	⋮	⋮
x^k	$a_k = b_k - cb_{k+1}$	$b_k = a_k + cb_{k+1}$
⋮	⋮	⋮
x^0	$a_0 = b_0 - cb_1$	$b_0 = a_0 + cb_1$

将 $x = c$ 替换入式(22),由于 $R_0 = b_0$,很容易得出结论 $P(c) = b_0$,如下所示:

$$P(c) = (c-c)Q_0(c) + R_0 = b_0 \tag{25}$$

●

借助于计算机,可以很容易地实现式(21)中计算 b_k 的递归公式。一个简单的算法如下:

```
b(n) = a(n);
for k = n − 1: −1: 0
    b(k) = a(k) + c * b(k + 1);
end
```

当手工计算霍纳方法时,将 $P(x)$ 的系数写在一行,在 a_k 所处的列下方计算 $b_k = a_k + cb_{k+1}$ 则更方便。这一处理过程如表 1.2 所示。

表 1.2　用于综合除法过程的霍纳表

输入	a_n	a_{n-1}	a_{n-2}	$\cdots$	a_k	$\cdots$	a_2	a_1	a_0
c		xb_n	xb_{n-1}	$\cdots$	xb_{k+1}	$\cdots$	xb_3	xb_2	xb_1
	b_n	b_{n-1}	b_{n-2}	$\cdots$	b_k	$\cdots$	b_2	b_1	$b_0 = P(c)$
									输出

例 1.9　利用综合除法(霍纳方法)求 $P(3)$,其中多项式

$$P(x) = x^5 - 6x^4 + 8x^3 + 8x^2 + 4x - 40$$

	a_5	a_4	a_3	a_2	a_1	a_0
输入	1	−6	8	8	4	−40
$c = 3$		3	−9	−3	15	57
	1	−3	−1	5	19	$17 = P(3) = b_0$
	b_5	b_4	b_3	b_2	b_1	输出

因此,$P(3) = 17$。 ■

1.1.6　习题

1. (a) 求解 $L = \lim_{n\to\infty}(4n+1)/(2n+1)$,确定 $\{\epsilon_n\} = \{L - x_n\}$ 并求解 $\lim_{n\to\infty}\epsilon_n$。

(b) 求解 $L=\lim_{n\to\infty}(2n^2+6n-1)/(4n^2+2n+1)$,确定 $\{\epsilon_n\}=\{L-x_n\}$ 并求解 $\lim_{n\to\infty}\epsilon_n$。

2. 设 $\{x_n\}_{n=1}^{\infty}$ 是满足 $\lim_{n\to\infty}x_n=2$ 的序列。
   (a) 求解式 $\lim_{n\to\infty}\sin(x_n)$。
   (b) 求解 $\lim_{n\to\infty}\ln(x_n^2)$。
3. 根据中值定理,求解下列函数在指定区间内满足值 L 的数 c。
   (a) $f(x)=-x^2+2x+3$,区间为 $[-1,0]$,$L=2$。
   (b) $f(x)=\sqrt{x^2-5x-2}$,区间为 $[6,8]$,$L=3$。
4. 根据极值定理,求解下列函数在指定区间内的上界和下界。
   (a) $f(x)=x^2-3x+1$,区间为 $[-1,2]$。
   (b) $f(x)=\cos^2(x)-\sin(x)$,区间为 $[0,2\pi]$。
5. 根据罗尔定理,求解下列函数在指定区间内的 c 值。
   (a) $f(x)=x^4-4x^2$,区间为 $[-2,2]$。
   (b) $f(x)=\sin(x)+\sin(2x)$,区间为 $[0,2\pi]$。
6. 根据均值定理,求解下列函数在指定区间内的 c 值。
   (a) $f(x)=\sqrt{x}$,区间为 $[0,4]$。
   (b) $f(x)=\dfrac{x^2}{x+1}$,区间为 $[0,1]$。
7. 给定区间 $[0,3]$,将广义罗尔定理应用于函数 $f(x)=x(x-1)(x-3)$。
8. 将第一基本定理应用于下列指定区间内的函数。
   (a) $f(x)=xe^x$,区间为 $[0,2]$。
   (b) $f(x)=\dfrac{3x}{x^2+1}$,区间为 $[-1,1]$。
9. 将第二基本定理应用于下列函数。
   (a) $\frac{\mathrm{d}}{\mathrm{d}x}\int_0^x t^2\cos(t)\,\mathrm{d}t$　　(b) $\frac{\mathrm{d}}{\mathrm{d}x}\int_1^{x^3}\mathrm{e}^{t^2}\,\mathrm{d}t$
10. 根据积分均值定理,求解下列函数在指定区间内的 c 值。
    (a) $f(x)=6x^2$,区间为 $[-3,4]$。
    (b) $f(x)=x\cos(x)$,区间为 $[0,3\pi/2]$。
11. 求解下列序列或级数的和。
    (a) $\left\{\dfrac{1}{2^n}\right\}_{n=0}^{\infty}$　　(b) $\left\{\dfrac{2}{3^n}\right\}_{n=1}^{\infty}$　　(c) $\displaystyle\sum_{n=1}^{\infty}\frac{3}{n(n+1)}$　　(d) $\displaystyle\sum_{k=1}^{\infty}\frac{1}{4k^2-1}$
12. 求解下列函数在 x_0 处的 4 阶泰勒多项式。
    (a) $f(x)=\sqrt{x}$,$x_0=1$
    (b) $f(x)=x^5+4x^2+3x+1$,$x_0=0$
    (c) $f(x)=\cos(x)$,$x_0=0$
13. 设 $f(x)=\sin(x)$,且 $P(x)=x-x^3/3!+x^5/5!-x^7/7!+x^9/9!$。证明 $P^{(k)}(0)=f^{(k)}(0)$,$k=1,2,\cdots,9$。
14. 利用综合除法(霍纳方法)求解 $P(c)$。
    (a) $P(x)=x^4+x^3-13x^2-x-12,\ c=3$
    (b) $P(x)=2x^7+x^6+x^5-2x^4-x+23,\ c=-1$

15. 求解中心位于原点,半径在 1 到 3 之间的所有圆的平均面积。
16. 设多项式 $P(x)$ 在区间 $[a,b]$ 内有 n 个实根,证明 $P^{(n-1)}(x)$ 在区间 $[a,b]$ 内至少有 1 个实根。
17. 设 f,f' 和 f'' 在区间 $[a,b]$ 内有定义,且 $f(a)=f(b)=0$,当 $c\in(a,b)$ 时有 $f(c)>0$。证明存在数 $d\in(a,b)$,满足 $f''(d)<0$。

1.2　二进制数

在日常生活中,人们通常使用十进制数进行计算,但大多数计算机内部通常采用二进制数。虽然从表面上看,计算机的通信(输入、输出操作)是基于十进制数的,但这并不表示计算机内部采用的是十进制数。事实上,计算机首先将输入的十进制数转换成二进制数,然后进行二进制运算,最后将答案再转换成十进制数显示出来。可以通过下述实验对此进行验证。例如,某计算机的最高精度为 9 位十进制数,则计算下述公式的结果为

$$\sum_{k=1}^{100000}\frac{1}{10}=9999.99447 \tag{1}$$

式(1)将数 $\frac{1}{10}$ 进行 100000 次累加,而它的算术精确值应当是 10000。通过上述分析,说明计算机的计算过程存在明显误差。本节最后将具体描述当计算机将十进制小数 $\frac{1}{10}$ 转换成二进制数时,误差是如何产生的。

1.2.1　二进制数

大多数数学计算都采用十进制数。例如,十进制数 1563 的**展开式**为

$$1563=(1\times10^3)+(5\times10^2)+(6\times10^1)+(3\times10^0)$$

通常,若 N 为一个正整数,则存在数 $a_0,a_1,\cdots,a_k$,使得 N 的十进制扩展表达式为

$$N=(a_k\times10^k)+(a_{k-1}\times10^{k-1})+\cdots+(a_1\times10^1)+(a_0\times10^0)$$

其中数 a_k 的取值范围为 $\{0,1,\cdots,8,9\}$,则 N 的十进制表示为

$$N=a_ka_{k-1}\cdots a_2a_1a_{0_{10}}\qquad(\text{十进制}) \tag{2}$$

如果已知该数为十进制数,则可以表示为

$$N=a_ka_{k-1}\cdots a_2a_1a_0$$

例如,$1563=1563_{10}$。

若用 2 的幂表示,则十进制数 1563 可以表示为:

$$\begin{aligned}1563=&(1\times2^{10})+(1\times2^9)+(0\times2^8)+(0\times2^7)+(0\times2^6)\\&+(0\times2^5)+(1\times2^4)+(1\times2^3)+(0\times2^2)+(1\times2^1)\\&+(1\times2^0)\end{aligned} \tag{3}$$

可以通过如下计算进行验证:

$$1563=1024+512+16+8+2+1$$

通常,若 N 为一个正整数,则存在数 $b_0,b_1,\cdots,b_J$,使得 N 的二进制扩展表达式为

$$N=(b_J\times2^J)+(b_{J-1}\times2^{J-1})+\cdots+(b_1\times2^1)+(b_0\times2^0) \tag{4}$$

其中数 b_j 为 0 或 1,则 N 的二进制表示为

$$N=b_Jb_{J-1}\cdots b_2b_1b_{0_2}\qquad(\text{二进制}) \tag{5}$$

利用上式可将式(3)表示为

$$1563 = 11000011011_2$$

批注 2 作为下标总出现在二进制数的末尾,以便读者区分十进制数和二进制数。因此,111 表示数一百一十一,而 111_2 表示数七。

通常,由于 2 的幂的增长小于 10 的幂的增长,因此数 N 的二进制表示位数远大于它的十进制表示位数。

通过式(4)可以得到一个计算 N 的二进制表示的有效算法。将式(4)的两边同除以 2 可得

$$\frac{N}{2} = (b_J \times 2^{J-1}) + (b_{J-1} \times 2^{J-2}) + \cdots + (b_1 \times 2^0) + \frac{b_0}{2} \tag{6}$$

这样,N 除以 2 的余数是 b_0。下面来计算 b_1。如果式(6)表示为 $N/2 = Q_0 + b_0/2$,则有

$$Q_0 = (b_J \times 2^{J-1}) + (b_{J-1} \times 2^{J-2}) + \cdots + (b_2 \times 2^1) + (b_1 \times 2^0) \tag{7}$$

将式(7)的两边同除以 2 可得

$$\frac{Q_0}{2} = (b_J \times 2^{J-2}) + (b_{J-1} \times 2^{J-3}) + \cdots + (b_2 \times 2^0) + \frac{b_1}{2}$$

这样,Q_0 除以 2 的余数是 b_1。将这个计算过程持续下去,可分别得到商序列 $\{Q_k\}$ 和余数序列 $\{b_k\}$。当找到整数 J,满足 $Q_J = 0$ 时,计算过程停止。这些序列符合下式:

$$\begin{aligned} N &= 2Q_0 + b_0 \\ Q_0 &= 2Q_1 + b_1 \\ &\vdots \\ Q_{J-2} &= 2Q_{J-1} + b_{J-1} \\ Q_{J-1} &= 2Q_J + b_J \qquad (Q_J = 0) \end{aligned} \tag{8}$$

例 1.10 完成 $1563 = 11000011011_2$ 的计算过程。

根据式(8),从 $N = 1536$ 开始,构造商序列和余数序列:

$$\begin{aligned} 1563 &= 2 \times 781 + 1, & b_0 &= 1 \\ 781 &= 2 \times 390 + 1, & b_1 &= 1 \\ 390 &= 2 \times 195 + 0, & b_2 &= 0 \\ 195 &= 2 \times 97 + 1, & b_3 &= 1 \\ 97 &= 2 \times 48 + 1, & b_4 &= 1 \\ 48 &= 2 \times 24 + 0, & b_5 &= 0 \\ 24 &= 2 \times 12 + 0, & b_6 &= 0 \\ 12 &= 2 \times 6 + 0, & b_7 &= 0 \\ 6 &= 2 \times 3 + 0, & b_8 &= 0 \\ 3 &= 2 \times 1 + 1, & b_9 &= 1 \\ 1 &= 2 \times 0 + 1, & b_{10} &= 1 \end{aligned}$$

则 1563 的二进制表示为

$$1563 = b_{10}b_9b_8 \cdots b_2b_1b_{0_2} = 11000011011_2$$ ■

1.2.2 序列与级数

当有理数表示成小数形式时,可能需要无穷多的位数。下面是一个常见的例子:

$$\frac{1}{3} = 0.\overline{3} \tag{9}$$

其中的符号 $\overline{3}$ 表示数 3 重复无穷次,形成无穷循环小数。式(9)的基为 10,而且从数学意义上讲,式(9)是式(10)的一种简化表示。

$$\begin{aligned} S &= (3\times 10^{-1}) + (3\times 10^{-2}) + \cdots + (3\times 10^{-n}) + \cdots \\ &= \sum_{k=1}^{\infty} 3(10)^{-k} = \frac{1}{3} \end{aligned} \tag{10}$$

如果只显示有限位数,则可得到 $\frac{1}{3}$ 的近似值。例如,$\frac{1}{3}\approx 0.333 = 333/1000$。近似值的误差为 $\frac{1}{3000}$。根据式(10),可验证 $\frac{1}{3} = 0.333 + \frac{1}{3000}$。

理解式(10)的扩展表示是非常重要的。一种基本方法是在两边都乘以 10,再相减,如下所示:

$$\begin{array}{rl} 10S = & 3 + (3\times 10^{-1}) + (3\times 10^{-2}) + \cdots + (3\times 10^{-n}) + \cdots \\ -S = & -(3\times 10^{-1}) - (3\times 10^{-2}) - \cdots - (3\times 10^{-n}) - \cdots \\ \hline 9S = & 3 + (0\times 10^{-1}) + (0\times 10^{-2}) + \cdots + (0\times 10^{-n}) + \cdots \end{array}$$

这样可以得到 $S = \frac{3}{9} = \frac{1}{3}$。在许多有关微积分的书中都可以找到证明两个无穷级数的差别的相关定理。下面将介绍一些基本概念,更详细的内容可参见与微积分相关的教材。

定义 1.6　*无穷级数*

$$\sum_{n=0}^{\infty} cr^n = c + cr + cr^2 + \cdots + cr^n + \cdots \tag{11}$$

*称为比率为 r 的*__几何级数__*,其中 $c\neq 0$,且 $r\neq 0$。* ▲

定理 1.14(几何级数)　*几何级数有如下性质:*

$$\text{如果 } |r|<1, \text{则 } \sum_{n=0}^{\infty} cr^n = \frac{c}{1-r} \tag{12}$$

$$\text{如果 } |r|>1, \text{则级数发散} \tag{13}$$

证明:有限几何级数的和可以表示为

$$S_n = c + cr + cr^2 + \cdots + cr^n = \frac{c(1-r^{n+1})}{1-r}, \quad r\neq 1 \tag{14}$$

为了建立式(12),可以看到

$$|r|<1 \qquad \text{意味着} \qquad \lim_{n\to\infty} r^{n+1} = 0 \tag{15}$$

这样,当 $n\to\infty$ 时,根据式(14)和式(15)可得

$$\lim_{n\to\infty} S_n = \frac{c}{1-r}\left(1 - \lim_{n\to\infty} r^{n+1}\right) = \frac{c}{1-r}$$

根据 1.1 节的式(15),可以证明式(12)成立。

如果 $|r|\geqslant 1$,则序列 $\{r^{n+1}\}$ 不收敛,因此式(14)中的序列 $\{S_n\}$ 不存在极限。这样,可以证明式(13)成立。 ●

根据定理 1.14 中的式(12),可以得到一个将无穷循环小数转换成分数的有效方法。

例 1.11

$$\begin{aligned} 0.\overline{3} = \sum_{k=1}^{\infty} 3(10)^{-k} &= -3 + \sum_{k=0}^{\infty} 3(10)^{-k} \\ &= -3 + \frac{3}{1-\frac{1}{10}} = -3 + \frac{10}{3} = \frac{1}{3} \end{aligned}$$

■

1.2.3 二进制分数

二进制(基数为 2)分数可以表示为 2 的负幂之和。如果 R 是一个实数,且 $0 < R < 1$,则存在 $d_1, d_2, \cdots, d_n, \cdots$,满足

$$R = (d_1 \times 2^{-1}) + (d_2 \times 2^{-2}) + \cdots + (d_n \times 2^{-n}) + \cdots \tag{16}$$

其中 $d_j \in \{0,1\}$。式(16)右边的值通常可以表示成二进制分数形式,如下所示:

$$R = 0.d_1 d_2 \cdots d_n \cdots_2 \tag{17}$$

许多实数的二进制表示有无穷位。分数 $\frac{7}{10}$ 的十进制表示为 0.7,但它的二进制表示有无穷位:

$$\frac{7}{10} = 0.1\overline{0110}_2 \tag{18}$$

式(18)中的二进制小数表示有无穷位,由 0110 这 4 个数重复循环构成。

现在可以开发一个求二进制表示的有效算法。如果式(16)的两边都乘以 2,则结果为

$$2R = d_1 + ((d_2 \times 2^{-1}) + \cdots + (d_n \times 2^{-n+1}) + \cdots) \tag{19}$$

式(19)右边的括号中的值是一个小于 1 的正数。因此 d_1 是 $2R$ 的整数部分,即 $d_1 = \text{int}(2R)$。继续上述步骤,可得到式(19)的小数部分,表示为

$$F_1 = \text{frac}(2R) = (d_2 \times 2^{-1}) + \cdots + (d_n \times 2^{-n+1}) + \cdots \tag{20}$$

其中 frac(2R)是实数 $2R$ 的小数部分。在式(20)两边乘以 2,可以得到

$$2F_1 = d_2 + ((d_3 \times 2^{-1}) + \cdots + (d_n \times 2^{-n+2}) + \cdots) \tag{21}$$

根据式(21)的整数部分可以得到 $d_2 = \text{int}(2F_1)$。

如果继续上述过程,无限执行下去(如果 R 是一个无限不重复的二进制表示),则可以递归地得到序列 $\{d_k\}$ 和 $\{F_k\}$:

$$\begin{aligned} d_k &= \text{int}(2F_{k-1}) \\ F_k &= \text{frac}(2F_{k-1}) \end{aligned} \tag{22}$$

其中 $d_1 = \text{int}(2R)$,且 $F_1 = \text{frac}(2R)$。通过下列收敛的几何级数:

$$R = \sum_{j=1}^{\infty} d_j (2)^{-j}$$

可描述出 R 的二进制小数表示。

例 1.12 通过式(22)可计算出式(18)中 $\frac{7}{10}$ 的二进制小数表示。设 $R = \frac{7}{10} = 0.7$,则

$2R = 1.4$	$d_1 = \text{int}(1.4) = 1$	$F_1 = \text{frac}(1.4) = 0.4$
$2F_1 = 0.8$	$d_2 = \text{int}(0.8) = 0$	$F_2 = \text{frac}(0.8) = 0.8$
$2F_2 = 1.6$	$d_3 = \text{int}(1.6) = 1$	$F_3 = \text{frac}(1.6) = 0.6$
$2F_3 = 1.2$	$d_4 = \text{int}(1.2) = 1$	$F_4 = \text{frac}(1.2) = 0.2$
$2F_4 = 0.4$	$d_5 = \text{int}(0.4) = 0$	$F_5 = \text{frac}(0.4) = 0.4$
$2F_5 = 0.8$	$d_6 = \text{int}(0.8) = 0$	$F_6 = \text{frac}(0.8) = 0.8$
$2F_6 = 1.6$	$d_7 = \text{int}(1.6) = 1$	$F_7 = \text{frac}(1.6) = 0.6$

注意，$2F_2 = 1.6 = 2F_6$。当 $k = 2, 3, 4, \cdots$ 时，可得到模式 $d_k = d_{k+4}$ 和 $F_k = F_{k+4}$，因此有 $\frac{7}{10} = 0.1\,\overline{0110}_2$。 ■

通过几何级数可以找到二进制有理数的十进制形式。

例 1.13　求解二进制有理数 $0.\overline{01}_2$ 的十进制形式。根据展开式有

$$\begin{aligned} 0.\overline{01}_2 &= (0 \times 2^{-1}) + (1 \times 2^{-2}) + (0 \times 2^{-3}) + (1 \times 2^{-4}) + \cdots \\ &= \sum_{k=1}^{\infty} (2^{-2})^k = -1 + \sum_{k=0}^{\infty} (2^{-2})^k \\ &= -1 + \frac{1}{1 - \frac{1}{4}} = -1 + \frac{4}{3} = \frac{1}{3} \end{aligned}$$ ■

1.2.4　二进制移位

可以通过移位操作，简化求解无穷循环的二进制有理数的过程。例如，S 为

$$S = 0.00000\overline{11000}_2 \tag{23}$$

在式(23)的两边乘以 2^5，将小数点右移 5 位，得到 $32S$，即

$$32S = 0.\overline{11000}_2 \tag{24}$$

同样，在式(23)两边乘以 2^{10}，将小数点右移 10 位，得到 $1024S$，即

$$1024S = 11000.\overline{11000}_2 \tag{25}$$

式(25)的左右两边减去式(24)的左右两边，可以得到 $992S = 11000_2$，或根据 $11000_2 = 24$ 得到 $992S = 24$，因此 $S = 8/124$。

1.2.5　科学计数法

将十进制小数点移位并乘以 10 的幂，是表示实数的标准方法之一，通常称之为**科学计数法**。例如，

$$0.0000747 = 7.47 \times 10^{-5}$$
$$31.4159265 = 3.14159265 \times 10$$
$$9700000000 = 9.7 \times 10^{9}$$

在化学领域中，阿伏伽德罗(Avogadro)数是一个重要的常量，可表示为 6.02252×10^{23}。它是单个分子的克原子重量下的原子个数。在计算机科学中，$1\mathrm{K} = 1.024 \times 10^3$。

1.2.6　机器数

计算机使用规格化浮点二进制数表示实数。这意味着实数 x 的算术值并没有实际存放在计算机中，计算机中实际存放的是 x 的近似值，

$$x \approx \pm q \times 2^n \tag{26}$$

其中数 q 称为**尾数**，它是一个有限的二进制数，满足不等式 $1/2 \leqslant q < 1$。整数 n 称为**阶码**。

在计算机中，只使用了实数系统的一小部分子集。典型的子集一般只包含式(26)表示的一部分。二进制数的个数受 q 和 n 的严格限制。例如，考虑下述正实数集合中的所有实数：

$$0.d_1d_2d_3d_{4_2} \times 2^n \tag{27}$$

其中,$d_1 = 1$,而 d_2, d_3 和 d_4 为 0 或 1,而且 $n \in \{-3, -2, -1, 0, 1, 2, 3, 4\}$。在式(27)中,尾数和阶码各有 8 种选择,可以产生 64 个数的集合:

$$\{0.1000_2 \times 2^{-3}, 0.1001_2 \times 2^{-3}, \cdots, 0.1110_2 \times 2^4, 0.1111_2 \times 2^4\} \tag{28}$$

这 64 个数的十进制形式如表 1.3 所示。当式(27)中的尾数和阶码受限时,计算机只能选用有限的数存储实数 x 的近似值。

表 1.3 尾数为 4 比特(bit),阶码为 $n = -3, -2, \cdots, 3, 4$ 的二进制数集合的十进制表示

尾数	阶码							
	$n=-3$	$n=-2$	$n=-1$	$n=0$	$n=1$	$n=2$	$n=3$	$n=4$
0.1000_2	0.0625	0.125	0.25	0.5	1	2	4	8
0.1001_2	0.0703125	0.140625	0.28125	0.5625	1.125	2.25	4.5	9
0.1010_2	0.078125	0.15625	0.3125	0.625	1.25	2.5	5	10
0.1011_2	0.0859375	0.171875	0.34375	0.6875	1.375	2.75	5.5	11
0.1100_2	0.09375	0.1875	0.375	0.75	1.5	3	6	12
0.1101_2	0.1015625	0.203125	0.40625	0.8125	1.625	3.25	6.5	13
0.1110_2	0.109375	0.21875	0.4375	0.875	1.75	3.5	7	14
0.1111_2	0.1171875	0.234375	0.46875	0.9375	1.875	3.75	7.5	15

如果计算机的尾数为 4 比特,如何计算$(\frac{1}{10} + \frac{1}{5}) + \frac{1}{6}$?假定计算机将实数四舍五入为表 1.3 中的二进制数,读者可根据表 1.3 查找各个实数所对应的最佳近似值。

$$\begin{array}{rcll} \frac{1}{10} & \approx & 0.1101_2 \times 2^{-3} & = \quad 0.01101_2 \times 2^{-2} \\ \frac{1}{5} & \approx & 0.1101_2 \times 2^{-2} & = \quad 0.1101_2 \times 2^{-2} \\ \hline \frac{3}{10} & & & \quad 1.00111_2 \times 2^{-2} \end{array} \tag{29}$$

计算机必须确定如何存储数 $1.00111_2 \times 2^{-2}$。若数被四舍五入为 $0.1010_2 \times 2^{-1}$,则下一步是

$$\begin{array}{rcll} \frac{3}{10} & \approx & 0.1010_2 \times 2^{-1} & = \quad 0.1010_2 \times 2^{-1} \\ \frac{1}{6} & \approx & 0.1011_2 \times 2^{-2} & = \quad 0.01011_2 \times 2^{-1} \\ \hline \frac{7}{15} & & & \quad 0.11111_2 \times 2^{-1} \end{array} \tag{30}$$

计算机必须确定如何存储数 $0.11111_2 \times 2^{-1}$。由于假定数被四舍五入,则数被存储为 $0.10000_2 \times 2^0$。因此,加法问题的计算机解为

$$\frac{7}{15} \approx 0.10000_2 \times 2^0 \tag{31}$$

计算机的计算误差为

$$\frac{7}{15} - 0.10000_2 \approx 0.466667 - 0.500000 \approx 0.033333 \tag{32}$$

如果用$\frac{7}{15}$的百分比来表示误差,则误差为 7.14%。

1.2.7 计算机精度

为了精确地存储数值,计算机必须使用二进制浮点数。其中,要表示 7 位十进制数,至少需要 24 比特的尾数;如果使用 32 比特的尾数,则可以存储 9 位十进制数。现在重新考虑在本节开始的式(1)中遇到的困难,即计算机累加$\frac{1}{10}$出现的误差。

设式(26)中的尾数 q 的位数为 32。根据条件 $\frac{1}{2} \leqslant q$,可得到第一个数 $d_1 = 1$。因此,q 有如下形式:

$$q = 0.1d_2d_3 \cdots d_{31}d_{32}{}_2 \tag{33}$$

当把小数部分用二进制形式表示时,通常是无限循环小数。例如,表达式

$$\frac{1}{10} = 0.0\overline{0011}_2 \tag{34}$$

当使用 32 位尾数时,计算机对数进行截断后的内在近似值为

$$\frac{1}{10} \approx 0.11001100110011001100110011001100_2 \times 2^{-3} \tag{35}$$

式(35)中近似值的误差,即式(35)与式(34)的差值为

$$0.\overline{1100}_2 \times 2^{-35} \approx 2.328306437 \times 10^{-11} \tag{36}$$

由于式(36)的存在,计算机对式(1)中的 $\frac{1}{10}$ 进行 100000 次累加后,必然存在误差。误差大于 $100000 \times 2.328306437 \times 10^{-11} = 2.328306437 \times 10^{-6}$。实际上,在计算过程中还存在着更大的误差。原因是部分和可能随时会被四舍五入,而且随着和的增加,$\frac{1}{10}$ 的后一个加数远小于逐渐增加的累加和,它在计算中的影响会被舍弃。这样,各种误差带来的组合效应使得实际产生的误差为 $10000 - 9999.99447 = 5.53 \times 10^{-3}$。

1.2.8　计算机浮点数

计算机采用整数模式和浮点模式来表示数。整数模式主要用来进行整数计算,在数值分析中的作用有限,而科学工程应用中主要采用浮点数。必须要明确如下事实:在式(26)的任何计算机实现中,尾数 q 的位数是有限的,而且阶码 n 的范围也肯定是有限的。

在以 32 位表示单精度实数的计算机中,阶码用 8 位表示,尾数用 24 位表示,因此可以表示的实数范围是

$$2.938736\text{E}-39 \quad \text{到} \quad 1.701412\text{E}+38$$

即 2^{-128} 到 2^{127},有 6 位十进制数值精度,例如,$2^{-23} = 1.2 \times 10^{-7}$。

在以 48 位表示单精度实数的计算机中,阶码用 8 位表示,尾数用 40 位表示,因此可以表示的实数范围是

$$2.9387358771\text{E}-39 \quad \text{到} \quad 1.7014118346\text{E}+38$$

即 2^{-128} 到 2^{127},有 11 位十进制数值精度,例如,$2^{-39} = 1.8 \times 10^{-12}$。

对于以 64 位表示双精度实数的计算机,可能会用 11 位表示阶码,用 53 位表示尾数,因此可以表示的实数范围是

$$5.562684646268003\text{E}-309 \quad \text{到} \quad 8.988465674311580\text{E}+307$$

即 2^{-1024} 到 2^{1023},有 16 位十进制数值精度,例如,$2^{-52} = 2.2 \times 10^{-16}$。

1.2.9　习题

1. 利用计算机求解下列表达式。提示:目的是让计算机重复执行减法计算,不要借助于乘法计算。

(a) $10000 - \sum_{k=1}^{100000} 0.1$　　(b) $10000 - \sum_{k=1}^{80000} 0.125$

2. 利用式(4)和式(5),将下列二进制数转换成十进制形式。

(a) 10101_2　　(b) 111000_2

(c) 11111110_2　　(d) 1000000111_2

3. 利用式(16)和式(17),将下列二进制小数转换成十进制形式。

(a) 0.11011_2　　(b) 0.10101_2

(c) 0.1010101_2　　(d) 0.110110110_2

4. 将下列二进制数转换成十进制形式。

(a) 1.0110101_2　　(b) 11.0010010001_2

5. 习题 4 中的数是$\sqrt{2}$和π的近似值。求解近似值的误差,即计算下列值:

(a) $\sqrt{2}-1.0110101_2$　　(利用$\sqrt{2}=1.41421356237309\cdots$)

(b) $\pi-11.0010010001_2$　　(利用$\pi=3.14159265358979\cdots$)

6. 参照例 1.10,将下列十进制数转换成二进制形式。

(a) 23　　(b) 87　　(c) 378　　(d) 2388

7. 参照例 1.12,将下列十进制数转换成如 $0.d_1d_2\cdots d_{n2}$ 形式的二进制小数。

(a) 7/16　　(b) 13/16　　(c) 23/32　　(d) 75/128

8. 参照例 1.12,将下列十进制数转换成二进制无限循环小数。

(a) 1/10　　(b) 1/3　　(c) 1/7

9. 求解下列具有 7 位有效数字的二进制近似值的误差 $R-0.d_1d_2d_3d_4d_5d_6d_{72}$。

(a) $1/10\approx 0.0001100_2$　　(b) $1/7\approx 0.0010010_2$

10. 证明二进制展开式 $1/7=0.\overline{001}_2$ 与$\frac{1}{7}=\frac{1}{8}+\frac{1}{64}+\frac{1}{512}+\cdots$等价。利用定理 1.14 来建立这个展开式。

11. 证明二进制展开式 $1/5=0.\overline{0011}_2$ 与$\frac{1}{5}=\frac{3}{16}+\frac{3}{256}+\frac{3}{4096}+\cdots$等价。利用定理 1.14 来建立这个展开式。

12. 证明对于任意的数 2^{-N},其中 N 是正整数,可以表示为 N 位十进制数,即 $2^{-N}=0.d_1d_2d_3\cdots d_N$。提示:$1/2=0.5$,$1/4=0.25$,…。

13. 根据表 1.3 来判断当用有 4 位尾数的计算机执行下列计算时,会得到什么结果。

(a) $\left(\frac{1}{3}+\frac{1}{5}\right)+\frac{1}{6}$　　(b) $\left(\frac{1}{10}+\frac{1}{3}\right)+\frac{1}{5}$

(c) $\left(\frac{3}{17}+\frac{1}{9}\right)+\frac{1}{7}$　　(d) $\left(\frac{7}{10}+\frac{1}{9}\right)+\frac{1}{7}$

14. 证明当将本节的式(8)中所有的 2 替换为 3 时,可得到一个求解正整数的三进制表示形式的方法,并求解下列整数的三进制表示形式。

(a) 10　　(b) 23　　(c) 421　　(d) 1784

15. 证明当将本节的式(22)中所有的 2 替换为 3 时,可得到一个求解正数 $R(0<R<1)$ 的三进制表示形式的方法,并求解下列正数的三进制表示形式。

(a) 1/3　　(b) 1/2　　(c) 1/10　　(d) 11/27

16. 证明当将本节的式(8)中所有的 2 替换为 5 时,可得到一个求解正整数的五进制表示形式的方法,并求解下列整数的五进制表示形式。

(a) 10　　(b) 35　　(c) 721　　(d) 734

17. 证明当将本节的式(22)中所有的 2 替换为 5 时,可得到一个求解正数 $R(0<R<1)$ 的五进制表示形式的方法,并求解下列正数的五进制表示形式。

(a) 1/3　　(b) 1/2　　(c) 1/10　　(d) 154/625

1.3　误差分析

读者将从数值分析的练习中学习到非常重要的一点,即认识到计算结果并不确切地等于数学结果,因为存在不少隐含方法会破坏数值结果的精度。对这一点的理解将有助于研究人员实现和开发正确的数值算法。

定义 1.7　设 $\hat{p}$ 是 p 的近似值,则**绝对误差**是 $E_p=|p-\hat{p}|$,简称误差。**相对误差**是 $R_p=|p-\hat{p}|/|p|$,其中 $p\neq 0$。 ▲

误差仅仅是真实值与近似值之间的差,而相对误差在很大程度上取决于真实值。

例 1.14　找出下面 3 种情况下的误差和相对误差。设 $x=3.141592$,$\hat{x}=3.14$,则误差为

$$E_x=|x-\hat{x}|=|3.141592-3.14|=0.001592 \tag{1a}$$

相对误差为

$$R_x=\frac{|x-\hat{x}|}{|x|}=\frac{0.001592}{3.141592}=0.00507$$

设 $y=1000000$,$\hat{y}=999996$,则误差为

$$E_y=|y-\hat{y}|=|1000000-999996|=4 \tag{1b}$$

相对误差为

$$R_y=\frac{|y-\hat{y}|}{|y|}=\frac{4}{1000000}=0.000004$$

设 $z=0.000012$,$\hat{z}=0.000009$,则误差为

$$E_z=|z-\hat{z}|=|0.000012-0.000009|=0.000003 \tag{1c}$$

相对误差为

$$R_z=\frac{|z-\hat{z}|}{|z|}=\frac{0.000003}{0.000012}=0.25$$ ■

在式(1a)中,E_x 和 R_x 之间没有太大差别,都可以用来判别 $\hat{x}$ 的精度。在式(1b)中,y 值的数量级为 10^6,误差 E_y 很大,相对误差很小。在这种情况下,可以认为 $\hat{y}$ 是 y 的好的近似值。在式(1c)中,z 值的数量级为 10^{-6},误差 E_z 是 3 种情况中最小的,但相对误差是最大的。如果用百分数形式来表示,相对误差为 25%,这样 $\hat{z}$ 是 z 的不良近似值。当 $|p|$ 远离 1 时(大于或小于),相对误差 R_p 比误差 E_p 能更好地表示近似值的精确程度。由于相对误差直接处理尾数,所以浮点表示主要采用相对误差。

定义 1.8　如果 d 是满足下列不等式的最大正整数,则称数 $\hat{p}$ 近似 p 时具有 d 位**有效数字**。

$$\frac{|p-\hat{p}|}{|p|}<\frac{10^{1-d}}{2} \tag{2}$$

▲

例 1.15　判断例 1.14 中近似值的有效数字。

(3a) 如果 $x=3.141592$,$\hat{x}=3.14$,则 $|x-\hat{x}|/|x|=0.000507<10^{-2}/2$。因此,$\hat{x}$ 近似 x 时具有 3 位有效数字。

(3b) 如果 $y=1000000$,$\hat{y}=999996$,则 $|y-\hat{y}|/|y|=0.000004<10^{-5}/2$。因此,$\hat{y}$ 近似 y 时具有 6 位有效数字。

(3c) 如果 $z=0.000012$,$\hat{z}=0.000009$,则 $|z-\hat{z}|/|z|=0.25<10^{-0}/2$。因此,$\hat{z}$ 近似 z 时具有 1 位有效数字。 ■

1.3.1 截断误差

截断误差通常指的是,用一个基本表达式替换一个相当复杂的算术表达式时所引入的误差。这个术语从用截断泰勒级数替换一个复杂表达式的技术中衍生而来。例如,无穷泰勒级数

$$e^{x^2}=1+x^2+\frac{x^4}{2!}+\frac{x^6}{3!}+\frac{x^8}{4!}+\cdots+\frac{x^{2n}}{n!}+\cdots$$

可以用它的前 5 项 $1+x^2+\frac{x^4}{2!}+\frac{x^6}{3!}+\frac{x^8}{4!}$ 来替代,这样就可以求数值积分的近似值。

例 1.16 设有 $\int_0^{1/2}e^{x^2}dx=0.544987104184=p$,当用截断泰勒级数 $P_8(x)=1+x^2+\frac{x^4}{2!}+\frac{x^6}{3!}+\frac{x^8}{4!}$ 替代被积函数 $f(x)=e^{x^2}$ 时,确定积分近似值的精度。

解:替换后的积分近似值为

$$\begin{aligned}\int_0^{1/2}\left(1+x^2+\frac{x^4}{2!}+\frac{x^6}{3!}+\frac{x^8}{4!}\right)dx&=\left(x+\frac{x^3}{3}+\frac{x^5}{5(2!)}+\frac{x^7}{7(3!)}+\frac{x^9}{9(4!)}\right)_{x=0}^{x=1/2}\\&=\frac{1}{2}+\frac{1}{24}+\frac{1}{320}+\frac{1}{5376}+\frac{1}{110592}\\&=\frac{2109491}{3870720}=0.544986720817=\hat{p}\end{aligned}$$

因为

$$10^{-5}/2>|p-\hat{p}|/|p|=7.03442\times10^{-7}>10^{-6}/2$$

近似值 $\hat{p}$ 近似真实值 $p=0.544987104184$ 时有 6 位有效数字。$y=f(x)=e^{x^2}$,$y=P_8(x)$ 的曲线图以及当 $0\leqslant x\leqslant\frac{1}{2}$ 时曲线下的区域面积如图 1.7 所示。 ■

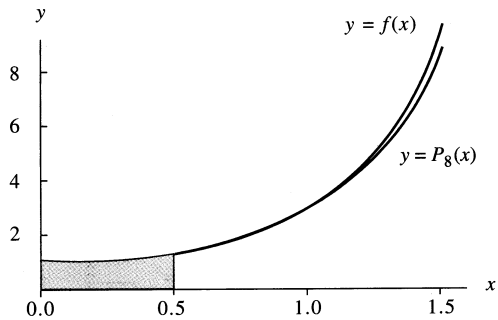


图 1.7 $y=f(x)=e^{x^2}$,$y=P_8(x)$ 的函数曲线图以及当$0\leqslant x\leqslant\frac{1}{2}$时曲线下的区域面积

1.3.2 舍入误差

计算机表示的实数受限于尾数的固定精度,因此有时并不能确切地表示真实值,这一类型的误差称为**舍入误差**。在上一节中,计算机对存储的实数 $\frac{1}{10}=0.0\,\overline{0011}_2$ 进行了截断,对实际存储在计算机中的数的最后一位进行了四舍五入。综上所述,由于计算机硬件只支持有限位机器数的运算,导致舍入误差在计算中被引入和传播。