数值方法

----误差的分析

秦 品 乐

中北大学大数据学院

误差的种类

模型误差

观测误差

舍入误差

截断误差

模型误差

　　在建立数学模型过程中，不可能将所有因素均考虑在内，必然要进行必要的简化，这就带来了与实际问题的误差。

观测误差

数学模型中的参数往往靠观测所得，由观测数据带来的误差。

舍入误差

计算机的字长是有限的，每一步运算均需要进行四舍五入，由此产生的误差称为传入误差。

截断误差

数值方法可能运用近似方法表示准确数值运算或数量而引入的误差，称为截断误差。

本课程需要考虑的误差

本课程学习基于以下假设：

　　模型误差和观测误差为0，主要讨论**舍入误差**和**截断误差**带来的影响。

舍入误差----有效数字

为何要引入有效数字，何为有效数字?

引入有效数字的概念是为了正式规定数值的可靠程度。一个数的有效数字是指可以放心使用的那些数字。

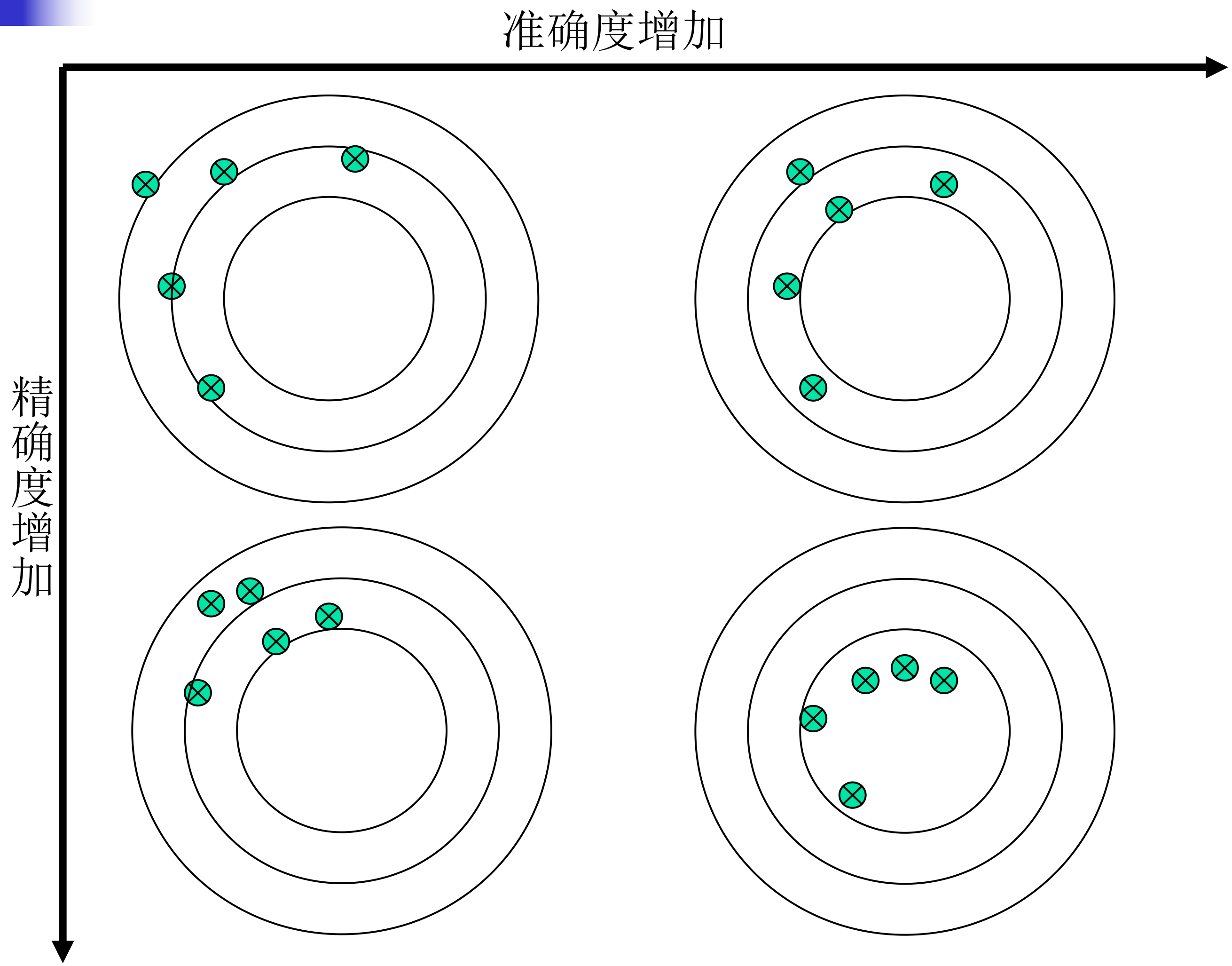
有效数字的概念有两个重要的含义:

1. 与伞兵问题中介绍的一样，数值方法得到的是近似解，所以必须建立准则来规定近似结果的可信度。
2. 因为计算机只能保留有限个有效数字，因此如π，$e,\sqrt{7}$ 之类的数无法准确表示。

舍入误差----有效数字

保留有效数字后，被省略的部分称为舍入误差。

准确度和精度



与计算和测量相关的误差可以用**精度**和**准确度**描述。

精度指测量值相互之间的集中程度；**准确度**指测量值与真值之间的接近程度。

误差的定义

数值误差包括截断误差和舍入误差，当用近似方法表示数学过程时就会出现截断误差；当用有限个有效数字表示确定的数时就会引起舍入误差。在两种情况下，准确结果与近似值之间的关系可表示为：

$$真值=近似值+误差$$

经变换可得

$$E_t = 真值-近似值$$

Et称为绝对误差

绝对误差限

我们设

$$|E_t|=|\text{真值-近似值}|\le \varepsilon_s$$

ε_s称为绝对误差限。四舍五入的绝对误差限是其末位的半个单位，即

$$\varepsilon_s = 0.5\times 10^{m-n}$$

其中**m**为科学计数法指数项，**n**为有效数字

相对误差的定义

绝对误差表示法有一定的缺陷，如同样是**1CM**的误差，测量一个铆钉，而不是一个桥梁，那么它就要大的多。如果要考虑这种数量级，方法之一就是将误差相对真值进行归一化处理。即

$$\text{真小数相对误差}=\frac{\text{真误差}}{\text{真值}}$$

经变换可得

$$\varepsilon_t=\frac{\text{真误差}}{\text{真值}}100\%$$

ε_t 称为真百分比相对误差

相对误差

对于数值方法，只有相关的函数可以用解析方法求解时，才能知道真值。因此有下式近似误差表示方法：

$$\varepsilon_{\mathrm{a}}=\frac{\text{真误差}}{\text{近似值}}100\%$$

思考：为什么可以用ε_{a}来近似表示真百分比相对误差ε_{t}?

相对误差限

相对误差限

$$\varepsilon_{\mathrm{a}}=\left|\frac{\text{真误差}}{\text{近似值}}\right|\leq\frac{\text{绝对误差限}}{|\text{近似值}|}=\varepsilon_{\mathrm{r}}$$

实例见书**6**页例**1-4**

误差实例

例：假设对一座桥梁和一个铆钉的长度进行了测量，测量的结果分别为**9999cm**和**9cm**。如果真值分别为**10000cm**和**10cm**，计算两种情况下（**1**）真误差；（**2**）真百分比相对误差

解：

桥梁的测量误差为 $E_t = 10000\text{-}9999 = 1cm$

铆钉的测量误差为 $E_t = 10\text{-}9 = 1cm$

桥梁的真百分比相对误差为 $\varepsilon_t = \dfrac{1}{10000} = 0.01\%$

铆钉的真百分比相对误差为 $\varepsilon_t = \dfrac{1}{10} = 10\%$

工程中的迭代

在数值计算中，缺乏真值信息的情况下需要确定误差的估计量，当前的近似值总是建立在前一个近似值基础上的。因此当前迭代结果的误差可以由下式确定：

$$\varepsilon_{\mathrm{a}} = \frac{\text{当前近似值-前个近似值}}{\text{当前近似值}} \times 100\%$$

注意，如有下面准则成立，那么我们就可以保证至少有**n**位有效数字是正确的：

$$\varepsilon_{\mathrm{s}} = (0.5 \times 10^{2-n})\%$$

迭代方法的误差估计

在数学中，常将函数表示为无穷级数，如指数函数可以用下式计算

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

可以看到，当加入的序列越多，对真值的估计就越好。上式称为麦克劳林级数展开

迭代方法的误差估计

如果想计算 $e^{0.5}$的值，在每加入一个新的项以后，分别真误差和近似百分比相对误差。注：$e^{0.5} = 1.648721$ 加入新项后，直到近似估计误差值小于预先设定的误差准则，其中误差准则必须符合3位有效数字的要求。

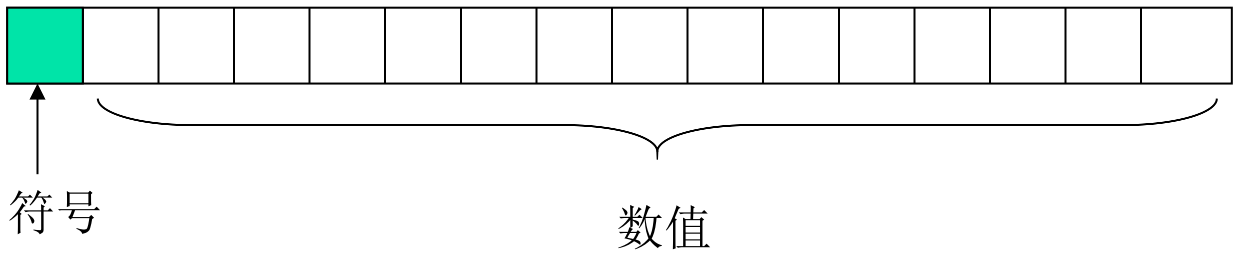
解：首先确定误差准则，以保证结果至少有3位有效数字

$$e_s = (0.5 \times 10^{2-3})\% = 0.05\%$$

舍入误差

计算机中表示信息的基本单位是字节，一个字节由**8**位组成。

整数的表示方法（以**16**位机讲解）



用**matlab**求**16**位机十进制整数可以表示的范围

舍入误差

计算机中整数一般采用补码形式表示。

浮点数的表示

在计算机中，小数一般采用浮点数形式表示。在这种表示方法中，一个数表示为小数部分----尾数，整数部分----指数

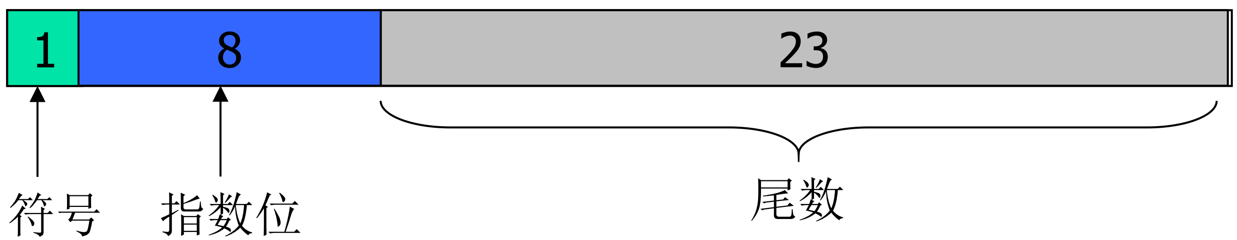
$$m \times b^e$$

M为尾数**(** $\frac{1}{b} \le m < 1$**)**，**b**为数制的基，**e**为指数。如**156.78**表示为十进制的浮点数为**:** 0.15678×10^3

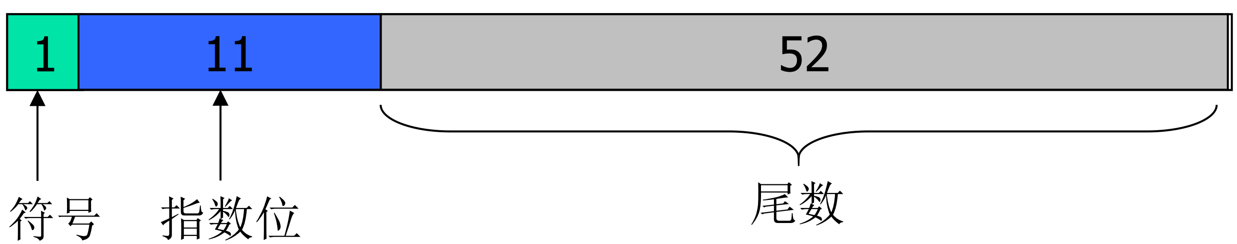
舍入误差

浮点数的计算机表示

Float类型的表示方法



Double类型的表示方法



舍入误差

小数**8.25**在计算机中如何表示

0	1000 0010	000 0100 0000 0000 0000 0000

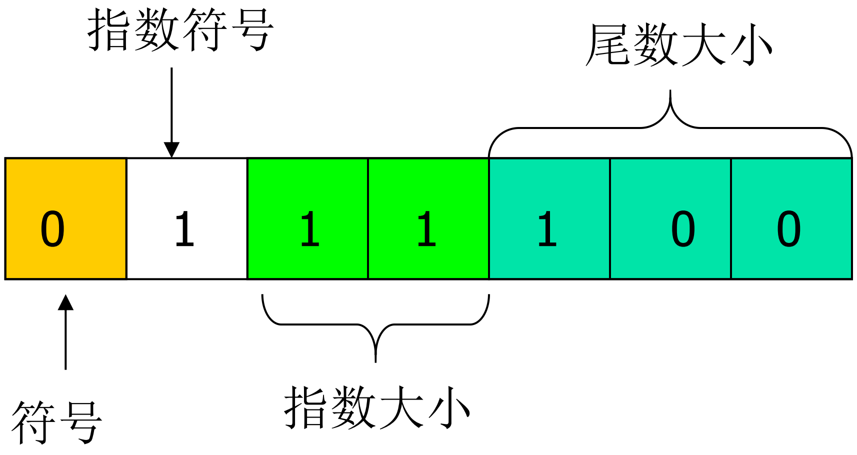
0	127+3=130	00001

0	3	00001

$$1.00001 \times 2^{3}$$

浮点溢出问题

浮点数表示方法仅能表示有限范围内的数。如果超过上限或下限则溢出。



指数大小为？，尾数为？

浮点溢出问题

$$0111100 = (1\times 2^{-1} + 0\times 2^{-2} + 0\times 2^{-3})\times 2^{-3} = (0.0625)_{10}$$

$$0111101 = (1\times 2^{-1} + 0\times 2^{-2} + 1\times 2^{-3})\times 2^{-3} = (0.078125)_{10}$$

$$0111110 = (1\times 2^{-1} + 1\times 2^{-2} + 0\times 2^{-3})\times 2^{-3} = (0.093750)_{10}$$

$$0111111 = (1\times 2^{-1} + 1\times 2^{-2} + 1\times 2^{-3})\times 2^{-3} = (0.109375)_{10}$$

等价的十进制数在空间上是均匀分布的，相邻两个数之间的间隔为0.015625。要想继续增大该数，必须将指数减小到10

$$0110100 = (1\times 2^{-1} + 0\times 2^{-2} + 0\times 2^{-3})\times 2^{-2} = (0.12500)_{10}$$

$$0110101 = (1\times 2^{-1} + 0\times 2^{-2} + 1\times 2^{-3})\times 2^{-2} = (0.156250)_{10}$$

$$0110110 = (1\times 2^{-1} + 1\times 2^{-2} + 0\times 2^{-3})\times 2^{-2} = (0.187500)_{10}$$

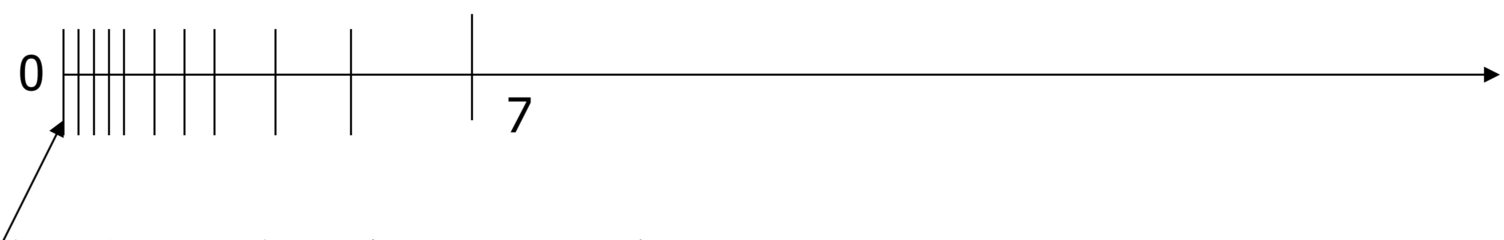
$$0110111 = (1\times 2^{-1} + 1\times 2^{-2} + 1\times 2^{-3})\times 2^{-2} = (0.218750)_{10}$$

浮点溢出问题

相邻两个数之间的间隔为**0.03125**，这种模式不断重复下去，直到达到最大数：

$$0011111 = (1\times 2^{-1} + 1\times 2^{-2} + 1\times 2^{-3})\times 2^{3} = (7)_{10}$$

数之间的间隔随着数大小的增加而增大



在**0**值附近有一个“洞”处会下溢

大数与小数相加

假设将一个小数**0.0010**与一个大数**4000**相加

$0.4000 \qquad 10^4$

$0.0000001 \qquad 10^4$

––––––––––––––––

$0.4000001 \qquad 10^4$

减性抵消

此术语表示当两个几乎相等的浮点数相减时引入的舍入误差

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

当 $b^2 >> 4ac$ ，分子可能会很小