Course: **Numerical Analysis**

1. ***Applied Numerical Methods for Engineers and Scientists***

   S. S. Rao, Prentice Hall

2. ***Numerical Analysis*/D. Kincaid and W. Cheney**

3. ***Numerical Analysis*/R.L. Burden and J.D. Faires**

Lecturer: 黃美嬌 教授 Rm.729 @ 33662696

mjhuang@ntu.edu.tw

Teaching Assistant: 翁健洲 Rm.533 @ 33664498

b96522106@ntu.edu.tw

Office hours: 週三下午2:00~3:30

週四上午 11:00~12:30

Contents:

學習目標:

1. 數值方法原理（設計、誤差分析、穩定性、效率等）
2. 高階程式語言（Fortran or C）與程式邏輯設計
3. 報告撰寫

學期評分:

1. 四份報告 80%（15%+20%+20%+25%）
2. 期末考 15%
3. 平常表現 5%

1 Introduction and Review

§ Notation

$(a,b) \equiv$ the set of all real numbers which are $> a$ and $< b$

$[a,b] \equiv$ the set of all real numbers which are $\geq a$ and $\leq b$

$(a,b] \equiv$ the set of all real numbers which are $> a$ and $\leq b$

R = the set of all real numbers

R^2 = the set of all points on the real two dimensional space

R^3 = the set of all points on the real three dimensional space

$C(X)$ = the set of all functions which are continuous in X.

$C^m(X)$ = the set of all functions which m derivatives exist and are continuous in X.

e.g. $f(x)=|x| \in C(R)$ but $\notin C^1(R)$

e.g. $f(x)=\exp(x) \in C(R), C^1(R), C^2(R), \cdots, C^\infty(R)$

e.g. all polynomials $\in C(R), C^1(R), C^2(R), \cdots, C^\infty(R)$

$$C^\infty(R) \subset \cdots \subset C^2(R) \subset C^1(R) \subset C(R) \equiv C^0(R)$$

§ Big *O*

$$f(x)=O(g(x)) \quad \text{as} \quad x \to x_0$$

if there exists a neighborhood D of x_0 and a constant C such that

$$|f(x)| \le C|g(x)| \quad \text{for} \quad x \in D$$

e.g. $\dfrac{n+1}{n^2}=O\left(\dfrac{1}{n}\right)$ as $n \to \infty$

e.g. $\sin(x)-\left(x-\dfrac{x^3}{6}\right)=O(x^5)$ as $x \to 0$

§ small *o*

$$f(x)=o(g(x)) \quad \text{as } x \to x_0 \quad \text{if} \quad \lim_{x \to x_0} \frac{f(x)}{g(x)}=0$$

e.g. $\dfrac{1}{x\log(x)}=o\left(\dfrac{1}{x}\right)$ as $x \to \infty$

e.g. $\cos(x)-1=o(x)$ as $x \to 0$

§ limit

$\lim_{x \to x_0} f(x)=L$ if given $\varepsilon > 0$, $\exists \delta > 0$ such that $|f(x)-L| < \varepsilon$ whenever $0 < |x-x_0| < \delta$

e.g. $\lim_{x \to 2} x^2 = 4$ (take $\delta=\varepsilon/(5+\varepsilon)$)

e.g. $\lim_{x \to 0} \dfrac{|x|}{x}$ does not exist.

§ Continuity

$f(x)$ is said to be continuous at x_0 if $\lim_{x \to 0} f(x) = f(x_0)$

e.g. $\lim_{x\to 2} x^2 = 4$ (take $\delta=\varepsilon/(5+\varepsilon)$)

for $|x-2|<\delta=\varepsilon/(5+\varepsilon)$: $\quad 2-\frac{\varepsilon}{5+\varepsilon}<x<2+\frac{\varepsilon}{5+\varepsilon}$

$$-\frac{4\varepsilon}{5+\varepsilon}+\frac{\varepsilon^2}{(5+\varepsilon)^2}<x^2-4<\frac{4\varepsilon}{5+\varepsilon}+\frac{\varepsilon^2}{(5+\varepsilon)^2}$$

$$RHS=\frac{4\varepsilon}{5+\varepsilon}+\frac{\varepsilon^2}{(5+\varepsilon)^2}=\frac{20\varepsilon+5\varepsilon^2}{(5+\varepsilon)^2}=\varepsilon\cdot\frac{5}{(5+\varepsilon)}\cdot\frac{4+\varepsilon}{(5+\varepsilon)}<\varepsilon$$

$$LHS=-\frac{4\varepsilon}{5+\varepsilon}+\frac{\varepsilon^2}{(5+\varepsilon)^2}=-\frac{20\varepsilon+3\varepsilon^2}{(5+\varepsilon)^2}>-\frac{20\varepsilon+5\varepsilon^2}{(5+\varepsilon)^2}=-\varepsilon\cdot\frac{5}{(5+\varepsilon)}\cdot\frac{4+\varepsilon}{(5+\varepsilon)}>-\varepsilon$$

$$-\varepsilon<x^2-4<\varepsilon \quad\Rightarrow\quad |x^2-4|<\varepsilon$$

§ Differentiability

$f(x)$ is said to be differentiable at x_0 if $\lim_{x\to x_0}\frac{f(x)-f(x_0)}{x-x_0}$ exists.

e.g. $f(x)=|x|$ is differentiable everywhere except at $x=0$

Differentiability implies continuity.

§ Riemann integral

$$\int_a^b f(x)\,dx=\lim_{n\to\infty}\sum_{i=1}^{n} f(z_i)\Delta x_i$$

where $a\le x_0\le x_1\le\cdots\le x_n\le b$

$\Delta x_i=x_i-x_{i-1},\ \ z_i\in[x_{i-1},x_i]$

§ Intermediate Value Theorem

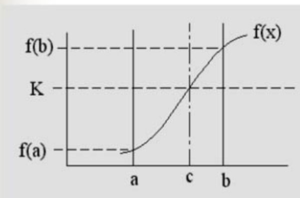
If $f(x)\in C[a,b]$ and K is any number between $f(a)$ and $f(b)$,

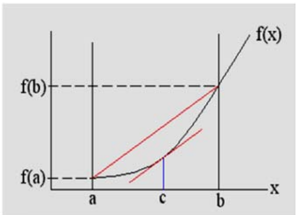
then $\exists c\in(a,b)\ni f(c)=K$

§ Mean Value Theorem

If $f(x)\in C[a,b]$ and f is differentiable on (a,b),

then $\exists c\in(a,b)\ni(b-a)f'(c)=f(b)-f(a)$





§ Taylor's Theorem with Integral Remainder

If $f(x)\in C^{n+1}[a,b]$, then for any points x and c in $[a,b]$,

$$f(x)=\sum_{k=0}^{n}\frac{1}{k!}f^{(k)}(c)(x-c)^k+R_n(x)$$

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

§ Taylor's Theorem with Lagrange Remainder

If $f(x) \in C^{n+1}[a,b]$, then for any points x and c in $[a,b]$,

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(c)(x-c)^k + E_n(x)$$

$$E_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi)(x-c)^{n+1}$$

for some ξ between x and c.

§ Taylor's Theorem in two variables

If $f(x,y) \in C^{n+1}([a,b]\times[c,d])$,

then for any point $(x+dx, y+dy)$ in $[a,b]\times[c,d] \subseteq R^2$,

$$f(x+dx, y+dy) = \sum_{k=0}^{n} \frac{1}{k!}\left(dx\frac{\partial}{\partial x} + dy\frac{\partial}{\partial y}\right)^k f(x,y) \;+\; E_n(x,y)$$

$$E_n(x,y) = \frac{1}{(n+1)!}\left(dx\frac{\partial}{\partial x} + dy\frac{\partial}{\partial y}\right)^{n+1} f(x+\theta dx, y+\theta dy)$$

for some $0 < \theta < 1$

$$\left(dx\frac{\partial}{\partial x} + dy\frac{\partial}{\partial y}\right)^k$$

$$\left(dx\frac{\partial}{\partial x} + dy\frac{\partial}{\partial y}\right)^0 f(x,y) = 1\cdot f(x,y) = f(x,y)$$

$$\left(dx\frac{\partial}{\partial x} + dy\frac{\partial}{\partial y}\right)^1 f(x,y) = dx\frac{\partial f}{\partial x} + dy\frac{\partial f}{\partial y}$$

$$\left(dx\frac{\partial}{\partial x} + dy\frac{\partial}{\partial y}\right)^2 f(x,y) = \left(dx^2\frac{\partial^2}{\partial x^2} + 2dxdy\frac{\partial}{\partial x}\frac{\partial}{\partial y} + dy^2\frac{\partial^2}{\partial y^2}\right) f(x,y)$$

$$= dx^2\frac{\partial^2 f}{\partial x^2} + 2dxdy\frac{\partial^2 f}{\partial x\partial y} + dy^2\frac{\partial^2 f}{\partial y^2}$$

§ Machine numbers

~ represented by a finite number of binary digits (bits)

e.g. a single-precision real number is usually represented by a word = 4 bytes = 32 bits

e.g. 16-base system: $\pm a \times 16^b$

X | XXXXXXX | XXX⋯XXX

1 bit	7 bits	24 bits
sign	exponential part	fractional part

• exponential part (7 bits):

# of numbers that can be composed = 2^7 = 128

$$\begin{cases} 64 \ \text{ for zero and positive exponents: } 0,1,2,\cdots 63 \\ 64 \ \text{ for negative exponents: } -1,-2,\cdots,-64 \end{cases}$$

0000000 ⇨ 0 ⇨ -64

0000001 ⇨ 1 ⇨ -63

...

1000000 ⇨ 64 ⇨ 0

…

1111111 ⇨ 127 ⇨ 63

• fractional part: (24 bits)

$$X_1X_2X_3\cdots X_{22}X_{23}X_{24} \equiv X_1\cdot 2^{-1}+X_2\cdot 2^{-2}+\cdots+X_{24}\cdot 2^{-24}$$

$$\text{maximum} \quad 111\cdots 111 \equiv 2^{-1}+2^{-2}+2^{-3}+\cdots+2^{-24} \approx \frac{2^{-1}}{1-2^{-1}} = 1$$

e.g. $\underline{0}\ \underline{1000010}\ \underline{101100\ldots000}$

$$1000010 = 2^1+2^6 = 66 \quad \Rightarrow 16^{66-64} = 16^2$$

$$101100...000 = 2^{-1}+2^{-3}+2^{-4} = 0.6875$$

$$\underline{0}\ \underline{1000010}\ \underline{101100...000} \equiv +0.6875\times 16^2 = 176 \quad (10\,\text{base})$$

§ Machine numbers --- 32 bits

~ # of machine numbers = $2^{32} = 1024^{3.2} \approx 4.3\times 10^9 \approx 4G$

the maximum one = 0(1111111)(11……1) $\approx 16^{63} \approx 10^{76}$

the minimum nonzero one = 0(0000000)(00……01)

$$= 2^{-24}\times 16^{-64} \approx 10^{-84}$$

WARMING MESSAGE:

OVERFLOW ~ appears a number which absolute value > 10^{76}

UNDERFLOW ~ appears a nonzero number which absolute value < 10^{-84}

~ a finite set of real numbers

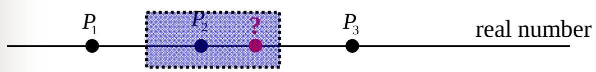
§ Machine numbers --- discrete number system

$\underline{0}\ \underline{10000000}\ \underline{101100\ ...000} = 0.6875 \ (P_2)$

The two nearby machine numbers are:

$\underline{0}\ \underline{10000000}\ \underline{101100\ ...001} = 0.6875 + 2^{-24} \ (P_3)$

$\underline{0}\ \underline{10000000}\ \underline{101011\ ...11} = 0.6875 - 2^{-24} \ (P_1)$



~ all represented by P_2

rounding error $\equiv |P - P_2|$

§ Rounding Errors

Suppose a machine can represent a number up to k digits in the following form: $\pm d_1 d_2 \cdots d_k \times 10^n$, $1 \le d_1 \le 9$ and $0 \le d_i \le 9, i = 2,3,\cdots,k$

How to present π=3.141592653589793…..? Machine-dependent!

e.g. k = 7

$$\text{chopping method}: fl(\pi) = 0.3141592 \times 10^1$$

$$\text{rounding method}: fl(\pi) = 0.3141593 \times 10^1$$

$$error = |\pi - fl(\pi)|$$

§ Rounding Errors

Round-off errors are unavoidable

and accumulate as computations go on.

$E_n \equiv$ magnitude of rounding error after n subsequent operations

$*$ linear growth : $E_n \approx CnE_0$ for some constant C

Usually unavoidable but acceptable as long as C and E_0 are sufficiently small.

$*$ exponential growth : $E_n \approx C^n E_0$ for some constant $C > 1$

Overflow!

example: compute the series $P_n = \dfrac{1}{3^n}$ with single-precision real numbers

Method 1

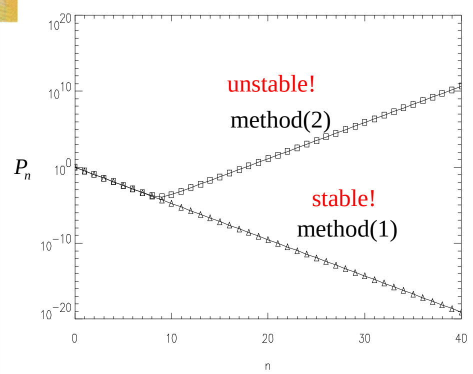
```
P(0)=1
DO n=1,100
    P(n)=1./3.*P(n-1)
END DO
```

Method 2

```
P(0)=1
P(1)=1./3.
DO n=2,100
    P(n)=10./3.*P(n-1)–P(n-2)
END DO
```

Method 1: $fl(P_n) = fl\left(fl\left(\frac{1}{3}\right) * fl(P_{n-1}) \right)$

Method 2: $fl(P_n) = fl\left(fl\left(\frac{10}{3}\right) * fl(P_{n-1}) - fl(P_{n-2}) \right)$



Method 1: $fl(P_n) = fl\left(fl\left(\frac{1}{3}\right) * fl(P_{n-1}) \right)$

Method 2: $fl(P_n) = fl\left(fl\left(\frac{10}{3}\right) * fl(P_{n-1}) - fl(P_{n-2}) \right)$

Method 2

```
P(0)=1
P(1)=1./3.
DO n=2,100
    P(n)=10./3.*P(n-1)–P(n-2)
END DO
```

$$P_n = \frac{1}{3^n}$$

formula correct!

$$P_n = \rho^n$$

$$\rho^n = \frac{10}{3}\rho^{n-1} - \rho^{n-2}$$

$$\rho^2 - \frac{10}{3}\rho + 1 = 0$$

$$(\rho - 3)\left(\rho - \frac{1}{3}\right) = 0$$

$$P_n = A \cdot 3^n + B \cdot \frac{1}{3^n}$$

$$\left.\begin{array}{l} P_0 = 1 = A + B \\ P_1 = \frac{1}{3} = 3A + \frac{B}{3} \end{array}\right\} \Rightarrow \left\{\begin{array}{l} A = 0 \\ B = 1 \end{array}\right.$$

Disasters due to rounding error

http://www.ma.utexas.edu/users/arbogast/disasters.html

1. The Patriot and the Scud.

On February 25, 1991, during the Gulf War, a Patriot missile defense system let a Scud get through. It hit a barrack, killing 28 people. The problem was in the differencing of floating point numbers obtained by converting and scaling an integer timing register.

Disasters due to rounding error

http://www.ma.utexas.edu/users/arbogast/disasters.html

2. The short flight of the Ariane 5.

On June 4, 1996, the first Ariane 5 was launched. All went well for 36 seconds. Then the Ariane veered off course and self-destructed. The problem was in the Inertial Reference System, which produced an operation exception trying to convert a 64-bit floating-point number to a 12-bit integer. It sent a diagnostic word to the On-Board Computer, which interpreted it as flight data. Finis.

Ironically, the computation was done by legacy software from the Ariane 4, and its results were not needed after lift-off.

Disasters due to rounding error

http://www.ma.utexas.edu/users/arbogast/disasters.html

3. The Vancouver Stock Exchange.

In 1982, the Vancouver Stock Exchange instituted a new index initialized to a value of 1000.000. The index was updated after each transaction. Twenty two months later it had fallen to 520. The cause was that the updated value was truncated rather than rounded. The rounded calculation gave a value of 1098.892.

Disasters due to rounding error

http://www.ma.utexas.edu/users/arbogast/disasters.html

4. Parliamentary elections in Schleswig-Holstein.

In German parliamentary elections, a party with less than 5.0% of the vote cannot be seated. The Greens appeared to have a cliff-hanging 5.0%, until it was discovered (after the results had been announced) that they really had only 4.97%. The printout was to two figures, and the actual percentage was rounded to 5.0%.

Ways of Avoiding Rounding Errors:

- **Reduce # of computations as many as possible.**

$$\pi + e = 3.1415926535... + 2.718281828... = 5.859874482...$$

$$\pi * e = 3.1415926535... * 2.718281828... = 8.539734222...$$

7 digits + rounding method:

$$fl(\,fl(\pi) + fl(e)) = fl(3.141593 + 2.718282) = 5.859875$$

$$\begin{aligned} fl(\,fl(\pi) * fl(e)) &= fl(3.141593 * 2.718282) \\ &= fl(8.5397375703...) = 8.539736 \end{aligned}$$

- **Avoid substraction of two nearly equal numbers.**

$$fl(x) - fl(y) = 0.3141593 \times 10^{1} - 0.3141291 \times 10^{1} = 0.3020000 \times 10^{-3}$$

~ lose 4 digits of significance

(Any further calculations can have only 3, instead of 7, digits of significance.)

- **Avoid dividing by a small number.**

original rounding error $= \delta$ exact number $= z = fl(z) + \delta$

divided by a small number $\varepsilon = 10^{-6}$

$$\text{rounding error} = \left|\frac{z}{\varepsilon} - \frac{fl(z)}{\varepsilon}\right| = \left|\frac{\delta}{\varepsilon}\right| = 10^{6}|\delta|$$