Dr. M.H. Shahnas

Computational Physics

Useful Books : Numerical Analysis

Richard L. Burden, J. Douglas Faires

Numerical Recipes

William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery

Mathematical Preliminaries

1-1) Review of Calculus

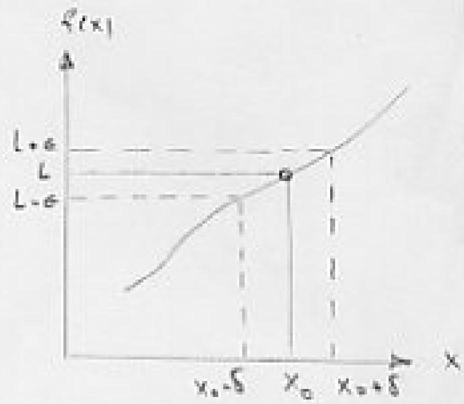
Def. 1.1)

Let ; f be a func. defined on a set X of real numbers;

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

if $\forall \; \epsilon > 0 \quad \exists \; \delta > 0$ such that $|f(x) - L| < \epsilon$

whenever $x \in X$ and $0 < |x - x_0| < \delta$



Def. 1.2)

Let ; f be a func. defined on a set X of real numbers, and $x_0 \in X$;

f is said to be <u>continuous</u> at x_0 if $\lim_{x \to x_0} f(x) = f(x_0)$

The func. f is said to be continuous on X if it is continuous at each number in X.

Def 1.3)

Let; $\{x_n\}_{n=1}^{\infty}$ be an infinite sequence of <u>real</u> or <u>complex</u> numbers. The sequence is said to <u>Converge</u> to a number x (called the limit); if $\forall \epsilon > 0 \ \exists \ N(\epsilon)$ (positive integer) such that $n > N(\epsilon)$ implies $|x_n - x| < \epsilon$.

The notation; $\lim_{n \to \infty} x_n = x$

or $x_n \to x$ as $n \to \infty$

means that the sequence $\{x_n\}_{n=1}^{\infty}$ Converges to x

Theo. 1.4)

If f is a func. defined on a set X of real numbers and $x_0 \in X$, then the following are equivalent:

a) f is continuous at x_0;

b) if $\{x_n\}_{n=1}^{\infty}$ is any sequence in X converging to x_0,

then; $$\lim_{n \to \infty} f(x_n) = f(x_0)$$

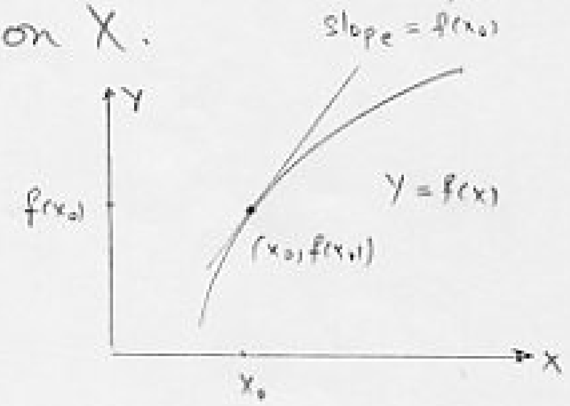
Def. 1.5)

If f is a func. defined in an open interval containing x_0, f is said to be differentiable at x_0, if

$$\lim \frac{f(x)-f(x_0)}{x-x_0} \qquad \text{exisits.}$$

$$\lim_{x \to x_0} \frac{f(x)-f(x_0)}{x-x_0} = f'(x_0)$$

A func. that has a derivative at each number in a set X is said to be differentiable on X.



Theo. 1.6)

If the func. f is differentiable at x_0, then f is continuous at x.

Remark: $C(X)$ denotes the set of all funcs. continuous on X.

$C^n(X)$ denotes the set of all funcs. that have n continuous derivatives on X

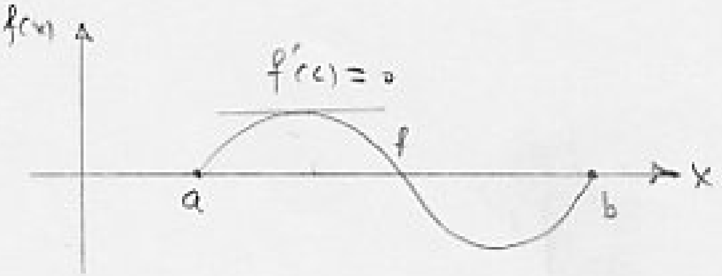
$C^\infty(X)$ denotes the set of all funcs. that have derivatives of all orders on X.

Theo. 1.7) (Rolle's Theorem)

Suppose $f \in \overset{\text{continuous}}{C}[a,b]$ and f is differentiable on (a,b).

If $f(a) = f(b) = 0 \xrightarrow{\text{then}} \exists\, c,\ a<c<b$ with

$f'(c) = 0$

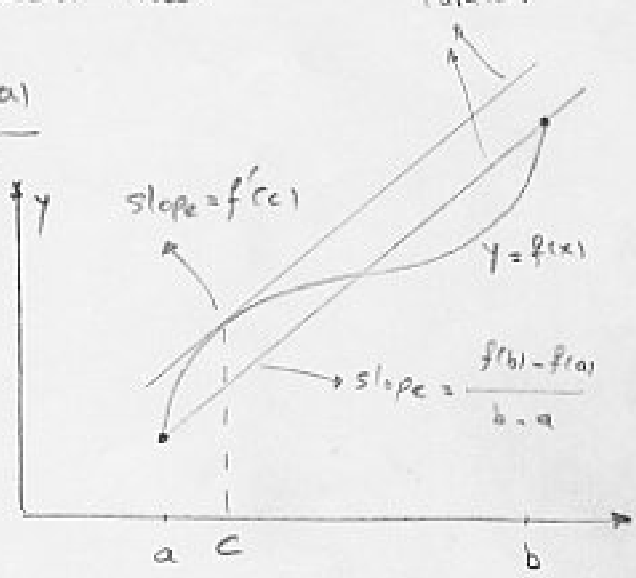


Theo. 1.8) (Mean Value Theorem)

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

$\xrightarrow{\text{then}} \exists\, c,\ a<c<b$ such that

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

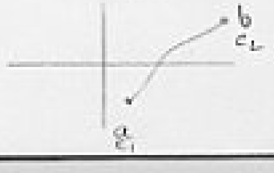


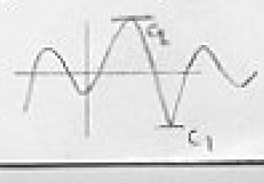
Theo. 1.9) (Extreme Value Theorem)

If $f \in C[a,b] \xrightarrow{\text{then}} \exists\, c_1, c_2 \in [a,b]$ with

$f(c_1) \leq f(x) \leq f(c_2)$, $\forall\ x \in [a,b]$.

If, in addition, f is differentiable on (a,b), $\xrightarrow{\text{then}}$ the numbers c_1 and c_2 occur either at endpoints of $[a,b]$ or where $f' = 0$





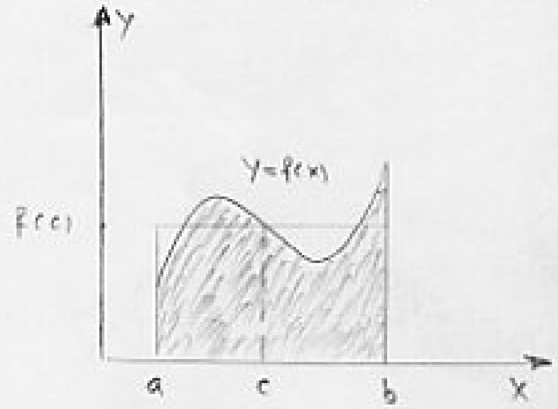
Theo 1.10) (Weighted Mean Value Theorem for Integrals)

If $f \in C[a,b]$, and, g is integrable on $[a,b]$ and $g(x)$ does not change sign on $[a,b]$ $\xrightarrow{\text{then}}$

$\exists\, c$, $a < c < b$, such that:

$$\int_a^b f(x)\, g(x)\, dx = f(c) \int_a^b g(x)\, dx$$

When $g(x) \equiv 1 \longrightarrow f(c) = \frac{1}{b-a} \int_a^b f(x)\,dx$ (average value of f over $[a,b]$)

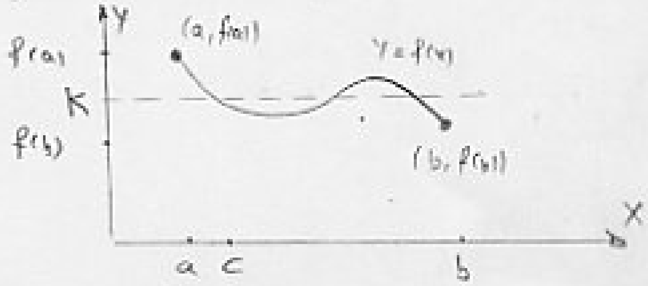


Theo. 1.11) (Generalized Rolle's Theorem)

Let: $f \in C[a,b]$ be n-times differentiable on (a,b).
If $f = 0$ at $n+1$ distinct numbers $x_0 \ldots x_n$ in $[a,b]$.
$\xrightarrow{\text{then}}$ $\exists\, c \in (a,b)$ with $f^{(n)}(c) = 0$

Theo. 1.12) (Intermediate Value Theorem)

If $f \in C[a,b]$ and K is any number between $f(a)$ and $f(b)$, $\xrightarrow{\text{then}}$ $\exists\, c \in (a,b)$ for which $f(c) = K$



Ex. — Show that $x^5 - 2x^3 + 3x^2 - 1 = 0$ has a sol. in the interval $[0,1]$

Sol.

$$f(x) = x^5 - 2x^3 + 3x^2 - 1 \qquad \text{a polynomial continuous on } [0,1]$$

Since $f(0) = -1 < 0 < +1 = f(1)$

$\xrightarrow{\text{Intermediate Value Theo.}} \exists\, x$ with $0 < x < 1$

for which $x^5 - 2x^3 + 3x^2 - 1 = 0$

Theo. 1.13) (Taylor's Theorem)

Suppose $f \in \underset{\text{continuous derivatives}}{C^n}[a,b]$ and $\underset{\text{derivative}}{f^{(n+1)}}$ exists on $[a,b]$.

Let; $x_0 \in [a,b]$.

$\forall\, x \in [a,b]$, $\exists\, \xi(x)$ between x_0 and x with

$$f(x) = P_n(x) + R_n(x)$$

where;

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \quad \left(\begin{array}{l}\text{n-th Taylor polynomial}\\ \text{for } f \text{ about } x_0\end{array}\right)$$

and

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)^{n+1} \quad \left(\begin{array}{l}\text{remainder term}\\ \text{or truncation error}\end{array}\right)$$

Remark: $n \to \infty$ (Taylor series)

$x_0 = 0$ (Maclaurin polynomial or series)

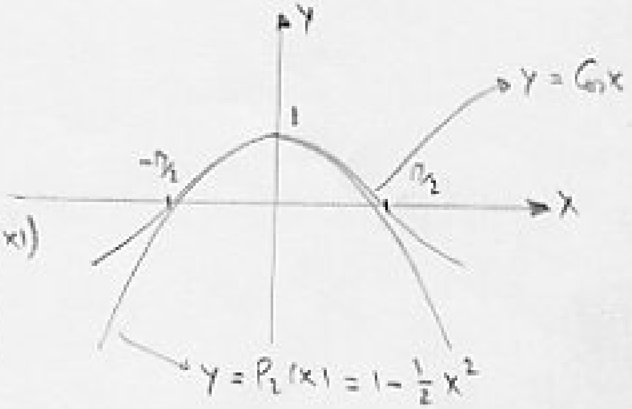
Ex. 2 – Determine a) the second and b) the third Taylor polynomial for $f(x) = \cos x$ about $x_0 = 0$, and use these polynomials to approximate $\cos(0.01)$

Sol.

Since; $f \in C^{\infty}(\mathbb{R})$, the previous theorem can be used for any $n > 0$

a) $n = 2$ and $x_0 = 0$; $\cos x = 1 - \frac{1}{2}x^2 + \frac{1}{6}x^3 \sin \xi(x)$

where $\xi(x)$ is a number between 0 and x.



with $x = 0.01$

$$\cos(0.01) = 1 - \frac{1}{2}(0.01)^2 + \frac{1}{6}(0.01)^3 \sin(\xi(x))$$

$$= 0.99995 + (0.1\bar{6}) \cdot 10^{-6} \sin(\xi(x))$$

where $0 < \xi(x) < 0.01$ (rad.)

Since $|\sin(\xi(x))| < 1$ $\longrightarrow$ We can use 0.99995 as an approximation to $\cos(0.01)$ with assurance of at least six-decimal place accuracy.

Using standard tables $\cos(0.01) = 0.99995000042$

b) For the third Taylor polynomial about $x_0 = 0$;

$$\cos x = 1 - \frac{1}{2}x^2 + 0 + \frac{1}{24}x^4 \cos(\xi(x))$$

where $0 < \xi(x) < 0.01$ and $f'''(0) = 0$

the approximation → $\cos(0.01) = 0.99995$ as before

the accuracy → nine-decimal-place, since

$$\left|\frac{1}{24}x^4 \cos\xi(x)\right| \leq \frac{1}{24}(0.01)^4(1) = 4.2 \times 10^{-10}$$

1.2 Round off Errors and Computer Arithmetic

Floating-point forms

$$a = \underbrace{\pm 0.d_1 d_2 \cdots d_k}_{\text{mantissa}} \times \underbrace{10^n}_{\text{characteristic}} \qquad 1 \leq d_1 \leq 9 \;,\; 0 \leq d_i \leq 9$$

$i = 2, \cdots k$

$-78 \leq n \leq 76$

$$y = 0.d_1 d_2 \cdots d_k d_{k+1} d_{k+2} \cdots \times 10^n \qquad \text{(any positive real number)}$$

There are two ways to obtain the floating-point form;

i) Chopping $\quad fl(y) = 0.d_1 d_2 \cdots d_k \times 10^n$

The mantissa is terminated at k decimal digits, just by chopping.

ii) Rounding $\quad fl(y) = 0.\delta_1 \delta_2 \cdots \delta_k \times 10^n$

$$\begin{cases} \text{if } d_{k+1} \geq 5 \; ; & d_k \to d_k + 1 \text{ and chop off right most digits} \\ \;\;\; d_{k+1} < 5 \; ; & d_k \to d_k \quad \text{" " " " " "} \end{cases}$$

<u>Ex.</u>

$\pi = 3.14159265 \cdots$

$\pi = 0.314159265 \cdots \times 10^1$ normalized decimal form

$fl(\pi) = 0.31415 \times 10^1 = 3.1415$ (chopping)

$fl(\pi) = (0.31415 + 0.00001) \times 10^1 = 3.1416$ (rounding)

Round off error:

$$= |y - fl(y)|$$ (regardless of whether the rounding or chopping method is used)

Def. 1.14) If p^* is a approximation to p:

absolute error $= |p - p^*|$

relative " $= \dfrac{|p - p^*|}{|p|}$ $\qquad p \neq 0$

Ex.

a) $p = 0.3000 \times 10^{1}$ $\qquad p^* = 0.3100 \times 10^{1}$

ab. error $= 0.1$, rel. error $= 0.333\bar{3} \times 10^{-1}$

b) $p = 0.3000 \times 10^{-3}$ $\qquad p^* = 0.3100 \times 10^{-3}$

ab. error $= 0.1 \times 10^{-4}$, rel. error $= 0.333\bar{3} \times 10^{-1}$

c) $p = 0.3000 \times 10^{4}$ $\qquad p^* = 0.3100 \times 10^{4}$

ab error $= 0.1 \times 10^{3}$, rel. error $= 0.333\bar{3} \times 10^{-1}$

Relative error in computer: (Chopping)

$$= \left| \frac{Y - fl(Y)}{Y} \right|$$

$$Y = 0.d_1 d_2 \ldots d_k d_{k+1} \ldots \times 10^n$$

$$\left| \frac{Y - fl(Y)}{Y} \right|_{\text{chopping}} = \left| \frac{0.d_1 d_2 \ldots d_k d_{k+1} \ldots \times 10^n - 0.d_1 d_2 \ldots d_k \times 10^n}{0.d_1 d_2 \ldots \times 10^n} \right|$$

$$= \left| \frac{0.d_{k+1} d_{k+2} \ldots \times 10^{n-k}}{0.d_1 d_2 \ldots \times 10^n} \right| = \left| \frac{0.d_{k+1} d_{k+2} \ldots}{0.d_1 d_2 \ldots} \right| \times 10^{-k}$$

Since $d_1 \neq 0 \longrightarrow \text{Min(denominator)} = 0.1$

and $\text{Max(numerator)} = 1$

$$\longrightarrow \left| \frac{Y - fl(Y)}{Y} \right| \leq \frac{1}{0.1} \times 10^{-k} = 10^{-k+1}$$

In a similar manner, a bound for the relative error when using k-digit rounding arithmetic is:

$$\left| \frac{Y - fl(Y)}{Y} \right|_{\text{rounding}} = \left| \frac{0.d_1 d_2 \ldots d_k d_{k+1} \ldots \times 10^n - 0.\delta_1 \delta_2 \ldots \delta_k \times 10^n}{0.d_1 d_2 \ldots \times 10^n} \right|$$

$$\leq 0.5 \times 10^{-k+1}$$

Def. 1.15)

The number p^* is said to approximate p to t significant digits (or figures) if t is the largest nonnegative integer for which

$$\left| \frac{p - p^*}{p} \right| < 5 \times 10^{-t}$$

Other Source of error:

In addition to inaccurate representation of numbers (floating point), the arithmatic performed in a computer is not exact.

Assume finite-digit arithmetic given by:

$$X \oplus Y = fl(fl(x) + fl(y)) \qquad X \otimes Y = fl(fl(x) \times fl(y))$$

$$X \ominus Y = fl(fl(x) - fl(y)) \qquad X \oslash Y = fl(fl(x) \div fl(y))$$

Ex.

Suppose; $x = \frac{1}{3}$, $y = \frac{5}{7}$ (use 5-digit chopping)

$fl(x) = 0.33333 \times 10^0$ $\qquad fl(y) = 0.71428 \times 10^0$

Operation	Result	Actual value	Abs. error	Relative error
$X \oplus Y$	0.10476×10^1	$22/21$	0.190×10^{-4}	0.182×10^{-4}
$Y \ominus X$	0.38095×10^0	$8/21$	0.238×10^{-5}	0.625×10^{-5}
$X \otimes Y$	0.23809×10^0	$5/21$	0.524×10^{-5}	0.220×10^{-4}
$Y \oslash X$	0.21428×10^1	$15/7$	0.571×10^{-4}	0.267×10^{-4}

Max. (Rel. error) $= 0.267 \times 10^{-4}$ satisfactory

Now consider;

$u = 0.714251$, $v = 98765.9$, $w = 0.111111 \times 10^{-4}$

$\longrightarrow fl(u) = 0.71425 \times 10^0$, $fl(v) = 0.98765 \times 10^5$, $fl(w) = 0.11111 \times 10^{-4}$

Operation	Result	Actual value	Abs. error	Rel. error
$y \ominus u$	0.30000×10^{-4}	0.34714×10^{-4}	0.471×10^{-5}	0.136
$(y \ominus u) \oslash w$	0.27000×10^{1}	0.31243×10^{1}	0.424	0.136
$(y \ominus u) \otimes v$	0.29629×10^{1}	0.34285×10^{1}	0.465	0.136
$u \oplus v$	0.98765×10^{5}	0.98766×10^{5}	0.161×10^{1}	0.163×10^{-4}

Other most common error source;

Subtraction of nearly equal numbers:

$$fl(x) = 0.d_1 d_2 \ldots d_p \alpha_{p+1} \alpha_{p+2} \ldots \alpha_k \times 10^n$$

$$fl(y) = 0.d_1 d_2 \ldots d_p \beta_{p+1} \beta_{p+2} \ldots \beta_k \times 10^n$$

$$fl(fl(x) - fl(y)) = \underbrace{0.\sigma_{p+1} \sigma_{p+2} \ldots \sigma_k}_{k-p \text{ digits of significance}} \times 10^{n-p}$$

where

$$0.\sigma_{p+1} \sigma_{p+2} \ldots \sigma_k = 0.\alpha_{p+1} \alpha_{p+2} \ldots \alpha_k - 0.\beta_{p+1} \beta_{p+2} \ldots \beta_k$$

However in most calculation devices;

$$fl(fl(x) - fl(y)) = 0.\sigma_{p+1} \sigma_{p+2} \ldots \sigma_k \underbrace{00000}_{p \text{ digits}} \times 10^{n-p}$$

Errors in chain;

suppose;

$$fl(\text{Some arithmetic calculation}) = Z + \delta$$

(having finite digit)

δ: error of arithmetic cal.

Z: actual floating poit form of arithmetic cal.

$$\frac{Z}{\epsilon} \approx fl\left[\frac{(Z+\delta)}{fl(\epsilon)}\right] \qquad (\epsilon \neq 0)$$

suppose; $\epsilon = 10^{-n}$ $\quad n > 0$

then $\longrightarrow$ $\frac{Z}{\epsilon} = Z \times 10^{n}$

and $fl\left[\frac{(Z+\delta)}{fl(\epsilon)}\right] = (Z+\delta) \times 10^{n}$

abs. error $= |\delta| \times 10^{n}$

Thus; if a finite-digit representation or calculation introduces an error, further enlargement of the error occurs when dividing by a number with small magnitude (or equivaletely, when multiplying by a number with large magnitude).

Ex.

$$ax^2+bx+c=0 \qquad a\neq 0$$

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

Consider; $x^2+62.10x+1=0$

$x_1=-0.01610723 \qquad x_2=-62.08390$ (approx.)

Since $b^2 >> 4ac \longrightarrow b\approx\sqrt{b^2-4ac}$

Now suppose 4-digit rounding arithmethic;

$$\sqrt{b^2-4ac}=\sqrt{(62.10)^2-4.000}=\sqrt{3856.-4.000}=\sqrt{3852.}=62.06$$

So;

$$fl(x_1)=\frac{-b+\sqrt{b^2-4ac}}{2a}=\frac{-62.10+62.06}{2.000}=\frac{-0.04000}{2.000}=-0.02000$$

is an approximation to $x_1=-0.01610723$;

$$\text{Rel. error}=\left|\frac{-0.01611-0.02000}{-0.01611}\right|\simeq 2.4\times10^{-1}$$

On the other hand;

$$fl(x_2)=\frac{-b-\sqrt{b^2-4ac}}{2a}=\frac{-62.10-62.06}{2.000}=\frac{-124.2}{2.000}=-62.10$$

is an approximation to $x_2=-62.08$

$$\text{Rel. error}=\left|\frac{-62.08+62.10}{-62.08}\right|\simeq 3.2\times10^{-4}$$

Remedy;

$$X_1 = \frac{-b+\sqrt{b^2-4ac}}{2a}\left(\frac{-b-\sqrt{b^2-4ac}}{-b-\sqrt{b^2-4ac}}\right) = \frac{-2c}{b+\sqrt{b^2-4ac}}$$

$$\rightarrow fl(x_1) = \frac{-2.000}{62.10+62.06} = \frac{-2.000}{124.2} = -0.01610$$

$$\rightarrow \text{Rel. error} = 6.2\times10^{-4}$$

The rationalization tequnique can also be applied to give the alternate form for x_2;

$$x_2 = \frac{-2c}{b-\sqrt{b^2-4ac}}$$

This would be the form to use if b were a negative number.

But notice;

$$fl(x_2) = \frac{-2c}{b-\sqrt{b^2-4ac}} = \frac{2.000}{62.10-62.06} = \frac{-2.000}{0.04000} = -50.00$$

here then are two sources of error;

$$\begin{cases} 1\text{- subtraction of nearly equal numbers} \\ 2\text{- division by the small number.} \end{cases}$$

Ex.

Evaluate $f(x) = x^3 - 6x^2 + 3x - 0.149$ at $x = 4.71$ using three-digit arithmetic.

	x	x^2	x^3	$6x^2$	$3x$
Exact	4.71	22.1841	104.487111	133.1046	14.13
Threedigit (chopping)	4.71	22.1	104.	132.	14.1
" " (rounding)	4.71	22.2	105.	133.	14.1

Exact: $f(4.71) = 104.487111 - 133.1046 + 14.13 - 0.149 = -14.636489$

Three-digit (chopping) $f(4.71) = 104. - 132. + 14.1 - 0.149 = -14.0$

Three-digit (rounding) $f(4.71) = 105. - 133. + 14.1 - 0.149 = -14.0$

Rel. error $= \left| \dfrac{-14.636489 - 14.0}{-14.636489} \right| \approx 0.04$

(for both method)

Alternative approach; (nesting method)

$f(x) = x^3 - 6x^2 + 3x - 0.149 = ((x-6)x + 3)x - 0.149$

$\rightarrow f(4.71) = ((4.71 - 6)4.71 + 3)4.71 - 0.149 = -14.5$ 3-digit chopping

$f(4.71) = -14.6$ 3-digit rounding

Rel. error $= \left| \dfrac{-14.636489 + 14.5}{-14.636489} \right| \simeq 0.0093$, Rel. error $\simeq \left| \dfrac{-14.636489 + 14.6}{-14.636489} \right| \simeq 0.0025$

chopping

1.3 Convergence:

Ex. The Taylor polynomial $P_N(x)$ for $f(x) = \ln x$ expanded about $x_0 = 1$ is

$$P_N(x) = \sum_{i=1}^{N} \frac{(-1)^{i+1}}{i} (x-1)^i$$

$\ln 1.5 = 0.40546511$ to the 8-decimal places.

We want to find $\text{Min.}(N) = ?$ for:

$$|\ln 1.5 - P_N(1.5)| < 10^{-5}$$

without using the Taylor polynomial truncation formula.

From calculus we know:

If $A_N = \sum_{n=1}^{N} a_n$ and $\lim_{N \to \infty} A_N = A$ (with decreasing terms)

then $\longrightarrow |A - A_N| \leq |a_{N+1}|$

Stablity:

One criterion we will impose on an algorithm whenever possible is that <u>small</u> changes in the <u>initial</u> data produce correspondingly <u>small</u> changes in the <u>final</u> results. An algorithm that satisfies this property is called <u>stable</u>; it is <u>unstable</u> when this criterion is not fulfilled.

Rounding Error Growth and its Connection to Algorithm stability:

Suppose: ϵ: an error introduced at some stage in the calculation

E_n: the error after n subsequent operations

Def. 1.6) Suppose that E_n represents the growth of an error after n subsequent operations.

If $|E_n| \approx C n \epsilon$ where C = const. indep. of n

$\to$ the growth of error is said to be <u>linear</u>.

If $|E_n| \approx k^n \epsilon$ for some $k > 1$

$\to$ the growth of error is called <u>exponential</u>.

Ex.— The sequence $P_n = (\frac{1}{3})^n$, $n>0$, can be generated recursively by letting $P_0=1$ and defining $P_n = (\frac{1}{3}) P_{n-1}$, whenever $n>1$.

In 5-digit rounding arithmetic:

$P_0 = 0.10000 \times 10^{1}$ $\quad P_1 = 0.33333 \times 10^{0}$ $\quad P_2 = 0.11111 \times 10^{0}$

$P_3 = 0.37036 \times 10^{-1}$ $\quad P_4 = 0.12345 \times 10^{-1}$ . . .

Rounding error replacing $\frac{1}{3}$ by 0.33333 produces an error of only $(0.33333)^n \times 10^{-5}$ in the nth term of the sequence.

$$\text{Rel. error} = \left| \frac{\frac{1}{3} - 0.33333}{\frac{1}{3}} \right| = 1 \times 10^{-5}$$

Abs. error = Rel. error × Data

Abs error $= (0.33333) \times 10^{-5}$ for $n=1$

Another way to generate the sequence:

Define $P_0 = 1$, $P_1 = \frac{1}{3}$

$$P_n = \left(\frac{10}{3}\right) P_{n-1} - P_{n-2} \qquad n \geq 2 \qquad (1)$$

This method is quite clearly <u>unstable</u>.

Note that formula (1) is satisfied whenever P_n is of the form:

$$P_n = C_1\left(\frac{1}{3}\right)^n + C_2 3^n$$

$P_n = \left(\frac{1}{3}\right)^n$ $P_0 = 1$
$\forall$ $n \geq 1$

n	Computed P_n	Correct value P_n
0	0.10000×10^{1}	0.10000×10^{1}
1	0.33333×10^{0}	0.33333×10^{0}
2	0.11110×10^{0}	0.11111×10^{0}
3	0.37000×10^{-1}	0.37037×10^{-1}
4	0.12230×10^{-1}	0.12346×10^{-1}
5	0.37660×10^{-2}	0.41152×10^{-2}
6	0.32300×10^{-3}	0.13717×10^{-2}
7	-0.26893×10^{-2}	0.45725×10^{-3}
8	-0.92872×10^{-2}	0.15242×10^{-3}

Verification:

$$\frac{10}{3}P_{n-1} - P_{n-2} = \frac{10}{3}\left[C_1\left(\frac{1}{3}\right)^{n-1} + C_2 3^{n-1}\right] - \left[C_1\left(\frac{1}{3}\right)^{n-2} + C_2 3^{n-2}\right]$$

$$= C_1\left[\frac{10}{3}\left(\frac{1}{3}\right)^{n-1} - \left(\frac{1}{3}\right)^{n-2}\right] + C_2\left[\frac{10}{3}3^{n-1} - 3^{n-2}\right]$$

$$= C_1\left(\frac{1}{3}\right)^n + C_2 3^n = P_n$$

To have $\begin{cases} P_0 = 1 \\ P_1 = \frac{1}{3} \end{cases}$ in equ (1) $\xrightarrow{\text{must}}$ $\begin{cases} C_1 = 1 \\ C_2 = 0 \end{cases}$

However, in the 5-digit approx. $\begin{cases} P_0 = 0.10000 \times 10^{1} \\ P_1 = 0.33333 \times 10^{0} \end{cases}$

$\xrightarrow{\text{then}}$ $\begin{cases} C_1 = 0.10000 \times 10^{1} \\ C_2 = -0.12500 \times 10^{-5} \end{cases}$

This small change C_1 and C_2 results a rounding error:

$$\underbrace{0.10000 \times 10^{1}\left(\frac{1}{3}\right)^n}_{\substack{\text{decreasing} \\ \text{with } n \text{ (exponential)}}} + \underbrace{(-0.12500 \times 10^{-5}) \times 3^n}_{\substack{\text{increasing} \\ \text{with } n \text{ (exponential)}}}$$

Remedy:

i) Using double- or multi-precision arithmetic.

ii) Using a more suitable algorithm.

Def. 1.17 -

Suppose $\{\alpha_n\}_{n=1}^{\infty}$ is a sequence that converges to a number α. We say that $\{\alpha_n\}_{n=1}^{\infty}$ converges to α with rate of convergence $O(\beta_n)$, where $\{\beta_n\}_{n=1}^{\infty}$ is another sequence with $\beta_n \neq 0$ for each n, if

$$\frac{|\alpha_n - \alpha|}{|\beta_n|} \leqslant K \qquad \text{for sufficiently large } n$$

and K = const. indep. of n. This is indicated by writing

$\alpha_n = \alpha + O(\beta_n)$

or

$\alpha_n \rightarrow \alpha$ with rate of convergence $O(\beta_n)$

Ex. -

Suppose that the sequences $\{\alpha_n\}$ and $\{\hat{\alpha}_n\}$ are described by

$$\alpha_n = \frac{n+1}{n^2} \qquad \forall\, n \geq 1 \qquad n\text{: int.}$$
$$\hat{\alpha}_n = \frac{n+3}{n^3}$$

Although;

$$\lim_{n\to\infty} \alpha_n = 0 \qquad \lim_{n\to\infty} \hat{\alpha}_n = 0$$

the sequence $\{\hat{\alpha}_n\}$ converges to this limit much faster than $\{\alpha_n\}$.

If we let $\beta_n = \frac{1}{n}$, $\hat{\beta}_n = \frac{1}{n^2}$

$$\left| \frac{\alpha_n - 0}{\beta_n} \right| = \left| \frac{(n+1)/n^2 - 0}{(1/n)} \right| = \frac{n+1}{n} \leq 2$$

$$\left| \frac{\hat{\alpha}_n - 0}{\hat{\beta}_n} \right| = \left| \frac{(n+3)/n^3 - 0}{(1/n^2)} \right| = \frac{n+3}{n} \leq 4$$

$\longrightarrow$ $\alpha_n = 0 + O(\frac{1}{n})$ while $\hat{\alpha}_n = 0 + O(\frac{1}{n^2})$

This implies that;

The rate of convergence of $\{\alpha_n\}$ is similar to the convergence of $\{\frac{1}{n}\}$ to zero

while " " " " " $\{\hat{\alpha}_n\}$ " " " " $\{\frac{1}{n^2}\}$ " "

This concept generalizes to functions as follows:

Def 1.18 –

If $\lim_{x \to 0} F(x) = L$,

the convergence is said to be $O(G(x))$

if $\exists$ $K > 0$ indep. of x for which

$$\frac{|F(x) - L|}{|G(x)|} \leq K \quad \text{for sufficiently \underline{small} } |x| > 0$$

This situation is indicated by writing:

$$F(x) = L + O(G(x))$$

or $\underset{x \to 0}{F(x) \longrightarrow L}$ with rate of convergence $O(G(x))$

Ex. — We found before using a third Taylor polynomial,

$$\cos x = 1 - \frac{1}{2}x^2 + \frac{1}{24}x^4 \cos \xi(x)$$

$$0 \leq \xi(x) \leq x \quad \text{if } x > 0$$

$$\longrightarrow \cos x + \frac{1}{2}x^2 = 1 + \frac{1}{24}x^4 \cos \xi(x) \qquad 0 \geq \xi(x) \geq x \quad \text{if } x \leq 0$$

$$\longrightarrow \cos x + \frac{1}{2}x^2 = 1 + O(x^4)$$

Since $\left| \frac{(\cos x + \frac{1}{2}x^2) - 1}{x^4} \right| = \left| \frac{1}{24} \cos \xi(x) \right| \leq \frac{1}{24}$

The implication is that $\cos x + \frac{1}{2}x^2$ converges to its limit, 1, at approximately the same rate that x^4 converges to zero.

Chapter 11

Boundary-Value Probs. for Ordinary Differential Equs.

11.3 Finite Difference Methods for Linear Probs.

$$y'' = p(x)\, y' + q(x)\, y + r(x) \qquad a \leq x \leq b \;,\; y(a) = \alpha \;,\; y(b) = \beta$$

First, we select an integer $N > 0$

and divide the interval $[a, b]$ into $(N+1)$ equal subintervals

$$x_i = a + ih \quad \text{(mesh points)} \qquad h = \frac{b-a}{N+1} \qquad i = 0, 1, \ldots N+1$$

$$x_0 = a \,,\; x_{N+1} = b$$

$$y''(x_i) = p(x_i)\, y'(x_i) + q(x_i)\, y(x_i) + r(x_i)$$

$i = 1, 2, \ldots N$ interior mesh points

Expanding y in third-deg. Taylor polynomial about x_i evaluated at x_{i+1} and x_{i-1} :

$$y(x_{i+1}) = y(x_i + h) = y(x_i) + h\, y'(x_i) + \frac{h^2}{2} y''(x_i) + \frac{h^3}{6} y'''(x_i) + \frac{h^4}{24} y^{(4)}(\xi_i^+) \qquad (1)$$

for some ξ_i^+, $\quad x_i \leq \xi_i^+ \leq x_{i+1}$

$$y(x_{i-1}) = y(x_i - h) = y(x_i) - h\, y'(x_i) + \frac{h^2}{2} y''(x_i) - \frac{h^3}{6} y'''(x_i) + \frac{h^4}{24} y^{(4)}(\xi_i^-) \qquad (2)$$

for some ξ_i^- $\quad x_{i-1} \leq \xi_i^- \leq x_i$

assuming $y \in C^4[x_{i-1}, x_{i+1}]$

$$(1)(2) \longrightarrow y''(x_i) = \frac{1}{h^2}\left[y(x_{i+1}) - 2y(x_i) + y(x_{i-1})\right] - \frac{h^2}{24}\left[y^{(4)}(\xi_i^+) + y^{(4)}(\xi_i^-)\right]$$

Using intermediate value Theo.; $\begin{cases} y^{(4)}(\xi_i^-) \equiv f(a), \; y^{(4)}(\xi_i^+) \equiv f(b) \\ \frac{1}{2}\left[y^{(4)}(\xi_i^-) + y^{(4)}(\xi_i^+)\right] \equiv k \text{ (between } f(a) \text{ and } f(b)) \\ \rightarrow \xi_i \equiv c \end{cases}$

$$y''(x_i) = \frac{1}{h^2}\left[y(x_{i+1}) - 2y(x_i) - y(x_{i-1})\right] - \frac{h^2}{12} y^{(4)}(\xi_i) \qquad (3)$$

$x_{i+1} \leqslant \xi_i \leqslant x_{i+1}$

Equ. (3) is called the Centered-difference formula for $y''(x_i)$

Similarly:

$$y'(x_i) = \frac{1}{2h}\left[y(x_{i+1}) - y(x_{i-1})\right] - \frac{h^2}{6} y'''(\eta_i) \qquad (4)$$

$x_{i-1} \leqslant \eta_i \leqslant x_{i+1}$ centered-diff. formula for $y'(x_i)$

The use of (3) and (4) in the main differential equ. yields:

$$\frac{y(x_{i+1}) - 2y(x_i) + y(x_{i+1})}{h^2} = p(x_i)\left[\frac{y(x_{i+1}) - y(x_{i-1})}{2h}\right] + q(x_i)\,y(x_i)$$

$$+ r(x_i) - \frac{h^2}{12}\left[2p(x_i)\,y'''(\eta_i) - y^{(4)}(\xi_i)\right]$$

A Finite-Difference method with truncation error of order $O(h^2)$ results by using this equ. together with the boundary conds. $y(a)=\alpha$, and $y(b)=\beta$ to define

$$w_0 = \alpha \quad , \quad w_{N+1} = \beta$$

and

$$\left(\frac{2w_i - w_{i+1} - w_{i-1}}{h^2}\right) + P(x_i)\left(\frac{w_{i+1} - w_{i-1}}{2h}\right) + q(x_i)\,w_i = -r(x_i)$$

$i = 1, \ldots N$

or;

$$-\left(1+\frac{h}{2}p(x_i)\right)w_{i-1} + \left(2+h^2 q(x_i)\right)w_i - \left(1-\frac{h}{2}p(x_i)\right)w_{i+1} = -h^2 r(x_i)$$

and the resulting system of equs. is expressed in the tridiagonal $N \times N$-matrix form;

$$A\,w = b$$

$$A = \begin{bmatrix} 2+h^2 q(x_1) & -1+\frac{h}{2}p(x_1) & 0 & \cdots & \cdots & 0 \\ -1-\frac{h}{2}p(x_2) & 2+h^2 q(x_2) & -1+\frac{h}{2}p(x_2) & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & -1+\frac{h}{2}p(x_{N-1}) \\ 0 & \cdots & \cdots & 0 & -1-\frac{h}{2}p(x_N) & 2+h^2 q(x_N) \end{bmatrix}$$

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ \\ w_N \end{bmatrix} \qquad b = \begin{bmatrix} -h^2 r(x_1) + (1 + \frac{h}{2} p(x_1))\, w_0 \\ -h^2 r(x_2) \\ \vdots \\ -h^2 r(x_{N-1}) \\ -h^2 r(x_N) + (1 - \frac{h}{2} p(x_N))\, w_{N+1} \end{bmatrix}$$

Linear Finite-Difference Algorithm 11.3

To approximate the sol. of the boundary-value problem:

$$y'' = p(x)\, y' + q(x)\, y + r(x) \qquad a \leq x \leq b, \quad y(a) = \alpha, \; y(b) = \beta$$

Input; a, b, α, β, N

Output: approxs. w_i to $y(x_i)$ for $i = 0, 1, \ldots N+1$

S1 $\quad h = (b-a)/(N+1)$

$x = a + h$

$a_1 = 2 + h^2 q(x)$

$b_1 = -1 + (h/2)\, p(x)$

$d_1 = -h^2 r(x) + (1 + (h/2)\, p(x))\, \alpha$

S2 $\quad$ Do $i = 2, N-1$

$x = a + ih$

$a_i = 2 + h^2 q(x)$

$b_i = -1 + (h/2)\, p(x)$

$c_i = -1 - (h/2)\, p(x)$

$d_i = -h^2 r(x)$

Continue

S3 $\quad x = b - h$

$a_N = 2 + h^2 q(x)$

$c_N = -1 - (h/2)\, p(x)$

$d_N = -h^2 r(x) + (1 - (h/2)\, p(x))\, \beta$

To be continued by the algorithm for solving linear system of equs..

Remark: Non-linear differential equ.:

$y'' = f(x, y, y')$ general form

Ex.: $y'' = \frac{1}{8}(32 + 2x^3 - yy')$

Ex. $y'' + \frac{1}{x} y' - \frac{y}{x^2} = 3$, $y(1) = 2$, $y(2) = 3$

$\longrightarrow a = 1,\ b = 2,\ \alpha = 2,\ \beta = 3$

$p(x) = \frac{-1}{x}$, $q(x) = +\frac{1}{x^2}$ $\quad r(x) = 3$

$N = 5 \rightarrow h = 0.2,\ x_1 = 1.2,\ x_2 = 1.4,\ x_3 = 1.6,\ x_4 = 1.8,\ x_5 = 2$

$$-2.0278\,Y_1 + 1.0833\,Y_2 = 0.12 - 1.8333$$

$$0.9286\,Y_1 - 2.0204\,Y_2 + 1.0714\,Y_3 = 0.12$$

$$0.9375\,Y_2 - 2.0156\,Y_3 + 1.0625\,Y_4 = 0.12$$

$$0.9444\,Y_3 - 2.0123\,Y_4 = 0.12 - 3.1667$$

X	analytic Sol.	Numerical Sol.
1	2	2
1.2	1.9067	1.9083
1.4	1.9886	1.9904
1.6	2.2100	2.2115
1.8	2.5511	2.5520
2	3	3

Approximating Eigenvalues:

Jacobi Method;

If A is symmetric, it can be shown there exists an orthogonal matrix P such that:

$$P^T A P = D$$

where D is diagonal matrix whose diagonal elements are the eigenvalues of A.

Jacobi's method transforms A into diagonal form by annihilating its off-diagonal elements one-by-one.

It makes use of plane rotation matrices $R(p,q)$ which are basically unit matrices except for elements:

$$r_{pp} = C\theta \qquad r_{pq} = -S\theta$$

$$r_{qp} = S\theta \qquad r_{qq} = C\theta$$

$R(p,q)$ orthogonal matrix;

then a tr. of the form $\bar{A} = R^T A R$ preserves the eigenvalues of A.

θ is chosen in such a way that $\bar{a}_{pq}$ is reduced to zero.

$$\bar{A} = R^T A R = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & c & \cdots & s & \cdots & 0 \\ \vdots & & \vdots & & & & \\ 0 & \cdots & -s & \cdots & c & \cdots & 0 \\ \vdots & & & & & & \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1p} & \cdots & a_{1q} & \cdots & a_{1n} \\ a_{p1} & \cdots & a_{pp} & \cdots & a_{pq} & \cdots & a_{pn} \\ \vdots & & & & & & \\ a_{q1} & \cdots & a_{qp} & \cdots & a_{qq} & \cdots & a_{qn} \\ a_{n1} & \cdots & a_{np} & \cdots & a_{nq} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} 1 & \cdots & 0 & & 0 & \cdots & 0 \\ 0 & \cdots & c & \cdots & -s & \cdots & 0 \\ \vdots & & & & & & \\ 0 & \cdots & s & \cdots & c & \cdots & 0 \\ \vdots & & & & & & \\ 0 & & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}$$

when $c = \cos\theta$, $s = \sin\theta$

$$(R^T A)_{pp} = c\,a_{pp} + s\,a_{qp}$$

$$(R^T A)_{pq} = c\,a_{pq} + s\,a_{qq}$$

and so

$$\bar{a}_{pq} = (c\,a_{pp} + s\,a_{qp})(-s) + (c\,a_{pq} + s\,a_{qq})c$$
$$= -cs\,a_{pp} - s^2 a_{qp} + c^2 a_{pq} + cs\,a_{qq}$$

Since A is symmetric,

$$\bar{a}_{pq} = (c^2 - s^2)a_{pq} + cs\,(a_{qq} - a_{pp}) = \tfrac{1}{2}\sin 2\theta\,(a_{qq} - a_{pp}) + \cos 2\theta\, a_{pq}$$

Hence; if

$$\theta = \begin{cases} \frac{1}{2}\tan^{-1}\left(\dfrac{2a_{pq}}{a_{pp} - a_{qq}}\right) & a_{pp} \neq a_{qq} \\ \pm\frac{\pi}{4} & a_{pp} = a_{qq} \end{cases} \qquad (1)$$

then the similarity tr. reduces $\bar{a}_{pq}$ to <u>zero</u>.

Unfortunately, as the calculation progresses, subsequent trs. will probably change the values of elements previously reduced to zero.

Thus the method becomes an iterative one in which we construct the sequence of matrices

$$A^{(k+1)} = R^T(p,q)\, A^{(k)}\, R(p,q) \qquad k = 0, 1, 2, \ldots$$

with $A^{(0)} = A$.

We take θ as:

$$-\frac{\pi}{4} \leqslant \theta \leqslant \frac{\pi}{4} \qquad (\text{if } a_{pp} \neq a_{qq}) \qquad \text{satisfying eqn(1)}$$

$$\theta = (\text{sign of } a_{pq}^{(k)})\,\frac{\pi}{4} \qquad (\text{if } a_{pp} = a_{qq})$$

For each value of k, we have to decide which off-diagonal element a_{pq} is to be reduced to zero.
It seems reasonable to choose the element of maximum modulus and this gives the standard Jacobi method.

These additional restrictions ensure that the iteration tends to a fixed diagonal matrix.

Because of the symmetry, only the elements above the diagonal need be considered.

$$R = R_1 R_2 R_3 \cdots R_n$$

Ex. $\quad A = \begin{pmatrix} 10 & 3 & 2 \\ 3 & 5 & 1 \\ 2 & 1 & 0 \end{pmatrix} \equiv A^{(0)}$

The element of $A^{(0)}$ with max. modulus and above the diagonal is in position $(p,q) = (1,2)$

$$R(1,2) = \begin{pmatrix} c\theta & -s\theta & 0 \\ s\theta & c\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{2a_{12}}{a_{11}-a_{22}}\right) = \frac{1}{2}\tan^{-1}\left(\frac{2\times 3}{10-5}\right) = 0.43803$$

$$\longrightarrow \begin{cases} s\theta = 0.42416 \\ c\theta = 0.90559 \end{cases}$$

$$A^{(1)} = R^T A R = \begin{pmatrix} c\theta & s\theta & 0 \\ -s\theta & c\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 10 & 3 & 2 \\ 3 & 5 & 1 \\ 2 & 1 & 0 \end{pmatrix}\begin{pmatrix} c\theta & -s\theta & 0 \\ s\theta & c\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 11.40518 & 0 & 0 \\ 0 & 3.59489 & 0.05727 \\ 2.23534 & 0.05727 & 0 \end{pmatrix}$$

Next we rotate in the (1,3) plane using

$$R(1,3) = \begin{pmatrix} c\theta & 0 & -s\theta \\ 0 & 1 & 0 \\ s\theta & 0 & c\theta \end{pmatrix}$$

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{2a_{13}}{a_{11}-a_{22}}\right) = \frac{1}{2}\tan^{-1}\left(\frac{2\times 2.23534}{11.40518-0}\right) = 0.18679$$

$$A^{(2)} = \begin{pmatrix} 11.82777 & 0.01064 & 0 \\ 0.01064 & 3.59489 & 0.05627 \\ 0 & 0.05627 & -0.42246 \end{pmatrix}$$

Note that the second iteration has destroyed the zero elements in position (1,2) and (2,1).

$$R(2,3) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & C\theta & -S\theta \\ 0 & S\theta & C\theta \end{pmatrix} \qquad \theta = \frac{1}{2}\tan^{-1}\left(\frac{2\times 0.05627}{3.59481-(-0.42246)}\right)$$

$$= 0.01401$$

giving;

$$A^{(3)} = \begin{pmatrix} 11.82777 & 0.01064 & 0 \\ 0.01064 & 3.59567 & 0 \\ 0 & 0 & -0.42325 \end{pmatrix}$$

Finally: $R(1,2)$, with $\theta = 0.00129$

$$A^{(4)} = \begin{pmatrix} 11.82780 & 0 & 0 \\ 0 & 3.59567 & 0 \\ 0 & 0 & -0.42325 \end{pmatrix} \qquad (\text{tollerance: } 0.0001)$$

$\lambda_1 = 11.8278$ $\qquad \lambda_2 = 3.5957$ $\qquad \lambda_3 = -0.4233$

Since $D = R^T A R$ $\qquad$ where $R = R_1 R_2 \cdots R_n$

$\rightarrow AR = RD$

$\longrightarrow$ Eigenvectors of A are the columns of R

$$R = R(1,2)\,R(1,3)\,R(2,3)\,R(1,2)$$

$$= \begin{pmatrix} 0.88929 & -0.42762 & -0.16222 \\ 0.41795 & 0.90386 & -0.09145 \\ 0.18573 & 0.01226 & 0.98251 \end{pmatrix}$$

The elements of $\bar{A}$ are:

$$\left.\begin{aligned} \bar{a}_{ip} &= a_{ip}c + a_{iq}s = \bar{a}_{pi} \\ \bar{a}_{iq} &= a_{ip}(-s) + a_{iq}c = \bar{a}_{qi} \end{aligned}\right\} i \neq p,q$$

$$\begin{aligned} \bar{a}_{pp} &= (c\,a_{pp} + s\,a_{qp})\,c + (c\,a_{pq} + s\,a_{qq})\,s \\ &= c^2 a_{pp} + 2cs\,a_{pq} + s^2 a_{qq} \end{aligned}$$

$$\begin{aligned} \bar{a}_{qq} &= (-s\,a_{pp} + c\,a_{qp})(-s) + (-s\,a_{pq} + c\,a_{qq})\,c \\ &= s^2 a_{pp} - 2cs\,a_{pq} + c^2 a_{qq} \end{aligned}$$

$$\bar{a}_{pq} = \bar{a}_{qp} = 0$$

$$\bar{a}_{ij} = a_{ij} \qquad i,j \neq p,q$$