Computational Methods

Review of Taylor series:

Familiar (and useful) examples of Taylor series are the following:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!} \quad (|x| < \infty)$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!} \quad (|x| < \infty)$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!} \quad (|x| < \infty)$$

$$\frac{1}{(1-x)} = 1 + x + x^2 + x^3 + \cdots = \sum_{k=0}^{\infty} x^k \quad (|x| < 1)$$

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots = \sum_{k=1}^{\infty} (-1)^{k-1} \frac{x^k}{k} \quad (-1 < x \leq 1)$$

All the above series are the examples of Taylor series of the given function about the point $c = 0$.

A Taylor series expanded about $c = 1$ is

$$\ln(x) = (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \cdots = \sum_{k=1}^{\infty} (-1)^{k-1} \frac{(x-1)^k}{k}$$

where $0 < x \leq 2$ $(\because -1 < x-1 \leq 1)$

Formal Taylor series for f about c:

$$f(x) \sim f(c) + (x-c)f'(c) + \frac{(x-c)^2}{2!} f''(c) + \frac{(x-c)^3}{3!} f'''(c) + \cdots$$

or $$f(x) \sim \sum_{k=0}^{\infty} \frac{(x-c)^k}{k!} f^{(k)}(c)$$

is called the "Taylor series of f at the point c."

In the special case $c=0$,

$$f(x) \sim f(0) + x f'(0) + \frac{x^2}{2!} f''(0) + \frac{x^3}{3!} f'''(0) + \cdots$$

or $$f(x) \sim \sum_{k=0}^{\infty} \frac{x^k}{k!} f^{(k)}(0)$$

it is also called a Maclaurin series.

Note: (1) Here, rather than using $=$, we have written $\sim$ to indicate that we are not allowed to assume that $f(x)$ equals the series on the right.

(2) Taylor series of f at the point c exists provided the successive derivative $f', f'', f''', \ldots$ exist at the point c.

Que Write the Taylor series of the function $f(x) = 3x^5 - 2x^4 + 15x^3 + 13x^2 - 12x - 5$ at the point $c = 2$?

Sol We know that the Taylor series of the function f at the point c is given by

$$f(x) \sim f(c) + (x-c) f'(c) + \frac{(x-c)^2}{2!} f''(c) + \frac{(x-c)^3}{3!} f'''(c) + \cdots$$

Given $c = 2$

and $f(x) = 3x^5 - 2x^4 + 15x^3 + 13x^2 - 12x - 5$ $\quad f(2) = 207$

$\therefore$ $f'(x) = 15x^4 - 8x^3 + 45x^2 + 26x - 12$ $\quad f'(2) = 396$

$f''(x) = 60x^3 - 24x^2 + 90x + 26$ $\quad f''(2) = 590$

$f'''(x) = 180x^2 - 48x + 90$ $\quad f'''(2) = 714$

$f^{(4)}(x) = 360x - 48$ $\quad f^{(4)}(2) = 672$

$f^{(5)}(x) = 360$ $\quad f^{(5)}(2) = 360$

$f^{(k)}(x) = 0 \quad \forall\, k \geq 6, k \in N$ $\quad f^{(k)}(2) = 0 \quad \forall\, k \geq 6, k \in N$

$\therefore$ The Taylor series of the given function at the point $c=2$ is

$$f(x) \sim 207 + (x-2).396 + \frac{(x-2)^2}{2!}.590 + \frac{(x-2)^3}{3!}.714 + \frac{(x-2)^4}{4!}.672 + \frac{(x-2)^5}{5!}.360$$

i.e., $$f(x) \sim 207 + 396(x-2) + 295(x-2)^2 + 119(x-2)^3 + 28(x-2)^4 + 3(x-2)^5$$ A

Note: In this example, $\sim$ may be replaced by $=$ but it is not possible in general.

Que Using the complete Horner's algorithm, find the Taylor expansion of the function $f(x) = 3x^5 - 2x^4 + 15x^3 + 13x^2 - 12x - 5$ at the point $c=2$?

Sol The work can be arranged as follows:

2)	3	−2	15	13	−12	−5
		6	8	46	118	212
	3	4	23	59	106	207
		6	20	86	290	
	3	10	43	145	396	
		6	32	150		
	3	16	75	295		
		6	44			
	3	22	119			
		6				
	3	28				

∴ By Horner's algorithm, the Taylor series of the given function at the point $c = 2$ is

$$f(x) = 3(x-2)^5 + 28(x-2)^4 + 119(x-2)^3 + 295(x-2)^2 + 396(x-2) + 207 \quad \underline{A}$$

Note: We can use Horner's algorithm for finding the Taylor expansion of a polynomial about any point. So, we can replace $\sim$ by $=$ in the Taylor expansion.

Taylor's theorem for $f(x)$:

If the function f possesses continuous derivatives of order $0, 1, 2, \cdots, (n+1)$ in a closed interval $I = [a, b]$, then for any c and x in I,

$$f(x) = \sum_{k=0}^{n} \frac{(x-c)^k}{k!} f^{(k)}(c) + E_{n+1} = f(c) + \frac{(x-c)}{1!} f'(c) + \frac{(x-c)^2}{2!} f''(c) + \cdots + \frac{(x-c)^n}{n!} f^{(n)}(c) + E_{n+1}$$

where E_{n+1} is called the remainder or error term and is given by $E_{n+1} = \frac{(x-c)^{n+1}}{(n+1)!} f^{(n+1)}(\xi)$ (Lagrange's form)

Here ξ is a point that lies between c and x and depends on both.

Note: The explicit assumption in this theorem is that $f(x), f'(x), f''(x), \cdots, f^{(n+1)}(x)$ are all continuous functions in the interval $I = [a, b]$ and the formula for E_{n+1} is valid when $f^{(n+1)}$ exists at each point of the open interval (a, b). Here the point ξ is in the open interval (c, x) or (x, c).

Other form of Taylor's theorem

Taylor's Theorem for f(x+h):

If the function f possesses continuous derivatives of order 0, 1, 2, ---, (n+1) in a closed interval $I = [a, b]$, then for any x in I,

$$f(x+h) = \sum_{k=0}^{n} \frac{h^k}{k!} f^{(k)}(x) + E_{n+1} = f(x) + \frac{h}{1!}f'(x) + \frac{h^2}{2!}f''(x) + \cdots + \frac{h^n}{n!}f^{(n)}(x) + E_{n+1}$$

where h is any value such that $x+h$ is in I and where

$$E_{n+1} = \frac{h^{n+1}}{(n+1)!} f^{(n+1)}(\xi)$$

for some ξ between x and $x+h$.

Note: (1) This form can be obtained from the previous form by replacing x by $x+h$ and replacing c by x.

(2) The requirement on ξ means $x < \xi < x+h$ if $h > 0$
or $x+h < \xi < x$ if $h < 0$

(3) The error term E_{n+1} depends on h in two ways: First h^{n+1} is explicitly present; second the point ξ generally depends on h. As $h \to 0$, $E_{n+1} \to 0$ with essentially the same rapidity with which h^{n+1} converges to zero. For large n, this is quite rapid. To express the qualitative fact, we write

$$E_{n+1} = O(h^{n+1})$$

as $h \to 0$. This is called big O notation.

Roughly speaking, $E_{n+1} = O(h^{n+1})$ means that the behavior of E_{n+1} is similar to the much simpler expression h^{n+1}.

Que Derive the Taylor series for e^x at $c=0$ and prove that it converges to e^x by using Taylor's Theorem.

Sol Let $f(x) = e^x$. Then

$$f^{(k)}(x) = e^x \text{ for } k \geq 0$$

$$\therefore f^{(k)}(c) = f^{(k)}(0) = e^0 = 1 \quad \forall\ k \geq 0$$

$\therefore$ By Taylor's Theorem for $f(x)$,

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

where $E_{n+1} = \frac{(x-c)^{n+1}}{(n+1)!} f^{(n+1)}(\xi)$ (ξ is a point that lies between c and x)

$\therefore$ We have $$e^x = \sum_{k=0}^{n} \frac{x^k}{k!} + \frac{x^{n+1}}{(n+1)!} e^{\xi} \qquad \text{——————(1)}$$

Now, let us consider all the values of x in some symmetric interval around the origin, for example, $-s \leq x \leq s$. Then $|x| \leq s$, $|\xi| \leq s$, and $e^{\xi} \leq e^s$. Hence

$$\lim_{n\to\infty} |E_{n+1}| = \lim_{n\to\infty} \left| \frac{x^{n+1}}{(n+1)!} e^{\xi} \right| \leq \lim_{n\to\infty} \frac{s^{n+1}}{(n+1)!} e^s = 0$$

$$\Rightarrow \lim_{n\to\infty} E_{n+1} = 0 \qquad \left(\because \lim_{n\to\infty} \frac{x^n}{n!} = 0\right)$$

$\therefore$ If we take the limit as $n \to \infty$ on both sides of eqn. (1), we get

$$e^x = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{x^k}{k!} = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

Note: The above example shows that in specific cases, a formal Taylor series actually represents the function.

Now, we examine another example to see how the formal series can fail to represent the function.

Example Derive the formal Taylor series for $f(x) = \ln(1+x)$ at $c=0$, and determine the range of positive x for which the series represents the function.

Sol Given $f(x) = \ln(1+x)$, $c=0$ $\quad f(0) = 0$

$$\therefore f'(x) = \frac{1}{(1+x)} \qquad f'(0) = 1$$

$$f''(x) = -\frac{1}{(1+x)^2} \qquad f''(0) = -1$$

$$f'''(x) = \frac{2}{(1+x)^3} \qquad f'''(0) = 2$$

$$f^{(4)}(x) = \frac{-2\cdot 3}{(1+x)^4} \qquad f^{(4)}(0) = -6$$

$$\vdots \qquad\qquad \vdots$$

$$f^{(k)}(x) = \frac{(-1)^{k-1}(k-1)!}{(1+x)^k} \qquad f^{(k)}(0) = (-1)^{k-1}(k-1)!$$

Hence by Taylor's Theorem, we get $f(x) = f(0) + x$

$$\ln(1+x) = \sum_{k=1}^{n}\left\{(-1)^{k-1}(k-1)!\right\}\frac{x^k}{k!} + E_{n+1} = \sum_{k=1}^{n}\frac{(-1)^{k-1}x^k}{k} + E_{n+1}$$

where $E_{n+1} = \dfrac{(-1)^n n!}{(1+\xi)^{n+1}}\dfrac{x^{n+1}}{(n+1)!} = \dfrac{(-1)^n}{(n+1)}\left(\dfrac{x}{1+\xi}\right)^{n+1}$

For the infinite series to represent $\ln(1+x)$, it is necessary and sufficient that the error term converge to 0 as $n \to \infty$.

Let us assume that $0 \le x \le 1$. Then $0 \le \xi \le x$ (because 0 is the point of expansion).

$$\therefore \quad 0 \le \frac{x}{(1+\xi)} \le 1$$

$$\therefore \lim_{n\to\infty} |E_{n+1}| = \lim_{n\to\infty} \left| \frac{(-1)^n}{(n+1)} \left(\frac{x}{1+\xi}\right)^{n+1} \right|$$

$$\leq \lim_{n\to\infty} \frac{1}{(n+1)} = 0$$

$$\Rightarrow \lim_{n\to\infty} E_{n+1} = 0$$

But if $x > 1$, the terms in the series do not approach 0, and the series does not converge.

Hence, the series represents $\ln(1+x)$ if $0 \leq x \leq 1$ but not if $x > 1$.

Note: The series also represents $\ln(1+x)$ for $-1 < x < 0$ but not if $x \leq -1$.

Que Evaluate $\sqrt{1+h}$ in powers of h. Then compute $\sqrt{1.00001}$ and $\sqrt{0.99999}$

Sol Let $f(x) = \sqrt{x}$. Then by Taylor's theorem

$$f(x+h) = f(x) + \frac{h}{1!}f'(x) + \frac{h^2}{2!}f''(x) + E_3 \quad \text{(by taking } n=2 \text{ for illustration)}$$

where $E_3 = \frac{h^3}{3!}f'''(\xi)$ for some ξ between x and $x+h$

By taking $x = 1$, we have

$$f(1+h) = \sqrt{1+h} = f(1) + \frac{h}{1!}f'(1) + \frac{h^2}{2!}f''(1) + \frac{h^3}{3!}f'''(\xi) \quad \text{——(1)}$$

where $1 < \xi < 1+h$ if $h > 0$

Now, $f(x) = \sqrt{x} = x^{\frac{1}{2}}$ $\quad f(1) = 1$

$\Rightarrow f'(x) = \frac{1}{2}x^{-1/2}$ $\quad f'(1) = \frac{1}{2}$

$f''(x) = -\frac{1}{4}x^{-3/2}$ $\quad f''(1) = -\frac{1}{4}$

$f'''(x) = \frac{3}{8}x^{-5/2}$ $\quad f'''(\xi) = \frac{3}{8}\xi^{-5/2}$

$\therefore$ By equ.(1),

$$\sqrt{1+h} = 1+\frac{1}{2}h-\frac{1}{8}h^2+\frac{1}{16}h^3\xi^{-5/2} \quad \text{where } 1<\xi<1+h \text{ if } h>0$$

Let $h = 0.00001 = 10^{-5}$. Then

$$\sqrt{1.00001} \approx 1+0.000005-0.125\times10^{-10} = 1.000004999987500$$

By substituting $-h$ for h in the series, we obtain

$$\sqrt{1-h} = 1-\frac{1}{2}h-\frac{1}{8}h^2-\frac{1}{16}h^3\xi^{-5/2}$$

Hence, by taking $h = 0.00001$, we have

$$\sqrt{0.99999} \approx 1-0.000005-0.125\times10^{-10} = 0.999994999987500$$

Now, $\frac{1}{16}h^3\xi^{-5/2} < \frac{1}{16}10^{-15}$ $(\because 1<\xi<1+h \Rightarrow \xi^{-5/2}<1)$

$$= 0.0625\times10^{-15}$$

$$= 0.000000000000000625$$

$\therefore$ Both numerical values are correct to all 15 decimal places shown.

Que Use five terms in Taylor series for $f(x) = \ln(1+x)$ about $x=0$ to approximate $\ln(1.1)$.

Sol We have, $f(x) = f(0)+xf'(0)+\frac{x^2}{2!}f''(0)+\frac{x^3}{3!}f'''(0)+\frac{x^4}{4!}f^{(4)}(0)+\cdots$ ——— (1)

Here $f(x) = \ln(1+x)$ $\quad f(0) = 0$

$f'(x) = \frac{1}{(1+x)}$ $\quad f'(0) = 1$

$f''(x) = \frac{-1}{(1+x)^2}$ $\quad f''(0) = -1$

$f'''(x) = +\frac{2}{(1+x)^3}$ $\quad f'''(0) = 2$

$f^{(4)}(x) = \frac{-6}{(1+x)^4}$ $\qquad f^{(4)}(0) = -6$

$f^{(5)}(x) = \frac{24}{(1+x)^5}$ $\qquad f^{(5)}(0) = 24$

∴ The Taylor series for $f(x) = \ln(1+x)$ upto five terms is

$$\ln(1+x) \approx x - \frac{x^2}{2} + \frac{x^3}{3!}(2) + \frac{x^4}{4!}(-6) + \frac{x^5}{5!}(24)$$

$$\Rightarrow \ln(1+x) \approx x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5}$$

By taking $x = 0.1$,

$$\ln(1.1) \approx 0.1 - \frac{(0.1)^2}{2} + \frac{(0.1)^3}{3} - \frac{(0.1)^4}{4} + \frac{(0.1)^5}{5}$$

$$= 0.1 - \frac{0.01}{2} + \frac{0.001}{3} - \frac{0.0001}{4} + \frac{0.00001}{5}$$

$$= 0.1 - 0.05 + 0.00033333 - 0.000025 + 0.000002$$

$$= 0.09531033333\ldots$$

This value is correct to six decimal places of accuracy.

Dr. Nishu Gupta

Rolle's Theorem: If

(i) f is a continuous function on the closed interval $[a,b]$

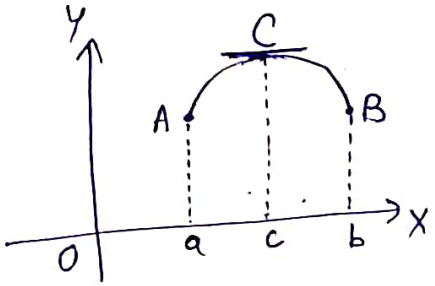
(ii) f' exists at each point of the open interval (a,b)

(iii) $f(a) = f(b)$,

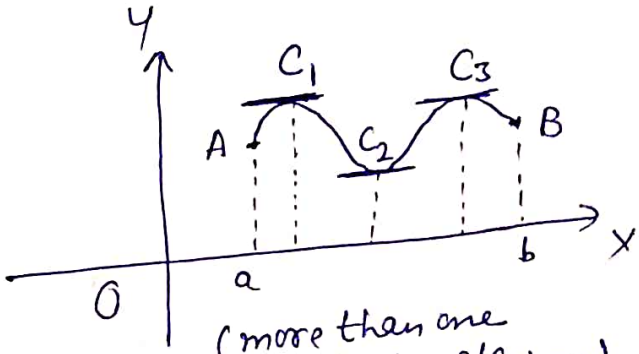
then there is at least one value $c \in (a,b)$ such that $f'(c) = 0$

Geometrical Interpretation:

Geometrically we can say that there is at least one point C (may be more) of the curve at which the tangent is parallel to the x-axis.



(exactly one point when $f'(c) = 0$)



(more than one point when $f'(c) = 0$)

Que: Verify Rolle's theorem for

(i) $\frac{\sin x}{e^x}$ in $[0, \pi]$ (ii) $(x-a)^m (x-b)^n$ where m, n are positive integers in $[a,b]$

Sol (i) Let $f(x) = \frac{\sin x}{e^x}$

(i) Clearly f is a continuous function in $[0,\pi]$ because $\sin x$ and e^x both are continuous in $[0,\pi]$ and $e^x \neq 0$ for any x.

(ii) Also, f is differentiable in $(0,\pi)$ as $\sin x$ and e^x both are differentiable in $(0,\pi)$ and $e^x \neq 0$ for any x.

(iii) $f(0) = \frac{\sin 0}{e^0} = \frac{0}{1} = 0$, $f(\pi) = \frac{\sin \pi}{e^\pi} = \frac{0}{e^\pi} = 0$

$\therefore f(0) = f(\pi)$

Hence the conditions of Rolle's theorem are satisfied.

Now, $f'(x) = \dfrac{e^x \cos x - e^x \sin x}{e^{2x}} = \dfrac{(\cos x - \sin x)e^x}{e^{2x}} = \dfrac{\cos x - \sin x}{e^x}$

$(\because e^x \neq 0)$

$\therefore$ $f'(x) = 0$ when $(\cos x - \sin x) = 0$

i.e., $\tan x = 1$

$\Rightarrow x = \frac{\pi}{4} \in (0, \pi)$

So, there exists a point $c = \frac{\pi}{4} \in (0, \pi)$ such that

$f'\left(\frac{\pi}{4}\right) = 0$

Hence Rolle's theorem is verified.

(ii) Let $f(x) = (x-a)^m (x-b)^n$ where m, n are positive integers in $[a, b]$.

Since every polynomial is continuous and differentiable for all values of x.

$\therefore$ f is continuous function in $[a, b]$ and differentiable in (a, b).

Also, $f(a) = f(b) = 0$

Hence the conditions of Rolle's theorem are satisfied.

Now, $f'(x) = m(x-a)^{m-1}(x-b)^n + (x-a)^m \cdot n(x-b)^{n-1}$

$= (x-a)^{m-1}(x-b)^{n-1}[m(x-b) + n(x-a)]$

$= (x-a)^{m-1}(x-b)^{n-1}[(m+n)x - (mb+na)]$

$\therefore$ $f'(x) = 0$ when $x = \dfrac{mb+na}{m+n}$

So, there exists a point $c = \dfrac{mb+na}{m+n} \in (a, b)$ such that $f'(c) = 0$

Hence Rolle's theorem is verified.

Dr. Nishu Gupta

Mean Value Theorem: If

(i) f is a continuous function on the closed interval $[a,b]$

(ii) f' exists at each point of the open interval (a,b),

then there is at least one value $c \in (a,b)$ such that

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

Que Verify Mean value theorem for

$$f(x) = x(x-1)(x-2) \text{ in } [0, \tfrac{1}{2}].$$

Sol Since every polynomial is continuous and differentiable for all values of x, therefore

$$f(x) = x^3 - 3x^2 + 2x.$$

$\Rightarrow$ f is continuous in $[0, \frac{1}{2}]$ and f' exists in (a,b).

Now, $f'(x) = 3x^2 - 6x + 2$

$$f(b) = f\left(\tfrac{1}{2}\right) = \left(\tfrac{1}{2}\right)^3 - 3\cdot\left(\tfrac{1}{2}\right)^2 + 2\cdot\left(\tfrac{1}{2}\right) = \frac{1}{8} - \frac{3}{4} + 1 = \frac{9}{8} - \frac{6}{8} = \frac{3}{8}$$

$$f(a) = f(0) = 0$$

$$\therefore \frac{f(b)-f(a)}{b-a} = \frac{\frac{3}{8}-0}{\frac{1}{2}-0} = \frac{3}{8} \times 2 = \frac{3}{4}$$

Now, $f'(c) = \dfrac{f(b)-f(a)}{b-a}$

When $3c^2 - 6c + 2 = \dfrac{3}{4}$

$\Rightarrow$ $12c^2 - 24c + 5 = 0$

$\Rightarrow$ $c = \dfrac{24 \pm \sqrt{(24)^2 - 4\times 12 \times 5}}{2\times 12} = \dfrac{24 \pm \sqrt{576-240}}{24}$

$\Rightarrow$ $c = \dfrac{24 \pm \sqrt{336}}{24} = \dfrac{24 \pm 18.33}{24} = 1.764, 0.236$

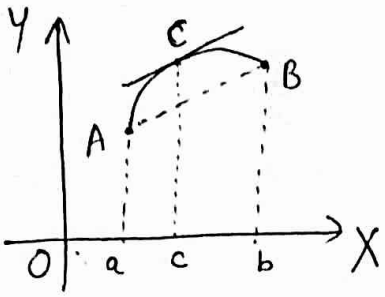
So, there exists $c = 0.236 \in (0, \frac{1}{2})$ such that

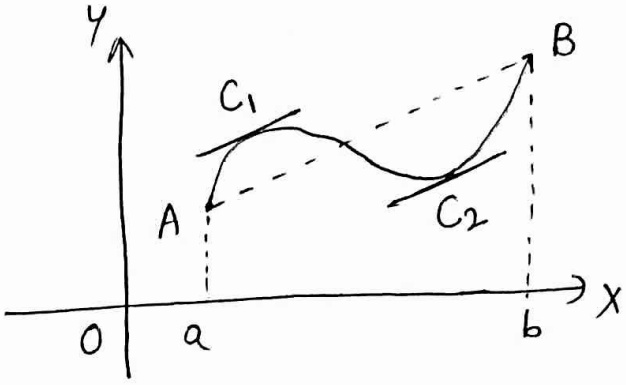
$$f'(c) = \frac{f(b)-f(a)}{b-a}$$

Hence Mean value-theorem is verified.

Geometrical Interpretation:

Geometrically, by mean value theorem we can say that there exists at least one point C (may be more) of the curve at which the tangent is parallel to the chord AB where $A = (a, f(a))$ and $B = (b, f(b))$.





Note: The special case $n = 0$ in Taylor's Theorem is known as the Mean-Value Theorem.

Approximation and Errors in numerical computations:

(1) **Approximate numbers:** There are two types of numbers exact and approximate. Exact numbers are 2, 4, 9, 13, 7/2, 6.45, etc. But there are numbers such as 4/3 (= 1.3333...), $\sqrt{2}$ (= 1.414213...) and π (= 3.141592...) which cannot be expressed by a finite number of digits. These may be approximated by numbers 1.3333, 1.4142 and 3.1416 respectively. Such numbers which represent the given numbers to a certain degree of accuracy are called approximate numbers.

(2) **Significant Digits of Precision:** Significant digits are digits beginning with the leftmost nonzero digit and ending with the rightmost correct digit, including final zeros that are exact.

Ex: 7845, 3.589, 0.4758 contains four significant digits

Ex: 0.00386, 0.000587, 0.00203 contains only three significant digits.

(3) **Accuracy and Precision:** Accurate to n decimal places means that you can trust n digits to the right of the decimal place. Accurate to n significant digits means that you can trust a total of n digits as being meaningful beginning with the leftmost nonzero digit.

(4) **Rounding off:** There are numbers with large number of digits eg., 22/7 = 3.14285714 3. In practice, it is desirable to limit such numbers to a manageable number of digits such as 3.14 or 3.143. This process of dropping unwanted digits is called rounding off.

Rule to round off a number

A number is rounded to position n by the following rule:

(i) Discard all digits to the right of the nth digit

(ii) If this discarded number is

(a) less than half a unit in the nth place, leave the nth digit unchanged;

(b) greater than half a unit in the nth place, increase the nth digit by unity;

(c) exactly half a unit in the nth place, increase the nth digit by unity if it is odd otherwise leave it unchanged.

For example (i) 84767 rounded to three significant figures = 84800

(ii) 3.567 rounded to three significant figures = 3.57

(iii) 8.73500 rounded to two decimal places = 8.74

(iv) 7.24500 rounded to two decimal places = 7.24

(v) 11.34576523 rounded to five decimal places = 11.34577

Errors: In any numerical computation we come across the following types of errors:

(1) **Inherent errors:** Errors which are already present in the statement of a problem before its solution are called inherent errors. Such errors arise either due to the given data being approximate or due to the limitation of mathematical tables, calculators or the digital computer. Inherent errors can be minimized by taking better data or by using high precision computer aids.

(2) Rounding errors: These errors arise from the process of rounding off the numbers during the computation. Such errors are unavoidable in most of the calculations due to the limitations of the computing aids. Rounding errors can, however be reduced:

(i) by changing the calculation procedure so as to avoid subtraction of nearly equal numbers or division by a small number;

(ii) by retaining at least one more significant figure at each step than that given in the data and rounding off at the last step.

(3) Truncation errors: These errors are caused by using approximate results or on replacing an infinite process by a finite one. For example, we consider the Taylor series expansion of $f(x)$ about $x = c$, $c \in [a,b]$. If we retain the first n terms, we get the approximation

$$f(x) \approx f(c) + (x-c)f'(c) + \frac{(x-c)^2}{2!}f''(c) + \cdots + \frac{(x-c)^{n-1}}{(n-1)!}f^{(n-1)}(c) \quad \text{---(1)}$$

and the truncation error (T.E.) is given by

$$T.E. = \frac{(x-c)^n}{n!}f^{(n)}(\xi), \text{ where } \xi \text{ lies between } c \text{ and } x$$

$$\Rightarrow |T.E.| \leq \frac{1}{n!} M_n \cdot \max_{[a,b]} |x-c|^n \quad \text{where } M_n = \max_{[a,b]} |f^{(n)}(x)| \quad \text{---(2)}$$

Assume that the value of M_n or its estimate is available. Then we can use it to determine an upper bound on the error.

Note: Suppose that we require $|T.E.| \leq \varepsilon$. Then we can determine (i) the number of terms (n) for a given x and ε by eq. (2).

(ii) $|x-c|$ for a given n and ε. This gives an interval about c in which the Taylor polynomial approximation given by eq (1) is valid to the prescribed accuracy.

(4) Absolute, Relative and Percentage errors: If X is the true value of a quantity and X^* is its approximate value, then

$$\text{Error} = \text{True value} - \text{Approximate value} = X - X^*$$

$$\text{Absolute Error} = |\text{True value} - \text{Approximate value}| = |X - X^*|$$

$$\text{Relative Error} = \frac{|X - X^*|}{|X|}$$

(True value is also known as exact value)

$$\text{Percentage error} = \frac{|X - X^*|}{|X|} \times 100$$

Note: (1) The relative and percentage errors are independent of the units used while absolute error is expressed in terms of these units.

(2) If a number is correct to n decimal places then the error

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

For Ex: If the number is 3.1416 correct to 4 decimal places, then the error $= \frac{1}{2} \times 10^{-4} = 0.00005$

(3) For practical reasons, the relative error is usually more meaningful than the absolute error. For example

if $x_1 = 1.333$, $x_1^* = 1.334$, and $x_2 = 0.001$, $x_2^* = 0.002$, then the absolute error of x_1^* as an approximation to x_1

$$= |x_1 - x_1^*| = |1.333 - 1.334| = 0.001 = 10^{-3}$$

and the absolute error of x_2^* as an approximation to x_2

$$= |x_2 - x_2^*| = |0.001 - 0.002| = 0.001 = 10^{-3}$$

Both are same.

But the relative error of $x_1^* = \frac{|x_1 - x_1^*|}{|x_1|} = \frac{1}{1.333} \times 10^{-3}$

$$= \frac{0.001}{1.333} = 0.750 \times 10^{-3}$$

and the relative error of x_2^*

$$= \frac{|x_2 - x_2^*|}{|x_2|} = \frac{0.001}{0.001} = 1$$

which indicates that x_1^* is a good approximation to x_1 but x_2^* is a poor approximation to x_2.

Note: If the approximate value of a number X having n decimal digits is X^* then

(i) Relative error due to rounding off to k digits

$$= \frac{|X - X^*|}{|X|} < \frac{1}{2} \cdot 10^{1-k}$$

(ii) Relative error due to truncation to k digits

$$= \frac{|X - X^*|}{|X|} < 10^{1-k}$$

Que Round off the numbers 865250 and 37.46235 to four significant figures and compute absolute error, relative error, percentage error in each case.

Sol (i) Number rounded off to four significant figures

$= 865200$

Absolute error = | True value − Approximate value |

$= |865250 - 865200| = 50$

$$\text{Relative Error} = \frac{|\text{True value} - \text{Appr. value}|}{|\text{True value}|} = \frac{50}{865250} = 6.71 \times 10^{-5}$$

Percentage Error = Relative Error $\times 100 = 6.71 \times 10^{-3}$

(ii) Number rounded off to four significant figures = 37.46

$\therefore$ Absolute Error = $|37.46235 - 37.46| = 0.00235$

$$\text{Relative Error} = \frac{0.00235}{37.46235} = 6.27 \times 10^{-5}$$

Percentage Error = $6.27 \times 10^{-5} \times 100 = 6.27 \times 10^{-3}$

Que Find the absolute error if the number $X = 0.00545828$ is

(i) truncated to three decimal digits

(ii) rounded off to three decimal digits

Sol Given $X = 0.00545828 = 0.545828 \times 10^{-2}$

(i) After truncated to three decimal digits, its approximate value $X^* = 0.545 \times 10^{-2}$

$\therefore$ Absolute error = $|X - X^*| = 0.000828 \times 10^{-2}$

$= 0.828 \times 10^{-5}$ A

(ii) Given $X = 0.0054582 8$

$= 0.545828 \times 10^{-2}$

After rounded off to three decimal places, its approximate value $X^* = 0.546 \times 10^{-2}$

$\therefore$ Absolute Error $= |X - X^*|$

$= |0.545828 \times 10^{-2} - 0.546000 \times 10^{-2}|$

$= 0.000172 \times 10^{-2}$

$= 0.172 \times 10^{-5}$ Ans

Que Find the relative error if the number $X = 0.004997$ is

(i) truncated to three decimal places

(ii) rounded off to three decimal digits

Sol We have $X = 0.004997 = 0.4997 \times 10^{-2}$

(i) After truncated to three decimal places, its approximate value $X^* = 0.499 \times 10^{-2}$

$\therefore$ Relative error $= \dfrac{|X - X^*|}{|X|}$

$= \dfrac{|0.4997 \times 10^{-7} - 0.499 \times 10^{-2}|}{|0.4997 \times 10^{-2}|}$

$= \dfrac{0.0007}{0.4997} = 0.00140 = 0.140 \times 10^{-2}$ Ans

(ii) After rounded off to three decimal places, its approximate value $X^* = 0.005 = 0.500 \times 10^{-2}$

$\therefore$ Relative error $= \dfrac{|X - X^*|}{|X|}$

$$= \frac{|0.4997 \times 10^{-2} - 0.5000 \times 10^{-2}|}{|0.4997 \times 10^{-2}|}$$

$$= \frac{0.0003}{0.4997} = 0.000600 = 0.600 \times 10^{-3} \quad \text{Ans}$$

Que Using Taylor series expansion of e^{-x} about $c = 0$. Determine

(i) maximum error for $x \in [-1, 1]$, when the first four terms are used in the approximation. Also find the maximum error when $x = 0.3$.

(ii) the least number of terms required in the approximation such that $|\text{error}| \leq 5 \times 10^{-4}$ for $x \in [-1, 1]$.

(iii) x, when the approximation obtained from the first four terms is accurate to 5×10^{-4}.

Sol Let $f(x) = e^{-x}$. Then $f(0) = 1$

$f^{(n)}(x) = (-1)^n e^{-x}$ $f^{(n)}(0) = (-1)^n$ and $f^{(n)}(\xi) = (-1)^n e^{-\xi}$

$\therefore$ Taylor series expansion of e^{-x} about $c = 0$ is given by

$$f(x) = f(0) + x f'(0) + \frac{x^2}{2!} f''(0) + \cdots + \frac{x^{n-1}}{(n-1)!} f^{n-1}(0) + E_n$$

where $E_n = \dfrac{x^n}{n!} f^{(n)}(\xi)$ where ξ lies between 0 and x —(1)

$\therefore$ When the first four terms are used, Taylor series expansion of e^{-x} is

$$e^{-x} \sim 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!}$$

and Error $= E_4 = \frac{x^4}{4!} f^{(4)}(\xi)$ where ξ lies between 0 and x

$\Rightarrow |\text{Error}| = \left| \frac{x^4}{4!} e^{-\xi} \right|$

$\leq \frac{e}{24}$ ($\because |x| \leq 1$ and $e^{-\xi} \leq e$ as $x \in [-1, 1]$ and ξ lies between 0 and x)

$\Rightarrow |\text{Error}| \leq 0.1133$

$\therefore$ Maximum error $= 0.1133$ A

For $x = 0.3$, $|\text{Error}| = \left| \frac{(0.3)^4}{4!} e^{-\xi} \right| \leq \frac{(0.3)^4}{24} e$ ($\because e^{-\xi} \leq e$ as above)

$= 0.00092$

$\therefore$ Maximum error $= 0.00092$ when $x = 0.3$ A

(ii) By eqn. (1), if n terms are required in the approximation, then Error $= \frac{x^n}{n!} f^{(n)}(\xi)$ where ξ lies between 0 and x

$\Rightarrow |\text{Error}| = \left| \frac{x^n}{n!} e^{-\xi} \right| \leq \frac{1}{n!} e$ ($\because |x| \leq 1$ and $e^{-\xi} \leq e$)

Given $|\text{error}| \leq 5 \times 10^{-4}$

$\therefore \frac{1}{n!} e \leq 5 \times 10^{-4} \Rightarrow n! \geq e \times \frac{10^4}{5}$ i.e, $n! \geq 2000e$

The inequality is satisfied for $n = 8$ A

(iii) Given $|E_4| \leq 5 \times 10^{-4}$ ($\because |E_4| \leq \frac{|x^4|}{24} e$)

$\therefore$ We choose x_1 such that $\frac{|x^4|}{24} e \leq 5 \times 10^{-4}$

or $|x|^4 \leq \frac{120}{e} \times 10^{-4} = 0.00441$

Hence $|x| \leq 0.2577$

$\Rightarrow x \in [-0.2577, 0.2577]$ A

Normalized floating-point representation or normalized scientific notation

In the decimal system, any real no. (other than 0) can be represented in normalized floating point form as

$$x = \pm 0.d_1 d_2 d_3 \cdots \times 10^n$$

where $d_1 \neq 0$ and n is an integer (positive, negative or zero). The numbers $d_1, d_2, \cdots$ are the decimal digits $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$.

Alternatively, any real no. (other than 0) can be represented in normalized floating point decimal form as

$$x = \pm r \times 10^n \quad \left(\frac{1}{10} \leq r < 1\right)$$

Here the number r is called the normalized mantissa and integer n the exponent.

In the binary system, if $x \neq 0$, it can be written as

$$x = \pm q \times 2^m \quad \left(\frac{1}{2} \leq q < 1\right), \; (-126 \leq m \leq 127)$$

where m is any integer

The mantissa q would be expressed as a sequence of zeros or ones in the form $(0.b_1 b_2 b_3 \cdots)_2$ where $b_1 \neq 0$. Hence $b_1 = 1$ and then necessarily $q \geq \frac{1}{2}$.

Note that every computer has only a finite word length and a finite total capacity, so only numbers with a finite number of digits can be represented.

- A number is alloted only one word of storage in the single precision mode (two or more words in double or extended precision).

In either case, the degree of precision is strictly limited.

- The most real numbers cannot be represented exactly in a computer.
- The real numbers that are representable in a computer are called its machine numbers.
- A number that has a terminating expansion in one base may have a nonterminating expansion in another.

For ex: $\frac{1}{10} = (0.1)_{10} = (0.063146314631463146314\ldots)_8$

$= (0.0001100110011001100110011\ldots)_2$

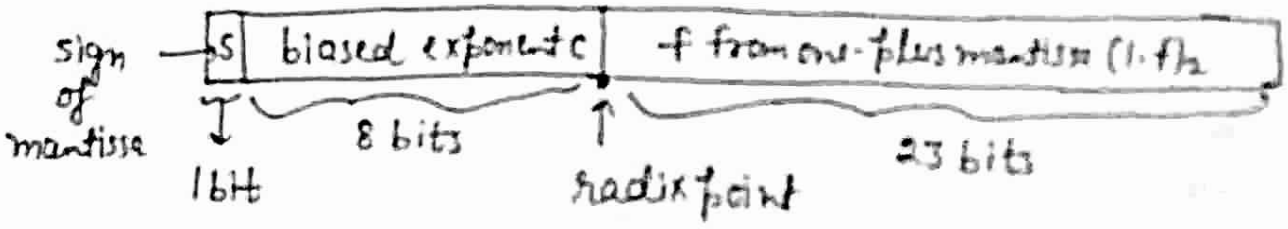
0.1 × 8 = 0.8
0.8 × 8 = 6.4
0.4 × 8 = 3.2
0.2 × 8 = 1.6
0.6 × 8 = 4.8
0.8 × 8 = 6.4

Single-Precision Floating-Point Form:

A machine number in standard single-precision floating-point form corresponds to

$$(-1)^s \times 2^{c-127} \times (1.f)_2 \quad , -126 \leq c-127 \leq 127$$

The leftmost bit is used for the sign of the mantissa, where s = 0 corresponds to + and s = 1 corresponds to −. The next eight bits are used to represent the number c in the exponent of 2^{c-127}, which is interpreted as an excess-127 code and the last 23 bits represent f from the fractional part of the mantissa in the 1-plus form: $(1.f)_2$. Each floating point single precision word is partitioned as



- **Note:** In the normalized representation of a nonzero floating-point number, the first bit in the mantissa is always 1 so that this bit does not have to be stored. This can be accomplished by shifting the binary point to a "1-plus" form $(1.f)_2$. The mantissa is the rightmost 23 bits and contains f with an understood binary point (radix point). So the mantissa (significand) actually corresponds to 24 binary digits since there is a hidden bit. (An important exception is the number ± 0)

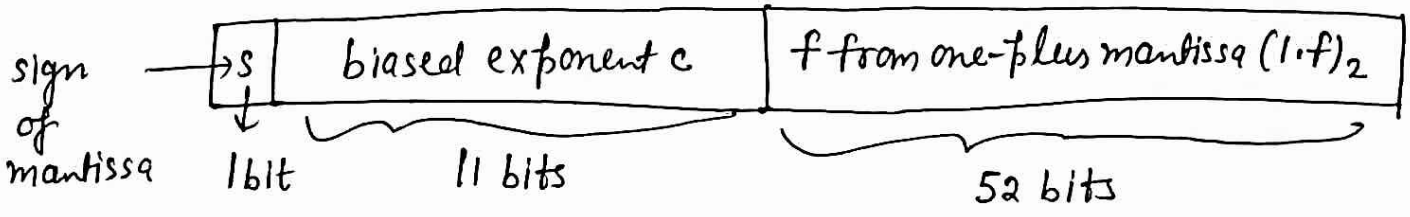
Double-Precision Floating-Point Form:

When more precision is needed, double precision can be used, in which case each double precision floating-point number is stored in two computer words in memory.

A machine number in standard double-precision floating-point form corresponds to

$$(-1)^s \times 2^{c-1023} \times (1.f)_2 \quad , \quad -1022 \leq c-1023 \leq 1023$$

which can be partitioned as



Que Determine the ~~single-precision~~ machine representation of the decimal number -52.234375 in both single precision and double precision.

Sol First we convert the integral part to binary, we have

$$(52.)_{10} = (64.)_8 = (110\ 100.)_2$$

8	52	
8	6	4 ↑
	0	6

Now, we convert the fractional part to binary

$(0.234375)_{10}$

$= (0.17)_8$

$= (0.001111)_2$

	0.234375
×	8
	1.875000
×	8
	7.000000

Note:

Binary:	000	001	010	011
Octal:	0	1	2	3
Binary:	**100**	**101**	**110**	**111**
Octal:	**4**	**5**	**6**	**7**

Now, $(52.234375)_{10}$

$= (110\ 100.001111)_2$

$= (1.101000\ 011\ 110)_2 \times 2^5$

is the corresponding one-plus form in base 2, and $(.101\ 000\ 011\ 110)_2$ is the stored mantissa.

Next the exponent is $(5)_{10}$ and since $c - 127 = 5$, we immediately see that $c = 132$ and $(132)_{10} = (204)_8$

8	132	
8	16	4 ↑
8	2	0
	0	2

$\therefore (10\ 000\ 100)_2$

is the stored exponent.

Thus, the single-precision machine representation of -52.234375 is

$$[1\ 10\ 000\ 100\ 101\ 000\ 011\ 110\ 000\ 000\ 000\ 00]_2$$

$= [1100\ 0010\ 0101\ 0000\ 1111\ 0000\ 0000\ 0000]_2$

$= [C250F000]_{16}$

Note

Binary :	0000	0001	0010	0011	0100	0101
Hexadecimal:	0	1	2	3	4	5
Binary :	6	7	8	9	A	B
Hexadecimal:	0110	0111	1000	1001	1010	1011
Binary :	1100	1101	1110	1111		
Hexadecimal:	C	D	E	F		

In double precision, for the exponent $(5)_{10}$, we let

$c - 1023 = 5 \Rightarrow c = 1028$

and we have $(1028)_{10} = (2004)_8$

$= (10\ 000\ 000\ 100)_2$

which is the store exponent.

8 | 1028		
8 | 128	4	↑
8 | 16	0	
8 | 2	0	
0	2	

Thus, the double-precision machine representation of -52.234375 is

$[1\ 10\ 000\ 000\ 100\ 101\ 000\ 011\ 110\ \underbrace{000\cdots00}_{40\ 0's}]_2$

$= [1100\ 0000\ 0100\ 1010\ 0001\ 1110\ \underbrace{0000\cdots0000}_{40\ 0's}]_2$

$= [C04A1E\underbrace{0000000000}_{10\ 0's}]_{16}$ A

Que Determine the decimal numbers that correspond to these machine words:

$[45DE4000]_{16}$ $[BA390000]_{16}$

Sol (i) $[45DE4000]_{16} = [0100\ 0101\ 1101\ 1110\ 0100\ 0000\ 0000\ 0000]_2$

0 → It shows sign is +ive; 10001011 → biased exponent; 11011110010000000000000 → f from one-plus mantissa $(1.f)_2$

Stored exponent $= (10001011)_2 = (1\times2^0+1\times2^1+1\times2^3+1\times2^7)_{10}$
$= (1+2+8+128)_{10} = (139)_{10}$

$\therefore c = 139 \Rightarrow 2^{c-127} = 2^{139-127} = 2^{12}$

Now, the mantissa is +ive and represents the number

$(1.101\ 111\ 001)_2 \times 2^{12} = (1101\ 1110\ 0100\ 00)_2$

$= 2^3+2^6+2^7+2^8+2^9+2^{11}+2^{12} = 2^3(1+2^3+2^4)+2^8(1+2+2^3+2^4)$

$= 8\times(1+8+16)+256(1+2+8+16)$

$= 8\times25+256\times27 = 200+6912 = (7112)_{10}$ ∆

or first we can write in octal form as $(213)_8 = 2\times8^2+8+3 = 128+8+3 = (139)_{10}$

(ii) $[BA390000]_{16}$

$= [1011\ 1010\ 0011\ 1001\ 0000\ 0000\ 0000\ 0000]_2$

1 → sign is -ve; 01110100 → biased exponent; 01110010000000000000000 → f from one-plus mantissa $(1.f)_2$

Stored exponent $= (01110100)_2 = (164)_8 = (1\times8^2+6\times8+4)_{10}$
$= (64+48+4)_{10} = (116)_{10}$

$\therefore c = 116 \Rightarrow 2^{c-127} = 2^{116-127} = 2^{-11}$

Now, the mantissa is negative and represents the number

$-(1.011\ 1001)_2 \times 2^{-11} = -(0.000\ 000\ 000\ 010\ 111\ 001)_2$

$= -(0.000271)_8 = -(2\times8^{-4}+7\times8^{-5}+8^{-6})_{10}$

$= -8^{-6}(1+56+128)$

$= -\frac{185}{8^6} = \frac{-185}{262144}$

$\approx -7.0571899\times10^{-4}$ ∆

Note:

(1) Machine epsilon: When we are using single precision, the binary machine floating-point number $\varepsilon = 2^{-23}$ is called the machine epsilon. It is the smallest positive machine number ε such that $1+\varepsilon \neq 1$. When we are using double precision, the machine epsilon is 2^{-52}.

Since $2^{-23} \approx 1.2 \times 10^{-7}$ and $2^{-52} \approx 2.2 \times 10^{-16}$

∴ Approximately 6 significant decimal digits of accuracy may be obtained in single precision while approximately 15 significant decimal digits of accuracy may be obtained in double precision.

(2) Single precision on a 64-bit computer is comparable to double precision on a 32-bit computer, whereas double precision on a 64-bit computer gives four times the precision available on a 32-bit computer.

Floating-point machine number:

Suppose that we are working with a five-place decimal machine and wish to add numbers.

For ex: $x = 0.37218 \times 10^4$ and $y = 0.71422 \times 10^{-1}$ are two normalized floating point machine numbers. Then we can find $x+y$ as

$$x = 0.3721800000 \times 10^4$$
$$y = 0.0000071422 \times 10^4$$
$$x+y = 0.3721871422 \times 10^4$$

(adjust the exponent of smaller number so that both exponents are same to add them)

The nearest machine number is $z = 0.37219 \times 10^4$

$\therefore$ Relative error (involved in this addition)

$$= \frac{|(x+y)-z|}{|x+y|} = \frac{0.0000028578 \times 10^4}{0.3721871422 \times 10^4} \approx 0.77 \times 10^{-5}$$

To facilitate the analysis of such errors, we introduce the notation $fl(x)$.

Notation $fl(x)$: The notation $fl(x)$ is used to denote the floating point machine number that corresponds to the real number x.

Note: $fl(x) = x(1+\delta)$, $|\delta| \leq 2^{-24}$

where δ is the relative error.

Computer Arithmetic: Let the symbol $\odot$ denote any one of the arithmetic operations $+, -, \times$ or $\div$. Suppose a 32-bit word length computer has been designed so that whenever two machine numbers x and y are to be combined automatically, the computer produces $fl(x \odot y)$ instead of $x \odot y$. We can imagine that $x \odot y$ is first correctly formed, then normalized, any finally rounded to become a machine number.

$\therefore$ $fl(x \odot y) = (x \odot y)(1+\delta)$; $|\delta| \leq 2^{-24}$

Hence $fl(x \pm y) = (x \pm y)(1+\delta)$ where δ is the relative error

$fl(xy) = xy(1+\delta)$

$fl\left(\frac{x}{y}\right) = \left(\frac{x}{y}\right)(1+\delta)$

where $-2^{-24} \leq \delta \leq 2^{24}$

Que If x, y and z are machine numbers in a 32-bit word-length computer, what upper bound can be given for the relative roundoff error in computing $z(x+y)$?

Sol In the computer, the calculation of $(x+y)$ would be done first and produces the machine number $fl(x+y)$, which differ from $x+y$ because of roundoff.

$$\therefore fl(x+y) = (x+y)(1+\delta_1) \quad (\text{for some } \delta_1 \text{ such that } |\delta_1| \leq 2^{-24})$$

When z multiplies the machine number $fl(x+y)$, the result is the machine number $fl[z\,fl(x+y)]$ because z is also a machine number.

$$\therefore fl[z\,fl(x+y)] = z\,fl(x+y)(1+\delta_2) \quad (\text{for some } \delta_2 \text{ such that } |\delta_2| \leq 2^{-24})$$

Hence
$$\begin{aligned} fl[z\,fl(x+y)] &= z(x+y)(1+\delta_1)(1+\delta_2) \\ &= z(x+y)(1+\delta_1+\delta_2+\delta_1\delta_2) \\ &\approx z(x+y)(1+\delta_1+\delta_2) \\ &\qquad (\because |\delta_1\delta_2| \leq 2^{-48} \text{ and so we ignore it}) \\ &= z(x+y)(1+\delta) \quad (\text{where } \delta = \delta_1+\delta_2) \end{aligned}$$

Now, relative roundoff error in computing $z(x+y)$

$$= |\delta|$$
$$= |\delta_1+\delta_2| \leq |\delta_1|+|\delta_2| \leq 2^{-24}+2^{-24} = 2^{-23}$$

Hence $|\delta| \leq 2^{-23}$ A

Que Show by an example that in computer arithmetic $a+(b+c)$ may differ from $(a+b)+c$.

Sol Let $a = 0.345$, $b = 0.245 \times 10^{-3}$ and $c = 0.432 \times 10^{-3}$ and we are using 3-digit rounding.

Here $a = 0.345 \times 10^{0}$

$b = 0.000245 \times 10^{0}$

$c = 0.000432 \times 10^{0}$

Now, $b+c = 0.000677 \times 10^{0} = 0.677 \times 10^{-3}$ ($\because$ write numbers in normalized floating form)

$\therefore a+(b+c) = (0.345+0.000677) \times 10^{0}$

$= 0.345677 \times 10^{0} \approx \boxed{0.346 \times 10^{0}}$

Also, $a+b = 0.345245 \times 10^{0} = 0.345 \times 10^{0}$

$\therefore (a+b)+c = (0.345+0.000432) \times 10^{0}$

$= 0.345432 \times 10^{0} \approx \boxed{0.345 \times 10^{0}}$

Hence, $a+(b+c) \neq (a+b)+c$

Que If x and y are real numbers within the range of a 32-bit word-length computer and if xy is also within the range, what relative error can there be in the machine computation of xy?

Sol Machine produces $fl[fl(x)fl(y)]$.

Now, $fl(x) = x(1+\delta_1)$ for some δ_1 such that $|\delta_1| \leq 2^{-24}$

$fl(y) = y(1+\delta_2)$ " " δ_2 " " $|\delta_2| \leq 2^{-24}$

$\therefore fl[fl(x)fl(y)] = fl(x)fl(y).(1+\delta_3)$ for some δ_3 such that $|\delta_3| \leq 2^{-24}$

$= xy(1+\delta_1)(1+\delta_2)(1+\delta_3)$

$\approx xy(1+\delta_1+\delta_2+\delta_3)$ (neglecting other terms as they are very small and will not affect the calculation)

$= xy(1+\delta)$; $\delta = \delta_1+\delta_2+\delta_3$

$\therefore$ relative error $= |\delta| \leq |\delta_1|+|\delta_2|+|\delta_3| = 3.2^{-24}$ A

Loss of Significance: Loss of significance occurs in numerical calculations when too many significant digits cancel.

Significant digits Suppose that x is a real number expressed in normalized scientific notation in the decimal system

$$x = \pm r \times 10^n \quad \left(\frac{1}{10} \leq r < 1\right)$$

For Ex: $x = 0.3721498 \times 10^5$

The digits 3, 7, 2, 1, 4, 9, 8 used to express r do not have the same significance because they represent different powers of 10.

Here 3 is the most significant digit
and 8 is the least significant digit

Note that the significance of the digits diminishes from left to right.

Que If $x = 0.3721448693$
$y = 0.3720214371$, then what is the relative error in the computation of $x-y$ in a computer that has five decimal digits of accuracy?

Sol Exact value of $x-y = 0.3721448693 - 0.3720214371$
$= 0.0001234322$

Approximate value of $x-y$ (with five decimal digits of accuracy)
$= 0.37214 - 0.37202 = 0.00012$

$$\therefore \text{Relative error involved} = \left|\frac{\text{Exact value} - \text{Approximate value}}{\text{Exact value}}\right|$$

$$= \left|\frac{0.0001234322 - 0.00012}{0.0001234322}\right| = \frac{0.0000034322}{0.0001234322}$$

$\approx 3 \times 10^{-2}$ (which is quite large as by the coarsest estimates it cannot exceed $\frac{1}{2} \times 10^{1-5} = \frac{1}{2} \times 10^{-4}$)

Loss of precision theorem: Let x and y be normalized floating-point machine numbers, when $x > y > 0$. If

$2^{-p} \leq 1 - \frac{y}{x} \leq 2^{-q}$ for some positive integers p and q, then at most p and at least q significant binary bits are lost in the subtraction $x - y$.

Que In the subtraction $37.593621 - 37.584216$, how many bits of significance will be lost?

Sol Let $x = 37.593621$

and $y = 37.584216$

Then $1 - \frac{y}{x} = 0.0002501754$

This lies between $2^{-12} = 0.000244$ and $2^{-11} = 0.000488$.

Hence, at least 11 but not more than 12 bits are lost.

Remark: To avoid loss of significance in subtraction, one may be able to reformulate the expression using rationalizing, series expansions, or mathematical identities.

Ex: Consider the function

$$f(x) = \sqrt{x^2+1} - 1$$

whose value may be required for x near zero.

Since $\sqrt{x^2+1} \approx 1$ when $x \approx 0$, we see that there is a potential loss of significance in the subtraction.

$\therefore$ We can rationalized the numerator to avoid loss of significance as

$$f(x) = (\sqrt{x^2+1} - 1) \times \frac{(\sqrt{x^2+1}+1)}{(\sqrt{x^2+1}+1)}$$

$$= \frac{(x^2+1)-1}{\sqrt{x^2+1}+1} = \frac{x^2}{\sqrt{x^2+1}+1} \quad (\because \text{It removes subtraction})$$

Que How can accurate values of the function

$$f(x) = e^x - e^{-2x} \qquad \text{——— (1)}$$

be computed in the vicinity of $x = 0$?

Sol Since e^x and e^{-2x} are both equal to 1 when $x = 0$, therefore there is a loss of significance in the subtraction when x is closed to zero.

One cure of this problem is to use the Taylor series as

$$f(x) = \left(1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots\right) - \left(1 - 2x + \frac{4x^2}{2!} - \frac{8x^3}{3!} + \cdots\right)$$

$$= 3x - \frac{3}{2}x^2 + \frac{3}{2}x^3 - \cdots \qquad \text{——— (2)}$$

Extra: Find the range in which series (2) should be used and the range in which formula (1) can be used.

Sol Using the Theorem on Loss of Precision, we see that the loss of bits in the subtraction of formula (1) can be limited to at most 1 bit by restricting x so that

$$2^{-1} \le 1 - \frac{e^{-2x}}{e^x} \Rightarrow \frac{1}{2} \le 1 - \frac{1}{e^{3x}} \Rightarrow \frac{1}{e^{3x}} \le \frac{1}{2} \quad (\text{when } x > 0)$$

$$\Rightarrow e^{3x} \ge 2$$

$$\Rightarrow 3x \ge \ln 2$$

$$\Rightarrow x \ge \frac{1}{3}\ln 2 = 0.23105$$

Similarly when $x < 0$, then for $x \le -0.23105$ at most 1 bit is lost.

Hence the series (2) should be used for $|x| \leq 0.23105$ and for $|x| > 0.23105$, formula (1) can be used.

Range Reduction:

Another cause of loss of significant figures is the evaluation of various library functions with large arguments. For ex:-

A basic property of the function $\sin x$ is its periodicity;

$$\sin x = \sin(x + 2n\pi)$$ for all real values of x and for all integer values of n.

Because of this relationship, we need to know only the values of $\sin x$ in some fixed interval of length 2π to compute $\sin x$ for arbitrary x. This property can be used in the computer evaluation of $\sin x$ and is called range reduction.

Que For $\sin x$, how many binary bits of significance are lost in range reduction to the interval $[0, 2\pi)$?

Sol Given an argument $x > 2\pi$, we find an integer n that satisfies $0 \leq x - 2n\pi < 2\pi$

Then in evaluating elementary trigonometric functions, we use $f(x) = f(x - 2n\pi)$

In the subtraction $x - 2n\pi$, there is a loss of significance.

By the Loss of Precision theorem, at least q bits are lost if

$$1 - \frac{2n\pi}{x} \leq 2^{-q}$$

Since $$1 - \frac{2n\pi}{x} = \frac{x - 2n\pi}{x} < \frac{2\pi}{x}$$

we conclude that at least q bits are lost if $\frac{2\pi}{x} \leq 2^{-q}$

or $2^q \leq \frac{x}{2\pi}$ A

Location of roots of an equation: Algebraic and Transcendental Equations

An equation $f(x)=0$ is called an algebraic equation of degree n, if $f(x)$ is a polynomial of degree n.

If $f(x)$ contains some other functions such as trigonometric, logarithmic, exponential, etc. then $f(x)$ is called a transcendental equation.

(I.V.P.) **Intermediate Value Property:** If $f(x)$ is continuous in $[a,b]$ and $f(a)\cdot f(b)<0$ then $f(x)=0$ has one real root in (a,b).

Convergence: Let $x_1, x_2, x_3, \ldots, x_{n+1}$ be successive approximations of root r of an equation. If there exists a constant C such that

$$|r-x_{n+1}| \leq C|r-x_n|^m \quad (n \geq 1) \text{ or } |e_{n+1}| \leq C|e_n|^m$$

then convergence is said to be of order m. (where e_{n+1} and e_n are errors at $(n+1)^{th}$ and n^{th} step)

Bisection Method or Bolzano Method or Halving Method:

This method is based on the repeated application of I.V.P. Suppose $f(x)$ is a continuous function of x and we are to find real root of $f(x)=0$

Let a and b be real numbers such that $f(a)\cdot f(b)<0$

then I^{st} approximation is $x_1 = \frac{1}{2}(a+b)$

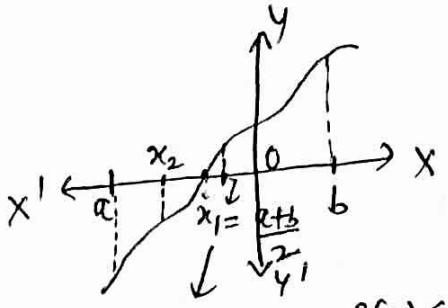
If $f(x_1)=0$, then x_1 is a root.

If $f(x_1) \neq 0$ then

either $f(a)\cdot f(x_1)<0$ in which case II^{nd} approximation $x_2 = \frac{a+x_1}{2}$

or $f(x_1)\cdot f(b)<0$ in which case II^{nd} approximation $x_2 = \frac{x_1+b}{2}$

Now, replace a or b by x_1 as the case be, then next approximation will be $x_3 = \frac{x_1 + x_2}{2}$ and so on.



$f(x_1) > 0$, $f(a) < 0$

$\therefore x_2 = \frac{a + x_1}{2}$

Convergence of Bisection method:

Suppose f is a continuous function in $[a, b]$ and $f(a).f(b) < 0$. Then there is a root r in $[a, b]$. If we use

$x_1 = \frac{a+b}{2}$ as Ist approximation, then

$$|r - x_1| \leq \frac{b-a}{2}$$

Now, we choose the next approximation $x_2 = \frac{a+x_1}{2}$ or $x_2 = \frac{b+x_1}{2}$ as the case may be, then

$$|r - x_2| \leq \frac{b-a}{2^2} \quad (\because \text{length of the interval at each step is } \tfrac{1}{2} \text{ the length of interval in previous step in which root lies})$$

$$\vdots$$

$$\therefore |r - x_n| \leq \frac{b-a}{2^n} \quad \text{——— (1)}$$

$\therefore$ At the end of n steps, when we obtain x_n, the root will lie in an interval of length $\frac{b-a}{2^n}$.

[Note: If we use $x_0 = \frac{a+b}{2}$ as initial estimate of r, and from next onward $x_1, x_2, \ldots x_n$, then $|r - x_n| \leq \frac{b-a}{2^{n+1}}$. Both are correct.]

Now, as the length of interval at each step is $\frac{1}{2}$ the length of the interval in the previous step in which root r lies, so

$$|r - x_{n+1}| = \frac{1}{2}|r - x_n| \quad (\text{this shows that error at } (n+1)^{th} \text{ step is } \tfrac{1}{2} \text{ of the error at } n^{th} \text{ step})$$

∴ Convergence is linear.

Hence the process is slow but must converge.

Number of iterations required to reach accuracy ε

By equation (1), the no. of iterations n required to reach accuracy ε, we must have

$$\frac{b-a}{2^n} \leq \varepsilon$$

$$\text{or } \log(b-a) - n\log 2 \leq \log \varepsilon$$

$$\text{or } n \geq \frac{\log(b-a) - \log\varepsilon}{\log 2} \quad \text{——— (2)}$$

∴ Smallest natural no. n satisfying this inequality gives the no. of iterations required to reach accuracy ε.

Bisection Method theorem If the bisection algorithm is applied to a continuous function f on an interval $[a,b]$, where $f(a)\cdot f(b) < 0$, then after n steps, an approximate root will have been computed with error at most $(b-a)/2^n$. (where we are considering first step as $x_1 = \frac{a+b}{2}$)

Que How many steps of the bisection algorithm are needed to compute a root of f to full machine single precision on a 32-bit word-length computer if $a = 16$ and $b = 17$?

Sol By equation (2), the no. of steps n required is given by

$$n \geq \frac{\log(b-a) - \log\varepsilon}{\log 2}$$

Here $a = 16$ and $b = 17$

$$\therefore\ n \geq \frac{\log 1 - \log \varepsilon}{\log 2} \Rightarrow n \geq -\frac{\log \varepsilon}{\log 2} \quad \text{——(1)}$$

Now, the root is between the two binary numbers $a = (10000.0)_2$ and $b = (10001.0)_2$. Thus, we already know five of the binary digits in the answer. Since we can use 23 bits for mantissa f in $(1.f)_2$ form and 4 digits for f are given that leaves 19 bits to determine. We want the last one to be correct, so we want the error to be less than 2^{-19} or 2^{-20} (being conservative). $\therefore \varepsilon = 2^{-20}$

$\therefore$ By eq. (1) above $n \geq -\dfrac{\log 2^{-20}}{\log 2} \Rightarrow \boxed{n \geq 20}$ Δ

Que How many steps of the bisection method are needed to find a root of the equation $xe^x = 1$ correct to three decimal places in the interval $[0,1]$.

Sol Given $a = 0$, $b = 1$, $\varepsilon = 0.0005$

$\therefore$ No. of steps n are given by

$$n \geq \frac{\log(b-a) - \log \varepsilon}{\log 2}$$

$$\Rightarrow n \geq \frac{\log 1 - \log(0.0005)}{\log 2}$$

$$\Rightarrow n \geq 10.97$$

$\therefore$ Minimum steps are required $n = 11$ Δ

Note: It will be verified by solving the problem in next question.

Que Find a real root of the equation $xe^x = 1$ correct to three decimal places using bisection method.

Sol Let $f(x) = xe^x - 1 = 0$, Then

clearly f is continuous as

$f(0) = -1$, $f(1) = e - 1 = 1.71828$

i.e, $f(0).f(1) < 0 \Rightarrow$ root lies between 0 and 1.

$\therefore x_1 = \frac{0+1}{2} = 0.5$

Approximate root	f(x)	Root lies between	Next Approximation
$x_1 = 0.5$	-ive	0.5 and 1	$\frac{0.5+1}{2} = 0.75$
$x_2 = 0.75$	+ive	0.5 and 0.75	$\frac{0.5+0.75}{2} = 0.625$
$x_3 = 0.625$	+ive	0.5 and 0.625	$\frac{0.5+0.625}{2} = 0.5625$
$x_4 = 0.5625$	-ive	0.5625 and 0.625	$\frac{0.5625+0.625}{2} = 0.59375$
$x_5 = 0.59375$	+ive	0.5625 and 0.59375	$\frac{0.5625+0.59375}{2} = 0.57812$
$x_6 = 0.57812$	+ive	0.5625 and 0.57812	$\frac{0.5625+0.57812}{2} = 0.57031$
$x_7 = 0.57031$	+ive	0.5625 and 0.57031	$\frac{0.5625+0.57031}{2} = 0.56640$
$x_8 = 0.56640$	-ive	0.56640 and 0.57031	$\frac{0.56640+0.57031}{2} = 0.56836$
$x_9 = 0.56836$	+ive	0.56640 and 0.56836	$\frac{0.56640+0.56836}{2} = 0.56738$
$x_{10} = 0.56738$	+ive	0.56640 and 0.56738	$\frac{0.56640+0.56738}{2} = 0.56689$
$x_{11} = 0.56689$			

Since $x_{10} \approx x_{11}$ (correct to three decimal places)

$\therefore$ root $= 0.567$ (correct to 3D) ∎

Newton method or Newton-Raphson method or Method of Tangents:

Let x_0 be an approximation to the root of $f(x)=0$. We find the equation of tangent at (x_0, y_0) to the graph of curve $y = f(x)$ where $y_0 = f(x_0)$.

Let this tangent meets x-axis at x_1 then x_1 will be next approximation and we find (x_1, y_1) on the graph and draw tangent at (x_1, y_1) to the curve $y = f(x)$. Its intersection with x-axis will be x_2.

Proceeding in this way, when approximation x_n is found then intersection of tangent at (x_n, y_n) to $y = f(x)$ with x-axis will give next approximation x_{n+1}.

Now, equation of tangent at (x_n, y_n) to $y = f(x)$ is

$$y - y_n = f'(x_n)(x - x_n)$$

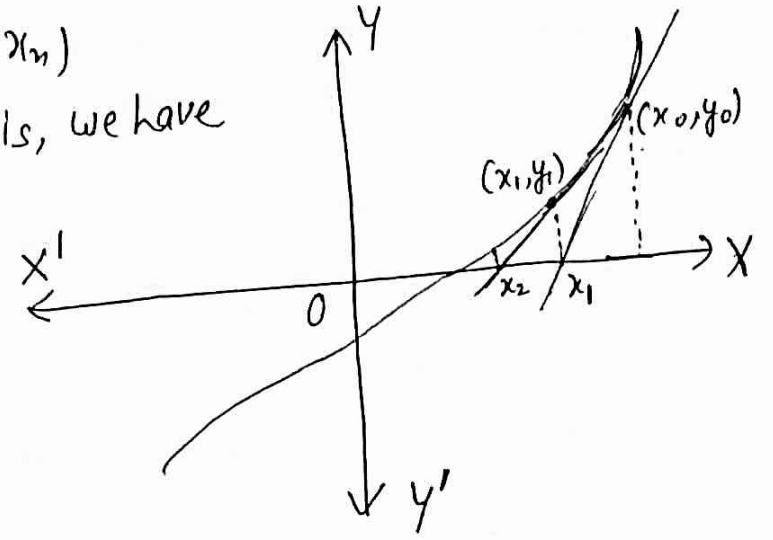
For its intersection with x-axis, we have

$y = 0$ and $x = x_{n+1}$

$$\therefore \; -y_n = f'(x_n)(x_{n+1} - x_n)$$

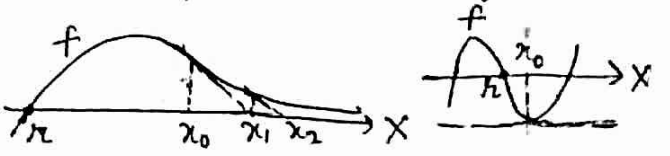
$$\Rightarrow \; x_{n+1} = x_n - \frac{y_n}{f'(x_n)}$$

or $$\boxed{x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}}$$



It is Newton's iterative formula to obtain the approximations.

Note: The method may fail if the initial approximation x_0 is far away from the root,
or the tangent at (x_0, y_0) does not intersect the x-axis.



Convergence of Newton Raphson method:

Let r be exact root of $f(x)$. Let x_n and x_{n+1} be its two successive approximations. Then by Newton Raphson iterative formula $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$

Substituting $x_n = r + e_n$ and $x_{n+1} = r + e_{n+1}$

$$r + e_{n+1} = r + e_n - \frac{f(r+e_n)}{f'(r+e_n)}$$

$$\therefore e_{n+1} = \frac{e_n f'(r+e_n) - f(r+e_n)}{f'(r+e_n)}$$

$$= e_n\left[f'(r) + \frac{e_n}{1!}f''(r) + \frac{e_n^2}{2!}f'''(r) + \cdots\right] - \left[f(r) + e_n f'(r) + \frac{e_n^2}{2!}f''(r) + \frac{e_n^3}{3!}f'''(r) + \cdots\right]$$

$$f'(r) + e_n f''(r) + \frac{e_n^2}{2!}f'''(r) + \cdots$$

(by Taylor series expansion)

But $f(r) = 0$ ($\because r$ is a root of $f(x) = 0$)

$$\therefore e_{n+1} = \frac{\frac{e_n^2}{2}f''(r) + e_n^3 f'''(r)\cdot\left(\frac{1}{2} - \frac{1}{3!}\right) + \cdots}{f'(r) + e_n f''(r) + \frac{e_n^2}{2!}f'''(r) + \cdots}$$

$$= \frac{1}{f'(r)}\left\{\frac{e_n^2}{2}f''(r) + \frac{e_n^3}{3}f'''(r) + \cdots\right\}\left\{1 + \left(e_n\frac{f''(r)}{f'(r)} + \frac{e_n^2 f'''(r)}{2!f'(r)} + \cdots\right)\right\}^{-1}$$

$$= \frac{1}{f'(r)}\left\{\frac{e_n^2}{2}f''(r) + \frac{e_n^3}{3}f'''(r) + \cdots\right\}\left\{1 - e_n\frac{f''(r)}{f'(r)} - \frac{e_n^2 f'''(r)}{2f'(r)} - \cdots\right\}$$

$$\therefore e_{n+1} = \frac{1}{2}e_n^2\frac{f''(r)}{f'(r)} + \cdots\cdots$$

$$\Rightarrow e_{n+1} = \frac{1}{2}e_n^2\frac{f''(r)}{f'(r)}$$ (If remaining terms are neglected)

$$\Rightarrow |e_{n+1}| \le C\cdot|e_n|^2 \text{ where } C = \frac{1}{2}\left|\frac{f''(r)}{f'(r)}\right|$$

Hence the convergence is of order 2 i.e, quadratic.

Que Using Newton-Raphson method evaluate $\sqrt[3]{41}$ correct to four places of decimals.

Sol Let $f(x) = x^3 - 41 = 0$

$\therefore f'(x) = 3x^2$

$\therefore$ Newton-Raphson iterative formula $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$

becomes, $x_{n+1} = x_n - \frac{x_n^3 - 41}{3x_n^2} = \frac{2x_n^3 + 41}{3x_n^2} = \frac{1}{3}\left(2x_n + \frac{41}{x_n^2}\right)$

Since $3^3 = 27$, $4^3 = 64$

$\therefore \sqrt[3]{27} = 3$, $\sqrt[3]{64} = 4$

Take $x_0 = 3.4$

n	x_n	$x_{n+1} = \frac{1}{3}\left(2x_n + \frac{41}{x_n^2}\right)$
0	3.4	3.4489
1	3.4489	3.44822
2	3.44822	3.44822

$\therefore \sqrt[3]{41} = 3.4482$ (correct to four places of decimals)

A

Que Use Newton-Raphson method to solve the equation

$$3x - \cos x - 1 = 0$$

Sol Let $f(x) = 3x - \cos x - 1 = 0$ $\therefore f(0) = -2$, $f(1) = 1.4597$

$\therefore f'(x) = 3 + \sin x$ $\therefore$ we take $x_0 = 0.6$

$\therefore$ Newton-Raphson iterative formula is

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

$$\Rightarrow \quad x_{n+1} = x_n - \frac{(3x_n - \cos x_n - 1)}{3 + \sin x_n} = \frac{x_n \sin x_n + \cos x_n + 1}{3 + \sin x_n}$$

n	x_n	$x_n \sin x_n + \cos x_n + 1$	$3 + \sin x_n$	x_{n+1}
0	0.6	2.1641	3.5646	0.6071
1	0.6071	2.1676	3.5705	0.607099

$\therefore$ Root to four decimal places = 0.6071 ∆

Multiplicity of the zero r of $f(x)=0$ is the least m such that

$$f^{(k)}(r) = 0 \text{ for } 0 \leq k < m \text{ but } f^{(m)}(r) \neq 0$$

For ex: $f(x) = x^2 - 2x + 1 = 0$ has a root at 1 of multiplicity 2
($\because f(x) = (x-1)^2$)

If we already know in advance that there is a zero of $f(x)=0$ with multiplicity m, then we can find it by modifying the Newton's method as,

$$x_{n+1} = x_n - \frac{m f(x_n)}{f'(x_n)}$$

and problem can be solved in a similar manner.

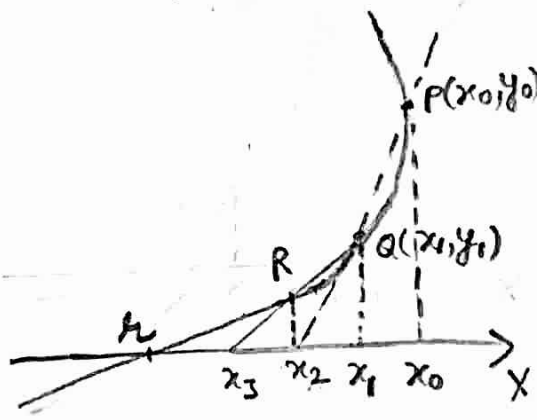
——x——x——

Secant Method: Let x_0, x_1 be two approximations of root of $y = f(x) = 0$. Then $P(x_0, y_0)$ and $Q(x_1, y_1)$ are two points on the curve $y = f(x)$ where $y_0 = f(x_0)$, $y_1 = f(x_1)$. Join PQ. We approximate the curve by secant (chord) PQ and take ~~Secant~~ the point of intersection of PQ with x-axis as the next approximation x_2 of the root. Then we take secant joining $Q(x_1, y_1)$ and $R(x_2, y_2)$ and repeat the same process to get the next approximation x_3.

Proceeding in this way, curve is approximated by secant joining (x_{n-1}, y_{n-1}) and (x_n, y_n) and its point of intersection with x-axis as the approximation x_{n+1} of the root.

Equation of secant joining (x_{n-1}, y_{n-1}) and (x_n, y_n) is

$$y - y_n = \frac{y_n - y_{n-1}}{x_n - x_{n-1}} (x - x_n) \quad \text{——————(1)}$$



Now, $y = 0$ and $x = x_{n+1}$

$$\Rightarrow \quad -y_n = \frac{y_n - y_{n-1}}{x_n - x_{n-1}} (x_{n+1} - x_n)$$

$$\therefore \quad x_{n+1} = x_n - \frac{(x_n - x_{n-1})}{(y_n - y_{n-1})} y_n$$

$$\Rightarrow \quad \boxed{x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n)}$$

Equation of secant joining (x_{n-1}, y_{n-1}) and (x_n, y_n) can also be written as

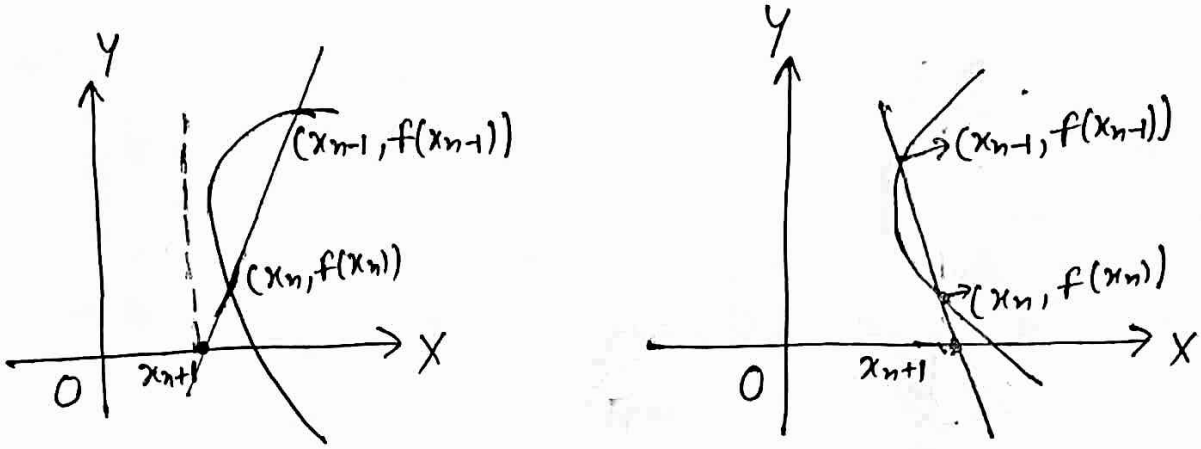
$$y - y_{n-1} = \frac{y_n - y_{n-1}}{x_n - x_{n-1}} (x - x_{n-1})$$

Now $y = 0$ and $x = x_{n+1}$

$$\Rightarrow \quad -y_{n-1} = \frac{y_n - y_{n-1}}{x_n - x_{n-1}} (x_{n+1} - x_{n-1})$$

$$\Rightarrow \boxed{x_{n+1} = x_{n-1} - \frac{(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})} f(x_{n-1})}$$

which is the iterative formula to find the approximations.



In figure (1), $f(x_{n+1})$ cannot be found and hence iteration process diverges but in figure (2) iteration process converges to root.

Note:
- It does not require the condition $f(x_0) \cdot f(x_1) < 0$.
- Two most recent approximations to the root are used to find the next approximation
- Also it is not necessary that the iteration process converge i.e, contain the root in (x_n, x_{n+1}).

Convergence of secant method : Let r be exact root of $f(x)$.

Let x_{n-1}, x_n and x_{n+1} be its successive approximations. Then by the iterative formula of secant method

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n)$$

Substituting $x_n = r + e_n$, we have

$$r + e_{n+1} = r + e_n - \frac{(r+e_n) - (r+e_{n-1})}{f(r+e_n) - f(r+e_{n-1})} f(r+e_n)$$

$$\Rightarrow e_{n+1} = \frac{e_n f(r+e_n) - e_n f(r+e_{n-1}) - (e_n - e_{n-1}) f(r+e_n)}{f(r+e_n) - f(r+e_{n-1})}$$

$$= \frac{e_{n-1} f(r+e_n) - e_n f(r+e_{n-1})}{f(r+e_n) - f(r+e_{n-1})}$$

$$= \frac{e_{n-1}\left[f(r) + e_n f'(r) + \frac{e_n^2}{2!} f''(r) + \cdots\right] - e_n\left[f(r) + e_{n-1} f'(r) + \frac{e_{n-1}^2}{2!} f''(r) + \cdots\right]}{\left[f(r) + e_n f'(r) + \frac{e_n^2}{2!} f''(r) + \cdots\right] - \left[f(r) + e_{n-1} f'(r) + \frac{e_{n-1}^2}{2!} f''(r) + \cdots\right]}$$

(using Taylor series)

$$= \frac{\frac{e_n e_{n-1}}{2}(e_n - e_{n-1}) f''(r) + \cdots}{(e_n - e_{n-1}) f'(r) + \frac{(e_n^2 - e_{n-1}^2)}{2} f''(r) + \cdots} \qquad (\because f(r) = 0)$$

$$\therefore e_{n+1} \equiv \frac{e_n e_{n-1}}{2} \frac{f''(r)}{f'(r)} + \cdots$$

$$\Rightarrow e_{n+1} = A e_n e_{n-1} \quad \text{where } A = \frac{f''(r)}{f'(r)} \quad \text{(if the remaining terms are neglected)} \qquad \text{---------- (1)}$$

Let m be the order of convergence, then we can find k such that $|e_n| = k|e_{n-1}|^m$ for some k ——— (2)

$\therefore$ From (1) and (2),

$$|e_{n+1}| = |A||e_n||e_{n-1}|$$

and $$|e_{n-1}| = \left(\frac{|e_n|}{k}\right)^{1/m}$$

$$\Rightarrow \quad |e_{n+1}| = \frac{|A|}{k^{1/m}} |e_n|^{1+\frac{1}{m}}$$

But order of convergence is m

$\therefore$ From this we get $m = 1 + \frac{1}{m}$

or $\quad m^2 - m - 1 = 0$

$$\Rightarrow \quad m = \frac{1 \pm \sqrt{1+4}}{2} = \frac{1 \pm \sqrt{5}}{2}$$

But $m > 0$

$$\therefore \quad m = \frac{1+\sqrt{5}}{2} = 1.62$$

$\therefore$ order of convergence is 1.62.

Que Find a root of the equation $x^3-2x-5=0$ using secant method correct to three decimal places.

Sol Let $f(x)=x^3-2x-5=0$

$\therefore f(2)=-1,\ f(3)=27-6-5=16$

$\therefore$ Taking initial approximations $x_0=2$ and $x_1=3$, by secant method, we have

$$x_{n+1}=x_n-\frac{x_n-x_{n-1}}{f(x_n)-f(x_{n-1})}f(x_n) \qquad \text{——— (1)}$$

$$\therefore x_2=x_1-\frac{x_1-x_0}{f(x_1)-f(x_0)}f(x_1)=3-\frac{(3-2)}{16+1}.16=3-\frac{16}{17}=2.058823$$

Now, $f(x_2)=-0.390799$

$$\therefore x_3=x_2-\frac{x_2-x_1}{f(x_2)-f(x_1)}f(x_2)=2.058823-\frac{2.058823-3}{-0.390799-16}(-0.390799)$$

$$\Rightarrow x_3=2.081263$$

$$f(x_3)=-0.147204$$

$$\therefore x_4=x_3-\frac{x_3-x_2}{f(x_3)-f(x_2)}f(x_3)=2.094824$$

and $f(x_4)=0.003042$

$$\therefore x_5=x_4-\frac{x_4-x_3}{f(x_4)-f(x_3)}f(x_4)=2.094549$$

Hence the root is 2.095 correct to 3 decimal places. ∎

Unconstrained one variable function minimization

In an unconstrained one variable minimization problem, a function $f: R \to R$ is defined and a point $z \in R$ is sought with the property that $f(z) \leq f(x) \quad \forall \; x \in R$

Note that if no assumptions are made about f, this problem is insoluble in its general form.

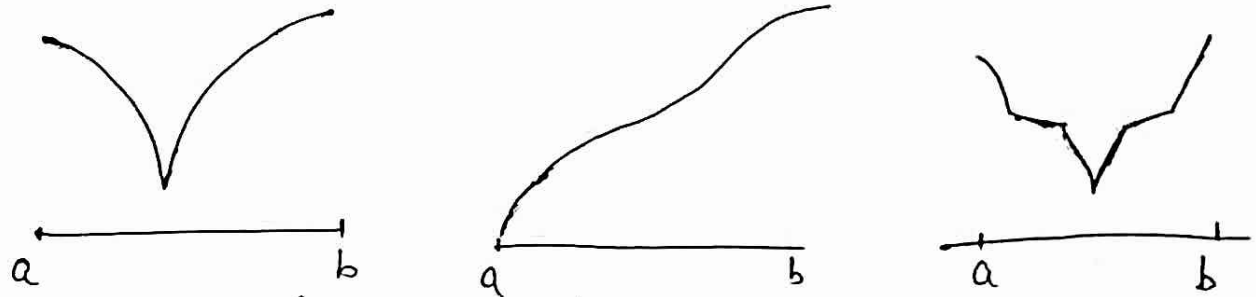
In attaking a minimization problem, one reasonable assumption is that on some interval $[a,b]$ given to us in advance, f has only a single local minimum. This property is often expressed by saying that f is unimodal on $[a,b]$.

Note:- A point z is a local minimum point of a function f there is some neighbourhood of z in which all points satisfy

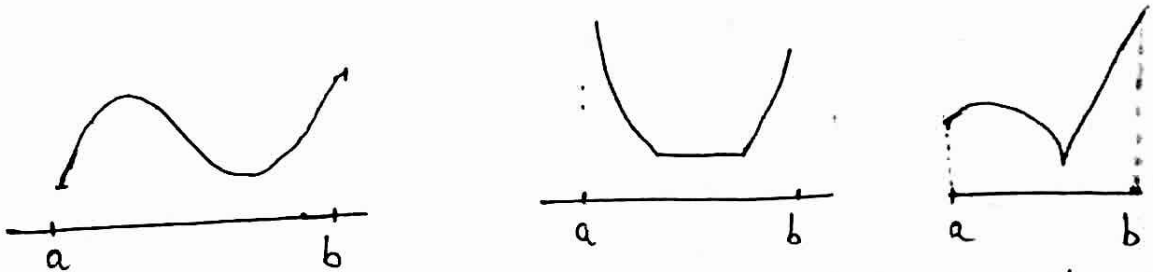
$$f(z) \leq f(x).$$

- An important property of a continuous unimodal function is that it is strictly decreasing up to the minimum point and strictly increasing thereafter.

Examples:



(Three unimodal functions)

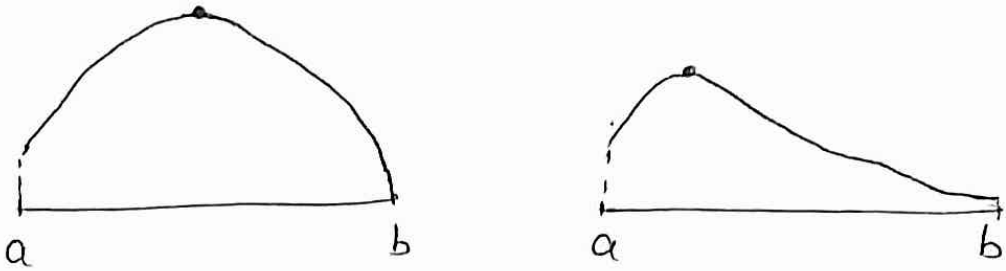


(Three functions that are not unimodal)

Unimodal function for maximization problem:

A function $f: R \to R$ that has only one local maximum in a given interval $[a, b]$ is also called a unimodal function.

Ex:



Fibonacci numbers: Fibonacci numbers are defined as

$$F_0 = F_1 = 1$$

$$F_n = F_{n-1} + F_{n-2} \; ; \; n \geq 2$$

$\therefore$ Fibonacci sequence $\{F_n\} = \{F_0, F_1, F_2, F_3, F_4, F_5, F_6, F_7, \ldots\}$

$= \{1, 1, 2, 3, 5, 8, 13, 21, \ldots\}$

Fibonacci search method:

- This method is an elimination technique and makes the use of Fibonacci numbers.
- Here we want to minimize a continuous unimodal function f over $[a, b]$ i.e., find $x \in [a, b]$ which minimize $f(x)$.

Note: The problem to maximize a continuous unimodal function f over $[a, b]$ can also be solved by this method.

- $L_0 = b - a$ is the length of the initial interval.
- L_n denote the length of the interval of uncertainty after n experiments.

- The number of steps n must be given in advance or the desired accuracy (tolerance) ε must be given in advance to find n.
- Divide the initial interval $[a,b]$ equally into F_n subintervals and hence the length of each subinterval is $\frac{1}{F_n}(b-a)$.

$$\therefore \quad L_n = \frac{1}{F_n}(b-a) = \frac{1}{F_n} L_0$$

$$\Rightarrow \boxed{\frac{L_n}{L_0} = \frac{1}{F_n}} \qquad \text{——————} (1)$$

- For a given tolerance ε with exact value of x, we can find n such that

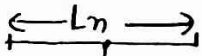
$$\boxed{\frac{L_n}{2} \leq \varepsilon}$$

and using (1), we get

$$\frac{L_0}{F_n} \leq 2\varepsilon$$

$$\Rightarrow F_n \geq \frac{L_0}{2\varepsilon}$$

$$\Rightarrow F_n \geq \frac{b-a}{2\varepsilon}$$



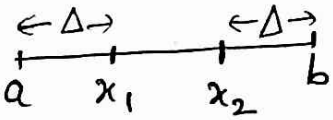
at mid-point we get the optimal value of x (i.e, $\hat{x}$) and exact value of x lies in the final interval of length L_n either on the left of $\hat{x}$ or right of $\hat{x}$ or exactly at $\hat{x}$.

the smallest value of n satisfying this inequality can be used as no. of steps n. $(n \in N)$

Fibonacci search algorithm

After fixing the no. of steps as n, we define a sequence of intervals starting with the given interval $[a, b]$ of length $L_0 = b-a$ and for $k = n, n-1, \dots 3$ use these formulas for updating

$$\Delta = \left(\frac{F_{k-2}}{F_k}\right)(b-a)$$



$$x_1 = a + \Delta, \quad x_2 = b - \Delta$$

$$\begin{cases} a = x_1, & \text{if } f(x_1) \geq f(x_2) \\ b = x_2 & \text{if } f(x_1) < f(x_2) \end{cases}$$

At the step $k = 2$,

$$x_1 = \frac{1}{2}(a+b) - 2\delta \qquad (\text{Take } 2\delta < \varepsilon)$$

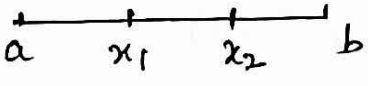
$$x_2 = \frac{1}{2}(a+b) + 2\delta$$

$$\begin{cases} a = x_1, & \text{if } f(x_1) \geq f(x_2) \\ b = x_2, & \text{if } f(x_1) < f(x_2) \end{cases}$$

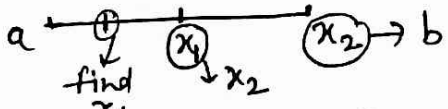
and we have the final interval $[a, b]$ from which we compute $\hat{x} = \frac{1}{2}(a+b)$.

This algorithm requires only one function evaluation per step after the initial step.

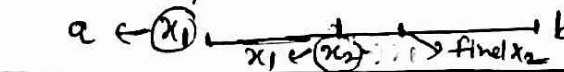
Note: (1) $x_1 + x_2 = a + b$ (always)



(2) If $f(x_1) < f(x_2)$ in $[a, b]$, then next interval of uncertainty $= [a, x_2]$



(3) If $f(x_1) \geq f(x_2)$ in $[a, b]$, then next interval of uncertainty $= [x_1, b]$



Que (1) Find the Minimum of $f(x) = x^2 - 6x + 2$ on $[0, 10]$ using Fibonacci search algorithm. Obtain the optimal value with tolerance $\varepsilon = \frac{1}{4}$.

or

(2) Find the minimum of $f(x) = x^2 - 6x + 2$ on $[0, 10]$ using Fibonacci search algorithm by taking $n = 7$.

or

(3) Find the minimum of $f(x) = x^2 - 6x + 2$ on $[0, 10]$ using Fibonacci search algorithm. Locate the value of x within 2.5% of exact value.

Sol Let L_0 be the length of the initial interval $[a, b] = [0, 10]$ and L_n be the length of final interval of uncertainty after n experiments.

Here $a_0 = 0$, $b = 10$, $L_0 = b - a = 10$

(1) For given $\varepsilon = \frac{1}{4}$, we have $\frac{L_n}{2} \leq \frac{1}{4} \Rightarrow L_n \leq \frac{1}{2}$

$$\Rightarrow \frac{L_0}{F_n} \leq \frac{1}{2} \qquad \left(\because \frac{L_n}{L_0} = \frac{1}{F_n}\right)$$

$$\Rightarrow F_n \geq 10 \times 2$$

$$\Rightarrow F_n \geq 20$$

$\therefore$ Smallest n for which this inequality is satisfied is $n = 7$. $(\because F_7 = 21)$

(3) Given $\frac{L_n}{2} \leq 2.5$ of L_0

$$\Rightarrow L_n \leq \frac{5}{100} \times 10 \Rightarrow L_n \leq \frac{1}{2} \Rightarrow F_n \geq 20 \text{ (as above)}$$

$$\Rightarrow n = 7$$

So, in all, we find $n = 7$.

• if $f(x_1) < f(x_2)$ in $[a,b]$ new interval of uncertainty $= [a, x_2]$

• if $f(x_1) \geq f(x_2)$ in $[a,b]$ then $[x_1, b]$

Here $n=7$, $a=0$, $b=10$ (for initial interval)

k	$\frac{F_{k-2}}{F_k}$	a	b	$x_1 = a + \frac{F_{k-2}}{F_k}(b-a)$	$x_2 = b - \frac{F_{k-2}}{F_k}(b-a)$	$f(x_1) = x_1^2 - 6x_1 + 2$	$f(x_2) = x_2^2 - 6x_2 + 2$	Min at
$k=7$	$\frac{F_5}{F_7} = \frac{8}{21}$	0	10	3.810	6.190	−6.344	3.176	x_1
$k=6$	$\frac{F_4}{F_6} = \frac{5}{13}$	0	6.190	2.380	3.810	−6.616	−6.344	x_1
$k=5$	$\frac{F_3}{F_5} = \frac{3}{8}$	0	3.810	1.43	2.380	−4.535	−6.616	x_2
$k=4$	$\frac{F_2}{F_4} = \frac{2}{5}$	1.43	3.810	2.380	2.860	−6.616	−6.980	x_2
$k=3$	$\frac{F_1}{F_3} = \frac{1}{3}$	2.380	3.810	2.860	3.330	−6.980	−6.891	x_1
$k=2$	$\frac{F_0}{F_2} = \frac{1}{2}$	2.380	3.330	Here $x_1 = \frac{1}{2}(a+b) - 0.2$ $= 2.655$	2.860	−6.881	−6.980	x_2

(by taking $2\delta = 0.2 < \varepsilon = 0.25$)

∴ Final interval of uncertainty $= [x_1, b] = [2.655, 3.330]$

∴ $\hat{x} = \frac{2.655 + 3.330}{2} = \frac{5.985}{2} = 2.99$

∴ for $x = 2.89$, Optimal value $= (2.99)^2 - 6 \times 2.99 + 2 = -6.999 \approx -7$

Note: If the problem is of maximization, then write last column as Max. at and make the changes accordingly.

Golden section search method:

In Fibonacci method, the ratio for the reduction of intervals is not constant and the number of subintervals (iterations) is predetermined which are based on the specified tolerance. while in golden section search the ratio of intervals is constant i.e., it depends on a ratio ρ known as the golden section ratio.

Golden Ratio:

The ratio of the smaller part of a line segment to the larger part is the same as the ratio of the larger part to the whole line segment.

For a line segment of length 1, denote the larger part by r and the smaller part by $1-r$ as shown here:

Hence, we have the ratios $\frac{1-r}{r} = \frac{r}{1}$ and we obtain the quadratic equation $r^2 = 1-r$ or $r^2+r=1$

The equation $r^2+r-1=0$ has two roots as

$$r = \frac{-1+\sqrt{5}}{2} \approx 0.61803\text{----} \text{ and } r = \frac{-1-\sqrt{5}}{2} \approx -1.61803\text{----}$$

The reciprocal of the positive root is the golden ratio ρ i.e.,

$$\rho = \frac{1}{r} \quad (r>0)$$

$$= \frac{2}{\sqrt{5}-1} = \frac{2(\sqrt{5}+1)}{5-1} = \frac{\sqrt{5}+1}{2} \approx 1.61803\text{----}$$

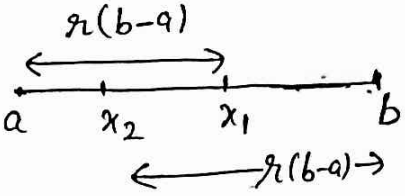
The Golden Section Search Algorithm:

Our problem is Min $f(x)$
s.t. $x \in [a,b]$ where $f(x)$ is continuous and unimodal.

We can follow the following algorithm:

Step 1 Find two intermediate points x_1 and x_2 such that

$$x_1 = a + r(b-a)$$
$$x_2 = b - r(b-a)$$
where $r = \frac{\sqrt{5}-1}{2} \approx 0.61803\ldots$



Step 2 Evaluate $f(x_1)$ and $f(x_2)$.

If $f(x_1) > f(x_2)$, then interval of uncertainty is $[a, x_1]$ and

$$a = a$$
$$b = x_1$$
$$x_1 = x_2$$
$$x_2 = b - r(b-a)$$

If $f(x_1) \leq f(x_2)$, then interval of uncertainty is $[x_2, b]$ and

$$b = b$$
$$a = x_2$$
$$x_2 = x_1$$
$$x_1 = a + r(b-a)$$

Step 3 If $b-a < \varepsilon$ (a sufficiently smaller number according to desire accuracy), then minimum occurs at $\frac{a+b}{2}$ and stop iterating, else go to step 2.

Note: If problem is of maximization, then choose the interval of uncertainty according to that and make the changes on the same way.

For eg: For maximization problem, if $f(x_1) > f(x_2)$ in step 2, then interval is $[x_2, b]$ and then $a = x_2$, $b = b$, $x_2 = x_1$, $x_1 = a + r(b-a)$.

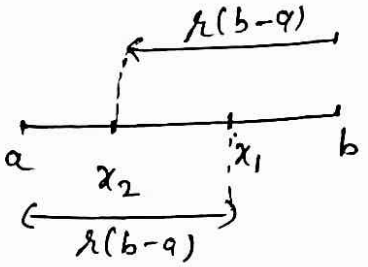
Que Use the golden section search to find the value of x that minimizes $f(x) = x^2 - 6x + 2$ in the range $[0, 10]$. Locate this value of x to within a range of 0.25.

Sol Given $f(x) = x^2 - 6x + 2$, $a = 0$, $b = 10$ (for initial interval)

x_1 and x_2 are two intermediate points in $[a, b]$ such that

$$\left.\begin{aligned} x_1 &= a + r(b-a) \\ x_2 &= b - r(b-a) \end{aligned}\right\} \Rightarrow x_1 + x_2 = a + b$$

where $r = \frac{\sqrt{5}-1}{2} \approx 0.61803$



No. of steps (n)	a	b	$x_1 = a + 0.61803(b-a)$	$x_2 = b - 0.61803(b-a)$	$f(x_1) = x_1^2 - 6x_1 + 2$	$f(x_2) = x_2^2 - 6x_2 + 2$	Min at x_1/x_2
n=1	0	10	6.1803	3.8197	3.1143	−6.3281	x_2
n=2	0	6.1803	3.8197	2.3606	−6.3281	−6.5912	x_2
n=3	0	3.8197	2.3606	1.4591	−6.5912	−4.6256	x_1
n=4	1.4591	3.8197	2.9182	2.3606	−6.9933	−6.5912	x_1
n=5	2.3606	3.8197	3.2621	2.9182	−6.9313	−6.9933	x_2
n=6	2.3602	3.2621	2.9182	2.7041	−6.9933	−6.9124	x_1
n=7	2.7041	3.2621	3.048	2.9182	−6.9977	−6.9933	x_1
n=8	2.9182	3.2621	3.1323	3.048	−6.9825	−6.9977	x_2
	2.9182	3.1323					

Since $[3.1323 - 2.9182) = 0.2141 < 0.25$, we stop here.

$\therefore x^* = \frac{2.9182 + 3.1323}{2} = 3.025$ and $f(x^*) = -6.999 \approx -7$

Number of steps required to reach accuracy ε by the golden section search method (where ε is the range for final interval)

After step 1, we get two evaluations and the length of new reduced interval $= r(b-a)$

After step 2, we get three evaluations and the length of new reduced interval $= r^2(b-a)$

⋮

After step n, we get $(n+1)$ evaluations and the length of new reduced interval $= r^n(b-a)$

To reach the accuracy ε, $r^n(b-a) \leq \varepsilon$ where $r = 0.61803$ the smallest value of n satisfying this inequality gives the no. of steps required. $(n \in N)$

Que Find the no. of steps required to reach the value of x within a range of 0.25 to minimize the function

$$f(x) = x^2 - 2x + 10 \; ; \; x \in [0, 10]$$

by using golden section search method.

Sol Let the no. of steps required be n.

Given $a = 0$, $b = 10$, $\varepsilon = 0.25$. To reach the accuracy ε,

$$r^n(b-a) \leq \varepsilon \quad \text{where } r = 0.61803$$

$$\Rightarrow (0.61803)^n \times 10 \leq 0.25 \Rightarrow (0.61803)^n \leq 0.025$$

$$\Rightarrow n \log(0.61803) \leq \log(0.025)$$

$$\Rightarrow n \times (-0.48122) \leq (-3.68888)$$

$$\Rightarrow n \geq \frac{3.68888}{0.48122} = 7.6657 \Rightarrow \boxed{n = 8}$$ ∎

Note: It is already verified in the previous question.

Newton's method for unconstrained one variable minimization:

Minimize $f(x)$ when x_0, the initial guess for x is given or some interval is given in which x_0 lies.

- We assume $f'(x)$ and $f''(x)$ exists for each measurement point x_n.
- We can fit a quadratic function through x_n that matches its first and second derivatives with that of the function f.

$$q(x) = f(x_n) + (x - x_n) f'(x_n) + \frac{1}{2}(x - x_n)^2 f''(x_n)$$

Clearly $q(x_n) = f(x_n)$, $q'(x_n) = f'(x_n)$ and $q''(x_n) = f''(x_n)$

- Instead of minimizing f, we minimize its approximation q.
- For minimizing q, the necessary condition is

$$q'(x) = 0$$

$$\Rightarrow f'(x_n) + (x - x_n) f''(x_n) = 0$$

$$\Rightarrow x = x_n - \frac{f'(x_n)}{f''(x_n)}$$

Setting $x = x_{n+1}$, we obtain

$$\boxed{x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}}, \quad n = 0, 1, 2, 3, \ldots$$

Note: • We stop when $|x_{n+1} - x_n| < \varepsilon$ for a given ε (accuracy).

- Newton's method work well if $f''(x) > 0$ everywhere. However, if $f''(x) < 0$ for some x, Newton's method may fail to converge to the minimizer.

Que Using Newton's method, find the minimum of

$$f(x) = \frac{1}{2}x^2 - \sin x \quad ; \; x_0 = 0.5$$

when the accuracy required $\varepsilon = 10^{-5} = 0.00001$ for x.

Sol Given $f(x) = \frac{1}{2}x^2 - \sin x$, $x_0 = 0.5$

$$\therefore f'(x) = x - \cos x$$

$$f''(x) = 1 + \sin x$$

According to Newton's method,

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}, \quad n = 0, 1, 2, 3, \ldots$$

$$\Rightarrow x_{n+1} = x_n - \frac{(x_n - \cos x_n)}{(1 + \sin x_n)}$$

$$= \frac{x_n + x_n \sin x_n - x_n + \cos x_n}{1 + \sin x_n}$$

$$\therefore x_{n+1} = \frac{x_n \sin x_n + \cos x_n}{1 + \sin x_n}$$

n	x_n	$x_n \sin x_n + \cos x_n$	$1 + \sin x_n$	x_{n+1}
0	0.5	1.117295	1.479426	0.755222
1	0.755222	1.245787	1.685450	0.739274
2	0.739274	1.237045	1.673752	0.739085
3	0.739085	1.236942	1.673612	0.739085

From last two iterations $|x_3 - x_4| = 0 < \varepsilon = 10^{-5}$

$\therefore$ $x^* = 0.739085$

and $f(x^*) = -0.400489$ ∎

Method of Steepest Descent for multivariate function minimization :

Minimize $f(X)$ with an initial guess X_0

where $f: R^n \to R$

[$R^n = \{(x_1, x_2, .., x_n) : x_1, x_2, .. x_n \in R\}$ is known as n-dimensional space over R.

An important property of the gradient vector $\nabla f(X)$ is that it points in the direction of the most rapid increase in the function f, which is the direction of steepest ascent.

[Note: If problem is of maximization, then convert it to minimization as $\min f(X) = -\max f(X)$ and proceed on the same way

Conversely, $-\nabla f(X)$ points in the direction of the steepest descent.

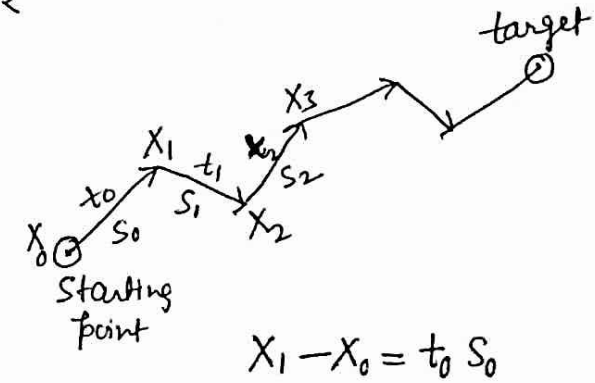
On the basis of this steepest descent method for minimizing the multivariate function $f: R^n \to R$ can be described as:

For $k = 0, 1, 2, \ldots$

Step 1: Start with initial guess X_0 i.e., for $k = 0$.

Step 2: Find the search direction S_k as

$$S_k = -\nabla f_k = -\nabla f(X_k)$$



$X_1 - X_0 = t_0 S_0$

Step 3: Determine the step length t_k by minimizing the function

$$\phi(t) = f(X_k + t S_k)$$

[Put $\phi'(t) = 0$ and find $t = t_k > 0$ such that $\phi''(t_k) > 0$]

Step 4: Find $X_{k+1} = X_k + t_k S_k$ and check the optimality for the new point X_{k+1} as

If $\nabla f(X_{k+1}) \cong 0$ or $t_k = 0$, then stop otherwise go to step (1) again for X_{k+1}.

Que By the method of steepest descent minimize the function $f(x_1, x_2) = x_1 - x_2 + 2x_1^2 + 2x_1x_2 + x_2^2$ starting from the point $X_0 = (0,0)$.

Sol Given $f(x_1, x_2) = x_1 - x_2 + 2x_1^2 + 2x_1x_2 + x_2^2$

$\therefore \nabla f(x_1, x_2) = (1 + 4x_1 + 2x_2, -1 + 2x_1 + 2x_2)$ ——— (1)

$\left(\because \nabla f(x_1, x_2) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}\right) \right)$

Iteration 1: $k = 0$, $X_0 = (0,0)$

$\therefore S_0 = -\nabla f(X_0) = -\nabla f(0,0) = -(1,-1) = (-1, 1)$

We now minimize the function

$\phi(t) = f(X_0 + tS_0) = f(-t, t)$

{ To find $\phi'(t)$ we can find $\phi(t)$ in terms of t by def. of f and then $\phi'(t)$ but it is complicated

$\therefore \phi'(t) = \nabla f(-t, t) \cdot S_0 = (1 - 4t + 2t, -1 - 2t + 2t) \cdot (-1, 1)$

$= (1 - 2t, -1) \cdot (-1, 1)$

$= -1 + 2t - 1 = 2(t - 1)$

$\therefore \phi'(t) = 0 \Rightarrow t = 1$

and $\phi''(t) = 2 > 0$ for $t = 1$

$\therefore \phi(t)$ is minimum at $t = 1 = t_0$

$\therefore$ the new point is

$X_1 = X_0 + t_0 S_0 = (0,0) + 1(-1,1) = (-1, 1)$

$\therefore \nabla f(X_1) = \nabla f(-1, 1) = (1 - 4 + 2, -1 - 2 + 2) = (-1, -1)$

by (1) $\neq (0,0)$

$\therefore$ We go to next iteration with $X_1 = (-1, 1)$ (i.e, for $k = 1$)

Iteration 2: $k = 1$, $X_1 = (-1, 1)$

$\therefore S_1 = -\nabla f(X_1) = (1, 1)$

We now minimize the function

$\phi(t) = f(X_1 + tS_1) = f(-1 + t, 1 + t)$

$\therefore \phi'(t) = \nabla f(-1+t, 1+t) \cdot S_1$

$= (1-4+4t+2+2t, -1+2t-2+2+2t) \cdot (1,1)$ by (1)

$= (-1+6t, -1+4t) \cdot (1,1)$

$= -1+6t-1+4t$

$= 2(5t-1)$

$\therefore \phi'(t) = 0 \Rightarrow t = \frac{1}{5}$ and $\phi''(t) = 10 > 0$

$\therefore \phi(t)$ is minimum at $t = \frac{1}{5} = t_1$

$\therefore$ the new point is

$X_2 = X_1 + t_1 S_1 = (-1,1) + \frac{1}{5}(1,1) = \left(-\frac{4}{5}, \frac{6}{5}\right)$

$\therefore \nabla f(X_2) = \nabla f\left(-\frac{4}{5}, \frac{6}{5}\right)$

$= \frac{1}{5}(5-16+12, -5-8+12) = \left(\frac{1}{5}, -\frac{1}{5}\right) \neq (0,0)$

$\therefore X_2$ is not optimum. So, we go to next iteration with X_2 (i.e, for $k=2$)

Iteration 3: $k = 2$, $X_2 = \left(-\frac{4}{5}, \frac{6}{5}\right)$

$\therefore S_2 = -\nabla f(X_2) = \left(-\frac{1}{5}, \frac{1}{5}\right)$

We now minimize the function

$\phi(t) = f(X_2 + tS_2) = f\left(\frac{-4-t}{5}, \frac{6+t}{5}\right)$

$\therefore \phi'(t) = \nabla f\left(\frac{-4-t}{5}, \frac{6+t}{5}\right) \cdot S_2$

$= \left(\frac{5-16-4t+12+2t}{5}, \frac{-5-8-2t+12+2t}{5}\right) \cdot \left(-\frac{1}{5}, \frac{1}{5}\right)$

$= \frac{1}{25}(1-2t, -1) \cdot (-1,1) = \frac{1}{25}(-1+2t-1) = \frac{2}{25}(t-1)$

$\therefore \phi(t)$ is minimum at $t = 1 = t_2$ $\left(\because \phi''(t) = \frac{2}{25} > 0\right)$

$\therefore$ the new point is $X_3 = X_2 + t_2 S_2$

$= \left(-\frac{4}{5}, \frac{6}{5}\right) + \left(-\frac{1}{5}, \frac{1}{5}\right) = \left(-1, \frac{7}{5}\right)$

$\therefore \nabla f(X_3) = \nabla f\left(-1, \frac{7}{5}\right)$

$= \left(1-4+\frac{14}{5}, -1-2+\frac{14}{5}\right)$ by (1)

$= \left(-\frac{1}{5}, -\frac{1}{5}\right) \neq (0,0)$

$\therefore$ We go to next iteration with $X_3 = \left(-1, \frac{7}{5}\right)$ (i.e, for $k=3$)

Iteration 4: $k=3$, $X_3 = \left(-1, \frac{7}{5}\right) = \left(-\frac{5}{5}, \frac{7}{5}\right)$

$\therefore S_3 = -\nabla f(X_3) = \left(\frac{1}{5}, \frac{1}{5}\right)$

We now minimize the function

$\phi(t) = f(X_3 + tS_3)$

$= f\left(\frac{-5+t}{5}, \frac{7+t}{5}\right)$

$\nabla f\left(\frac{-5+t}{5}, \frac{7+t}{5}\right) = \frac{1}{5}\left(5-20+4t+14+2t, -5-10+2t+14+2t\right)$

$= \frac{1}{5}(-1+6t, -1+4t)$

$\therefore \phi'(t) = \nabla f\left(\frac{-5+t}{5}, \frac{7+t}{5}\right) \cdot S_3$

$= \frac{1}{5}(-1+6t, -1+4t) \cdot \left(\frac{1}{5}, \frac{1}{5}\right)$

$= \frac{1}{25}(-1+6t-1+4t) = \frac{2}{25}(5t-1)$

$\therefore \phi'(t) = 0 \Rightarrow t = \frac{1}{5}$ and $\phi''(t) = \frac{2}{5} > 0$

$\therefore \phi(t)$ is minimum at $t = \frac{1}{5} = t_3$

$\therefore$ the new point is $X_4 = X_3 + t_3 S_3$

$= \left(-\frac{5}{5}, \frac{7}{5}\right) + \frac{1}{5}\left(\frac{1}{5}, \frac{1}{5}\right)$

$= \left(-\frac{24}{25}, \frac{36}{25}\right)$

$\therefore \nabla f(X_4) = \nabla f\left(-\frac{24}{25}, \frac{36}{25}\right) = \left(\frac{25-96+72}{25}, \frac{-25-48+72}{25}\right)$

$= \left(\frac{1}{25}, -\frac{1}{25}\right) = (0.04, -0.04)$

$\therefore \nabla f(X_4) = (0.04, -0.04) \cong (0,0)$ (upto one decimal place)

$\therefore$ At this step optimum value of X is

$$X^* = \left(\frac{-24}{25}, \frac{36}{25}\right) = (-0.96, 1.44) \quad A_2$$

Note: If we want to find more exact answer then we can proceed further for next iteration as:

Iteration 5: $k=4$, $X_4 = \left(\frac{-24}{25}, \frac{36}{25}\right)$

$\therefore S_4 = -\nabla f(X_4) = \left(\frac{-1}{25}, \frac{1}{25}\right)$

We now minimize the function

$$\phi(t) = f(X_4 + tS_4) = f\left(\frac{-24-t}{25}, \frac{36+t}{25}\right)$$

$$\therefore \phi'(t) = \nabla f\left(\frac{-24-t}{25}, \frac{36+t}{25}\right) \cdot S_4$$

$$= \left(\frac{25-96-4t+72+2t}{25}, \frac{-25-48-2t+72+2t}{25}\right) \cdot \left(\frac{-1}{25}, \frac{1}{25}\right)$$

$$= \left(\frac{1-2t}{25}, \frac{-1}{25}\right) \cdot \left(\frac{-1}{25}, \frac{1}{25}\right)$$

$$= \frac{-1+2t-1}{625} = \frac{2(t-1)}{625}$$

$\therefore \phi'(t) = 0 \Rightarrow t = 1$ and $\phi''(t) = \frac{2}{625} > 0$ for $t = 1$

$\therefore \phi(t)$ is minimum at $t = 1 = t_4$

$\therefore$ the new point is $X_5 = X_4 + t_4 S_4$

$$= \left(\frac{-24}{25}, \frac{36}{25}\right) + \left(\frac{-1}{25}, \frac{1}{25}\right) = \left(-1, \frac{37}{25}\right)$$

$$\nabla f(X_5) = \nabla f\left(-1, \frac{37}{25}\right) = \left(1-4+\frac{74}{25}, -1-2+\frac{74}{25}\right)$$

$$= \left(-\frac{1}{25}, -\frac{1}{25}\right) = (-0.04, -0.04)$$

$$\neq (0,0)$$

$\therefore$ We go to next iteration with X_5 i.e, for $k=5$

Iteration 6: $k=5$, $X_5 = \left(-1, \frac{37}{25}\right)$

$\therefore S_5 = -\nabla f(X_5) = \left(\frac{1}{25}, \frac{1}{25}\right)$

We now minimize the function

$$\phi(t) = f(X_5 + tS_5) = f\left(\frac{-25+t}{25}, \frac{37+t}{25}\right)$$

$$\therefore \phi'(t) = \nabla f\left(\frac{-25+t}{25}, \frac{37+t}{25}\right) \cdot S_5$$

$$= \left(\frac{25-100+4t+74+2t}{25}, -\frac{25-50+2t+74+2t}{25}\right) \cdot \left(\frac{1}{25}, \frac{1}{25}\right) \quad \text{by (1)}$$

$$= \frac{1}{625}(-1+6t, -1+4t) \cdot (1,1)$$

$$= \frac{1}{625}(-1+6t-1+4t) = \frac{2}{625}(5t-1)$$

$\therefore \phi'(t) = 0 \Rightarrow t = \frac{1}{5}$ and $\phi''(t) > 0$

$\therefore \phi(t)$ is minimum at $t = \frac{1}{5} = t_5$

$\therefore$ the new point is $X_6 = X_5 + t_5 S_5 = \left(\frac{-25}{25}, \frac{37}{25}\right) + \frac{1}{5}\left(\frac{1}{25}, \frac{1}{25}\right)$

$$= \left(\frac{-124}{125}, \frac{186}{125}\right)$$

$$= (0.992, 1.46)$$

and $\nabla f(X_6) = \nabla f\left(\frac{-124}{125}, \frac{186}{125}\right)$

$$= \left(\frac{125-496+372}{125}, \frac{-125+248+372}{125}\right)$$

$$= \left(\frac{1}{125}, \frac{1}{125}\right)$$

$$= (0.008, 0.008) \cong (0,0)$$

$\therefore$ We stop here and hence $X^* = X_6 \simeq (1.0, 1.5)$ Ans

$f(x_1, x_2) = 1 - 1.5 + 2 + 3 + 2.25$ (correct to one decimal places)

$= 6.75$