CEN-204 Numerical Analysis

Textbook: Numerical Analysis
Mathematics of Scientific Computing
David Kincaid
Ward Cheney

-Numerical Analysis involves the study, development, and analysis of algorithms for obtaining numerical solutions to various mathematical problems.

Chapter-1 Mathematical Preliminaries

- **-Limit**
- If *f* is a real-valued function of a real variable, then the limit of the function *f* at *c* is defined as follows:
- $\lim_{x \to c} f(x) = L$
- means that *f* can be made to be as close to *L* as desired by making *x* sufficiently close to *c*.
- Or $|f(x) - L| < \varepsilon$ whenever $|x - c| < \delta$
- Or for each *ε>0* there is a *δ>0* such that $0 < |x - c| < \delta$ makes $|f(x) - L| < \varepsilon$
- **Example:** Show that $\lim_{x \to 2} x^2 = 4$
- **Solution:** Let $\delta = -2 + \sqrt{4 + \varepsilon} > 0 \Rightarrow \delta(\delta + 4) = (-2 + \sqrt{4 + \varepsilon})(2 + \sqrt{4 + \varepsilon}) = \varepsilon$

For each **ε>0** there is $\delta=-2+\sqrt{4+\varepsilon}$ $(0<|x-2|<\delta)$ that makes $|x^2-4|<\varepsilon$

Example: $\lim_{x\to 0}\frac{|x|}{x}$ does not exist (left and right limits are not the same)

- The function f(x) is said to be continuous at c if $\lim_{x\to c} f(x)=f(c)$

Example: $f(x)=x^2$ is continuous at *x=2.*

Example: $f(x)=\frac{|x|}{x}$ is not continuous at *x=0.*

- Derivative of *f(x)* at *c* is defined by the equation
-
- $f'(x)=\lim_{x\to c}\frac{f(x)-f(c)}{x-c}$

If *f(x)* is a function for which $f'(c)$ exists, we say *f(x)* is differentiable at c. If f(x) is differentiable at c, then *f(x)* must be continuous at c. But the

reverse is not true.
Proof: $f(x)=|x|$ is a continuous function but not a differentiable function.

$$\lim_{x\to c}[f(x)-f(c)]=\lim_{x\to c}\frac{f(x)-f(c)}{x-c}(x-c)=f'(c)\lim_{x\to c}(x-c)=f'(c).0$$

$\Rightarrow if\ f'(c)$ exists then f(x) is continuous at c.

- The set of all functions that are continuous on the real line R is denoted by *C(R)*
- The set of functions for which $f'(x)$ is continuous on R is denoted by $C^1(R)$
- The set of functions for which $f''(x)$ is continuous on R is denoted by $C^2(R)$
- The set of functions for which $f^n(x)$ is continuous on R is denoted by $C^n(R)$
- $C^\infty(R)$ is the set of functions, each of whose derivatives is continuous.

$$\Rightarrow C^\infty(R)\subset\cdots\subset C^n(R)\subset\cdots\subset C^3(R)\subset C^2(R)\subset C^1(R)\subset C(R)$$

- $\Rightarrow C^n[a,b]$ To be the set of functions *f(x)* for which $f^{(n)}(x)$ exists and is continuous on the interval *[a,b]*.

Taylor's Theorem
Taylor's Theorem with Lagrange Remainder

-If $f(x) \in C^n[a,b]$ and $f^n(x)$ exists on *(a,b)*, then for any points *c* and *x* in *[a,b]*,

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

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

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

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

Example: Find the Taylor series of *f(x)=ln(x)* for *a=1, b=2,* and *c=1*.

Solution:

$$f'(x)=x^{-1},\ \ f^{(2)}=-x^{-2},\ \ f^{(3)}=2x^{-3},\ \ f^{(4)}=-6x^{-4},\ \ f^{(5)}=24x^{-5}$$

$$\Rightarrow f^{(k)}(x)=(-1)^{k-1}(k-1)!\,x^{-k}$$

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

$$E_n(x)=\frac{(-1)^{n}}{(n+1)}\xi^{-(n+1)}(x-1)^{n+1}\qquad\qquad(1<\xi<x)$$

$$|E_n(x)|=\frac{1}{(n+1)}\xi^{-(n+1)}(x-1)^{n+1}<\frac{(x-1)^{n+1}}{(n+1)}$$

Example: Assume that we want to compute *ln(2)* with the formulea given in the previous example. We want accuracy to be less than 10^{-8}. How many terms do we need to use?

Solution: $|E_n(x)| < \frac{(x-1)^{n+1}}{(n+1)} = \frac{(2-1)^{n+1}}{n+1}$

$$\Rightarrow 10^{-8} < \frac{1}{n+1} \Rightarrow (n+1) > 10^8 \Rightarrow n \geq 10^8 = 100 \quad million \quad terms.$$

Example: How many terms do we need to use to compute *ln(1.5)* with the same accuracy?

Solution:

$$\Rightarrow 10^{-8} < \frac{(1.5-1)^{n+1}}{n+1} = \frac{0.5^{n+1}}{n+1} \Rightarrow n \geq 22$$

Mean Value Theorem

If *f(x)* in *C[a,b]* and $f^{(1)}(x)$ exist on *(a,b)*, then for *x* and *c* in *[a,b]*

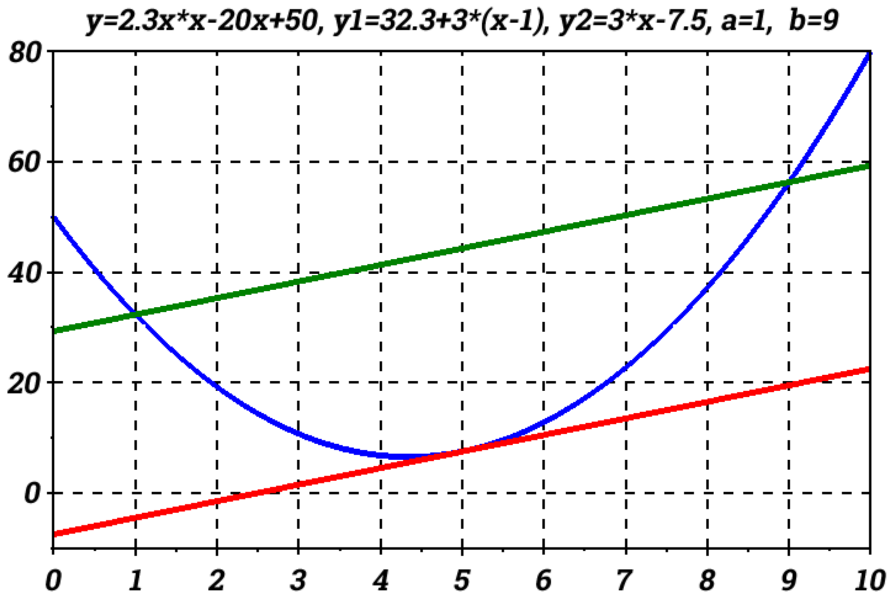
Zeroth order Taylor series expansion

$f(x)=f(c)+f'(\xi)(x-c)$ Where ξ is between *c* and *x*.

If x=b and c=a then

$f(b)-f(a)=f'(\xi)(b-a) \quad where \quad a<\xi<b$

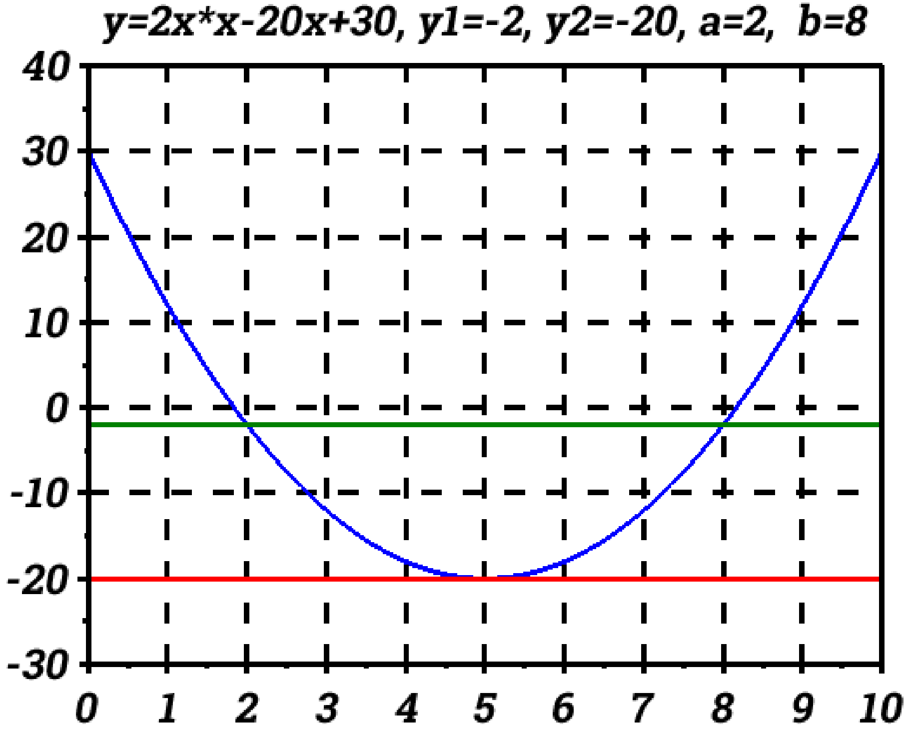
$$\Rightarrow f'(\xi)=\frac{f(b)-f(a)}{b-a}$$



f(a)=32.3, f(b)=56.3, (f(b)-f(a))/(b-a)=3

Rolle's Theorem

-If $f(x)$ in $C[a,b]$, if $f^{(1)}(x)$ exist on (a,b), and if $f(a)=f(b)$, then $f'(\xi)=0$ for some ξ in (a,b).



f(a)=f(b)=-2, (f(b)-f(a))/(b-a)=0

Taylor's Theorem with Integral Remainder

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

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

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

Proof: let $u = \frac{(x-t)^n}{n!}$ and $dv = f^{n+1}(t)\,dt \Rightarrow v = f^n(t)$ and $du = \frac{-(x-t)^{n-1}}{(n-1)!}dt$

$$\Rightarrow R_n(x) = \frac{1}{n!}\int_c^x f^{n+1}(t)(x-t)^n\,dt = \int_c^x u\,dv = uv\Big|_c^x - \int_c^x v\,du$$

$$\Rightarrow R_n(x) = \frac{(x-t)^n}{n!} f^n(t)\Big|_c^x + \frac{1}{(n-1)!}\int_c^x f^n(t)(x-t)^{n-1}\,dt = \frac{-(x-c)^n}{n!} f^n(c) + R_{n-1}(x)$$

$$\Rightarrow R_n(x) = \frac{-(x-c)^n}{n!} f^n(c) - \frac{(x-c)^{n-1}}{(n-1)!} f^{n-1}(c) + R_{n-2}(x)$$

If we repeat integration, we get

$$\Rightarrow R_n(x) = -\sum_{k=1}^{n} \frac{f^k(c)}{k!}(x-c)^k + R_0(x)$$

$$\Rightarrow R_0(x) = \int_c^x f'(t)\,dt = f(t)\Big|_c^x = f(x) - f(c)$$

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

Q.E.D.

Alternative Form of Taylor's Theorem

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

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

$$E_n(h) = \frac{1}{(n+1)!} f^{n+1}(\xi) h^{n+1}$$ where ξ lies between x and $(x+h)$.

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

$x+h \rightarrow x$ and $x \rightarrow c$ substitude $x+h$ for x and x for c

Taylor's Theorem in two Variables

If $f(x,y)\in C^{n+1}([a,b]x[c,d])$,then for any points $(x+h)$ and $(y+k)$ in $[a,b]x[c,d]\subseteq R^2$,

$$f(x+h,y+k)=\sum_{i=0}^{n}\frac{1}{i!}\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^i f(x,y)+E_n(h,k)\text{ where}$$

$$E_n(h,k)=\frac{1}{(n+1)!}\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^{n+1} f(x+\theta h,y+\theta k)+\text{ where } 0<\theta<1$$

The meaning of the terms: $\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^0 f(x,y)=f(x,y)$

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

$$\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^2 f(x,y)=\left(h^2\frac{\partial^2 f}{\partial x^2}+2hk\frac{\partial^2 f}{\partial x\partial y}+k^2\frac{\partial^2 f}{\partial y^2}\right)(x,y)\quad\text{and so on.}$$

1.2 Orders of Convergence and Additional Basic Concepts

-In numerical calculations, it is often happens that the answer to a problem is not produced all at once. Rather, a sequence of approximate answers is produced.

Convergent Sequences:

We write $\lim_{n\to\infty} x_n = L$ if there corresponds to each positive *ε* a real number *r* such that $|x_n - L| < \varepsilon$ whenever n>r. (Here n is an integer number.)

Example: Show that $\lim_{n\to\infty} \frac{n+1}{n} = 1$

Solution:

$$\lim_{n\to\infty} \frac{n+1}{n} = 1 \quad \text{because} \quad \left|\frac{n+1}{n} - 1\right| < \varepsilon \text{ whenever } n > \varepsilon^{-1} = r$$

Example: $e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$. If we compute the sequence $x_n = \left(1+\frac{1}{n}\right)^n$

$$x_1 = 2.000000, \quad x_{10} = 2.593742, \quad x_{30} = 2.674319, \quad x_{50} = 2.691588,$$
$$x_{1000} = 2.716924, \quad e = 2.7182818$$

This is an example of a sequence that is converging slowly. Using double-precision computations, we find numerical evidence that

$$\frac{|x_{n+1}-e|}{|x_n-e|} \rightarrow 1$$

This property is worse than linear convergence.

Example: $x_{n+1}=x_n-(x_n^2-2)\dfrac{x_n-x_{n-1}}{x_n^2-x_{n-1}^2} \rightarrow \sqrt{2}$ (converge to $\sqrt{2}$)

Let $x_1=2$, $x_2=1.5$, $\rightarrow x_3=1.428571$, $x_4=1.414634$,
$x_5=1.414244$, $x_6=1.414216$, $\sqrt{2}=1.414213562$

Using double-precision computations, we find numerical evidence that

$$\frac{|x_{n+1}-\sqrt{2}|}{|x_n-\sqrt{2}|} \leqslant 0.77$$

which is called superlinear convergence.

Example:

$$\left\{\begin{array}{ll} x_1 = 2, & \\ x_{n+1} = \dfrac{1}{2} x_n + \dfrac{1}{x_n} & n \geqslant 1 \end{array}\right\} \quad \text{converges to} \quad \sqrt{2}$$

→ $x_2 = 1.5, \quad x_3 = 1.416667, \quad x_4 = 1.414216, \quad \sqrt{2} = 1.414213562$

Using double-precision computations, we find numerical evidence that

$$\frac{|x_{n+1} - \sqrt{2}|}{|x_n - \sqrt{2}|^2} \leqslant 0.36 \quad \text{which is called quadratic convergence.}$$

Orders of Convergence

-Let $\lim_{n\to\infty} x_n = x^*$. We say that the rate of convergence is at least linear if there are constant *C<1* and an integer *N* such that

$$|x_{n+1}-x^*| \leqslant C|x_n-x^*| \qquad (n \geqslant N)$$

-We say that the rate of convergence is at least superlinear if there exist a sequence ε_n tending to *0* and an integer *N* such that

$$|x_{n+1}-x^*| \leqslant \varepsilon_n|x_n-x^*| \qquad (n \geqslant N)$$

-The rate of convergence is at least quadratic if there are constant *C* (not necessarily less than one) and an integer *N* such that

$$|x_{n+1}-x^*| \leqslant C|x_n-x^*|^2 \qquad (n \geqslant N)$$

-In general, if there are positive constant *C* and *α* and an integer *N* such that

$$|x_{n+1}-x^*| \leqslant C|x_n-x^*|^\alpha \qquad (n \geqslant N)$$

we say that the rate of convergence is of order *α* at least.

Big O and Little o Notation

Let x_n and α_n be two different sequences. We write $x_n = O(\alpha_n)$ if there are constants C and n_0 such that $|x_n| \leq C|\alpha_n|$ when $n \geq n_0$. Here we say that x_n is "Big oh" of α_n.

-The equation $x_n = o(\alpha_n)$ means that $\lim_{n \to \infty} (x_n / \alpha_n) = 0$. Here we say that x_n is "little oh" of α_n.

Example: $x_n = \frac{n+1}{n^2}$, $\alpha_n = \frac{1}{n}$ $\Rightarrow x_n = O(\alpha_n)$

Example: $x_n = \frac{1}{n \ln(n)}$, $\alpha_n = \frac{1}{n}$ $\Rightarrow x_n = o(\alpha_n)$

Example: $x_n = \frac{1}{n \ln(n)}$, $\alpha_n = \frac{1}{n}$ $\Rightarrow x_n = o(\alpha_n)$

Example: $x_n = \frac{5}{n} + e^{-n}$, $\alpha_n = \frac{1}{n}$ $\Rightarrow x_n = O(\alpha_n)$

Example: $x_n = e^{-n}\ ,\ \alpha_n = \frac{1}{n^2} \quad \Rightarrow x_n = o(\alpha_n)$

Example: $x_n = \ln(2) - \sum_{k=1}^{n-1} (-1)^{k-1} \frac{1}{k}\ ,\ \alpha_n = \frac{1}{n} \quad \Rightarrow x_n = O(\alpha_n)$

$$f(x) = \ln(x), \quad x=2, c=a=1, b=2 \quad \Rightarrow \quad x_n = E_{n-1}(x) = \frac{1}{n}(-1)^n \xi^{-n} (x-1)^n$$

$$\Rightarrow \quad x_n = E_{n-1}(2) = \frac{1}{n}(-1)^n \xi^{-n} \qquad 1 < \xi < 2$$

Example: $x_n = e^x - \sum_{k=0}^{n-1} \frac{1}{k!} x^k\ ,\ \alpha_n = \frac{1}{n!}, \qquad |x| \leqslant 1 \qquad \Rightarrow x_n = O(\alpha_n)$

$$f(x) = e^x,\ c=0, a=-1, b=1 \ \Rightarrow\ x_n = E_{n-1}(x) = \frac{1}{n!} e^{\xi} x^n \Rightarrow |x_n| \leq \frac{e}{n!} \qquad -1 \leq \xi \leq 1$$

-O and o notations can be used also for functions.

Example: $\sin(x) = x - \frac{x^3}{6} + O(x^5) \qquad (x \to 0)$

Means that $\left|\sin(x) - x + \frac{x^3}{6}\right| \le C|x^5|$ as $x \to 0$ $\qquad C$ is a psoitive constant

-An equation of the form $f(x) = O(g(x)) \quad (x \to \infty)$ means that there exist constants *r* and *C* so that $|f(x)| \le C|g(x)|$ whenever $x \ge r$

Example: $\sqrt{x^2+1} = O(x) \quad x \to \infty \qquad$ since $\sqrt{x^2+1} \le 2x$ when $x \ge 1$

-In general, we write $f(x) = O(g(x)) \quad (x \to x^*)$ when there is a positive constant *C* and a neighborhood of x^* such that $|f(x)| \le C|g(x)|$ in that neighborhood. Similarly, $f(x) = o(g(x)) \quad (x \to x^*)$ means that

$$\lim_{x \to x^*} [f(x)/g(x)] = 0$$

Mean Value Theorem for Integrals

Theorem: Let *u* and *v* be continuous real-valued functions on an interval *[a,b]*, and suppose that $v \geq 0$. Then there exists a point ξ in *[a,b]* such that

$$\int_a^b u(x)v(x)\,dx = u(\xi)\int_a^b v(x)\,dx = Iu(\xi)$$

Proof: Let α and β denote the least and greatest value of *u(x)* on *[a,b]*, respectively. Then

$$\alpha \leq u(x) \leq \beta \qquad (a \leq x \leq b) \qquad \Rightarrow \alpha \leq u(\xi) \leq \beta$$

since $v(x) \geq 0$ we have

$$\alpha v(x) \leq u(x)v(x) \leq \beta v(x) \qquad (a \leq x \leq b) \qquad \text{Let} \quad I = \int_a^b v(x)\,dx$$

$$\Rightarrow \int_a^b \alpha v(x)\,dx \leq \int_a^b u(x)v(x)\,dx \leq \int_a^b \beta v(x)\,dx$$

$$\Rightarrow \alpha I \leq \int_a^b u(x)v(x)\,dx \leq \beta I. \qquad \text{If } I = 0 \Rightarrow v(x) = 0. \quad \text{The result is trivial.}$$

If $I \neq 0 \quad \Rightarrow \alpha \leq I^{-1} \int_a^b u(x)v(x)\,dx \leq \beta$

By the intermediat-value theorem for continuous functions, there exists a point ξ in *[a,b]* for which

$$u(\xi) = I^{-1} \int_a^b u(x)v(x)\,dx$$

Chapter-2 Computer Arithmetic

2.1 Floating-Point Numbers and Roundoff Errors

-Most computers deal with real numbers in the binary system.

$$(427.325)_{10}=4*10^{2}+2*10^{1}+7*10^{0}+3*10^{-1}+2*10^{-2}+5*10^{-3}$$

$$(1001.11101)_{2}=1*2^{3}+0*2^{2}+0*2^{1}+1*2^{0}+1*2^{-1}+1*2^{-2}+1*2^{-3}+0*2^{-4}+1*2^{-5}$$

$$=(9.90625)_{10}$$

-The word length of the computer places restriction on the precision with which real numbers can be represented.

-Even 1/10 cannot be stored exactly in the computer.

$$1/10=(0.0001\,1001\,1001\,1001\,1001\,1001\ldots)_{2}$$

If we print 1/10 out to 40 decimal places, we obtain the following result:

$$0.1000000014901161193847656250000000000000$$

-There are two conversions: From decimal to binary, and from binary to decimal. Because of conversions, there will be errors.
-The product of two numbers that have eight digits to the right of the decimal point will be a number that has 16 digits to the right of the decimal point. (need rounding)

-Rounding to the nearest number

Consider a positive decimal number x of the form $0.a_1 a_2 a_3 \cdots a_{m-2} a_{m-1} a_m$ with m digits to the right of the decimal point. One rounds x to n decimal places (n<m) in a manner that depends on the value of (n+1)st digit. If the digit is less than 5, the digits after the *n*th decimal place, are discarded. Otherwise, *n*th digit is increased by one and the remaining digits are discarded.

Example: Seven-digit numbers are rounded to four-digit numbers:

$0.1735499 \rightarrow 0.1735$

$0.9999500 \rightarrow 1.0000$

$0.4321609 \rightarrow 0.4322$

If x is the number and $\tilde{x}$ is the rounded number, then

$|x-\tilde{x}| \leq \frac{1}{2} 10^{-n}$ where n is the number of decimal places after the decimal point for the rounded number.

Truncation: Discard all the digits beyond the *n*th digit.

If x: number , $\hat{x}$: truncated number $\Rightarrow |x-\hat{x}| < 10^{-n}$

Normalized Scientific Notation

$$732.5051 = 0.7325051 * 10^{3}$$

$$-0.005612 = -0.5612 * 10^{-2}$$

In general, a nonzero real number *x* can be represented in the form:

$x = \pm r\, 10^{n}$, $\frac{1}{10} \leq r < 1$, n is an integer number

For binary numbers:

$x = \pm q\,2^m,\quad \frac{1}{2} \le q < 1,\quad q$ is called mantissa and the integer m is the exponent.

A slightly different scientific notation:

$$q = (1.f)_2 \qquad \Rightarrow 1 \le q < 2$$

Hypothetical Computer Marc-32 (Same as float number in the programming languages such as C/C++)

Sign of mantissa 1 bit (s)	Biased exponent (e) 8 bits (unsigned integer)	Normalized Mantissa (f) 23 bits

imlicit radix point

$$q = (1.f)_2 \qquad \Rightarrow 1 \le q < 2$$

$$m = e - 127$$

$$x = (-1)^s q\,2^m$$ Normalized floating point form.

-Most real numbers are not precisely re-presentable within the Marc-32.

$$0 < e < (1111\ \ 1111)_2 = 2^8 - 1 = 255$$

e=0 and *e=255* are reserved for special cases such as $\pm 0,\ \ \pm\infty$ and NaN

$$m = e - 127 \qquad \Rightarrow -126 \le m \le 127$$

-Marc-32 can handle numbers as small as $2^{-126} \approx 1.2 * 10^{-38}$ and as large as $\left(2 - 2^{-23}\right) * 2^{127} \approx 3.4 * 10^{38}$

-Mantissa has 23 bits. Therefore, our machine numbers have a limited precision of roughly six decimal places, since the least significant bit in the mantissa represents unit of 2^{-23} (or approximately $1.2 * 10^{-7}$). Thus, numbers expressed with more than six decimal digits will be approximated when given as input to the computer.

Zero: $+0 = [0000\ \ 0000]_{16} \qquad -0 = [8000\ \ 0000]_{16}$

Infinity: $+\infty = [7F80\ \ 0000]_{16} \qquad -\infty = [FF80\ \ 0000]_{16}$

Not a Number (NaN): $0/0, \quad \infty - \infty, \quad x + NaN \qquad \Rightarrow e = 255, \quad f \neq 0$

Machine Rounding

-In addition to rounding input data, rounding is needed after most arithmetic operations.

-The default rounding mode is round to nearest. The maximum error is half a unit of rounding in the least significant place.

Nearby Machine Numbers

Let $x = q * 2^m, \quad 1 \le q < 2, \quad -126 \le m \le 127$. What is the machine number closest to *x*?

$x = (1.a_1 a_2 a_3 a_4 a_5 \ldots a_{23} a_{24} a_{25} \ldots)_2 * 2^m$ in which a_i is either 0 or 1.

-One nearby machine number is obtained by simply discarding the excess bits $a_{24} a_{25} a_{26}$. (called chopping)

$\Rightarrow \; x_- = (1.a_1 a_2 a_3 a_4 a_5 \ldots a_{21} a_{22} a_{23})_2 * 2^m$

-Another nearby machine number lies to the right of *x* (rounding up)

$\Rightarrow \; x_+ = ((1.a_1 a_2 a_3 a_4 a_5 \ldots a_{21} a_{22} a_{23})_2 + 2^{-23}) * 2^m$

$$|x-x_-|\le\frac{1}{2}|x_+-x_-|=\frac{1}{2}*2^{m-23}=2^{m-24}$$

The relative error is bounded as follows:

$$\left|\frac{x-x_-}{x}\right|\le\frac{2^{m-24}}{q\,2^m}=\frac{1}{q}2^{-24}\le 2^{-24}$$

-When *x* is closer to x_+ than to x_-

$$|x-x_+|\le\frac{1}{2}|x_+-x_-|=\frac{1}{2}*2^{m-23}=2^{m-24}\qquad\Rightarrow\left|\frac{x-x_+}{x}\right|\le\frac{2^{m-24}}{q\,2^m}=\frac{1}{q}2^{-24}\le 2^{-24}$$

The relative error cannot be greater than 2^{-24}

If *x* is a nonzero real number and x^* is the machine number (marc-32) closest to x, then

$$\left|\frac{x-x^*}{x}\right|\le 2^{-24},\quad \text{let }\ \delta=\left(\frac{x^*-x}{x}\right)\quad \text{and }\ fl(x)=x(1+\delta)\qquad\Rightarrow|\delta|\le 2^{-24}$$

-The notation *fl(x)* is used to denote the floating point machine number x^* closest to *x*.

The number 2^{-24} is called the unit roundoff error for the marc-32.

Relative Error Analysis

Theorem: Let $x_0 x_1 x_2 x_3 \ldots x_n$ be positive machine numbers in a computer whose unit roundoff error is ε. Then the relative roundoff error in computing $\sum_{i=0}^{n} x_i$ is at most $(1+\varepsilon)^n - 1 \simeq n\varepsilon$

2.2 Absolute and Relative Errors: Loss of Significance

The absolute error $= |x - x^*|$ where x^* is approximation of x

The relative error $= \left|\dfrac{x - x^*}{x}\right|$

Loss of Significance

-Large relative error can occur after substraction of two numbers that are close to each other.

Example: Let x=0.3721478693 and y=0.3720230572 (ten digit after the decimal point)

Do not represent additional accuracy

$x - y = 0.1248121000 * 10^{-3}$ (7 digits after the decimal point. 3 digits are lost.

If this calculation were to be performed in a decimal computer having a five-digit mantissa :

fl(x)=0.37215

fl(y)=0.37202

Three digits are lost.

fl(x)-fl(y)=0.00013= $0.13000 * 10^{-3}$

The relative error= $\left| \frac{x - y - [fl(x) - fl(y)]}{x - y} \right| = \left| \frac{0.0001248121 - 0.00013}{0.0001248121} \right| \simeq 4\%$

Subtraction of Nearly Equal Quantities

Example: $y=\sqrt{x^2+1}-1.$ Assume *x* is close to zero. It involves loss of significance for small values of *x*. How can we avoid this trouble?

Solution:

$$y=\sqrt{x^2+1}-1=\left(\sqrt{x^2+1}-1\right)\frac{\sqrt{x^2+1}+1}{\sqrt{x^2+1}+1}=\frac{x^2}{\sqrt{x^2+1}+1}$$

Loss of Precision

Theorem: If *x* and *y* are positive normalized floating-point binary machine numbers such that *x*>*y* and $2^{-q}\leq 1-\frac{y}{x}\leq 2^{-p}$. Then at most q and at least p significant binary bits are lost in the subtraction *x*-*y*.

Example: $y = x - \sin x$

Since $\sin x \simeq x$ for small values of x, this calculation will involve a loss of significance. How can this be avoided?

Solution: The Taylor series of $\sin x$ can be used.

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} \ldots$$

$$\Rightarrow y = x - \sin x = \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} - \frac{x^9}{9!} \ldots$$

We need to use only a couple of terms.

2.3 Stable and Unstable Computations: Conditioning

-A numerical process is unstable if small errors made at one stage of the process are magnified in subsequent stages and seriously degrade the accuracy of the overall calculation.

Example: $x_0 = 1, \quad x_1 = \frac{1}{3}, \qquad x_n = \frac{13}{3} x_{n-1} - \frac{4}{3} x_{n-2} \qquad n \geq 1$

In fact $\quad x_n = \frac{1}{3^n}$

Proof (by induction): It is true for n=0 and n=1.

Assume it is true for $n \le m$. Lets show that it is true for n=m+1.

$$\text{If} \quad n=m \quad \Rightarrow x_m = \frac{13}{3} x_{m-1} - \frac{4}{3} x_{m-2} = \frac{1}{3^m}$$

$$\text{If} \quad n=m+1 \quad \Rightarrow x_{m+1} = \frac{13}{3} x_m - \frac{4}{3} x_{m-1} = \frac{13}{3}\frac{1}{3^m} - \frac{4}{3}\frac{1}{3^{m-1}} = \frac{1}{3^{m-1}}\left[\frac{13}{9} - \frac{4}{3}\right] = \frac{1}{3^{m+1}}$$

Lets compute $x_n = \frac{13}{3} x_{n-1} - \frac{4}{3} x_{n-2}$ using marc-32 computer (float)

$x_0 = 1.0000000$

$x_1 = 0.3333333$ (seven correctly rounded significant digit)

$x_2 = 0.1111112$ (six correctly rounded significant digit)

$x_3 = 0.00370375$ (five correctly rounded significant digit)

$\vdots$

$x_7 = 0.0005131$ (one correctly rounded significant digit)

$x_{14} = 0.9143735$

$x_{15} = 3.657493$ incorrect with relative error 10^8

This algorithm is unstable. Any error present in x_n is multiplied by 13/3 in computing x_{n+1}. Error in x_1 may propagate in x_{15} with a factor $(13/3)^{14}$

-**The general solution of the previous equation is**

$$x_n = A\left(\frac{1}{3}\right)^n + B\,4^n$$

Example:

If $x_0 = 1, \quad x_1 = 4 \quad \text{then} \quad x_n = 4^n \quad$ (correct solution)

If we compute x_n using marc-32

$x_1 = 4.000006, \qquad x_{10} = 1.048576 * 10^{6,} \qquad x_{20} = 1.099512 * 10^{12}$

-The absolute errors are undoubtedly large as before but they are relatively negligible.

Example: (Numeric instability) Compute y_n

$$y_0=\int_0^1 e^x\,dx=e-1$$

$$y_n=\int_0^1 x^n e^x\,dx \qquad n\geq 0$$

Apply integration by part

$$y_{n+1}=\int_0^1 x^{n+1}e^x\,dx \qquad \text{Let } u(x)=x^{n+1}, \quad e^x\,dx=dv \qquad \Rightarrow v=e^x, \quad du=(n+1)x^n\,dx.$$

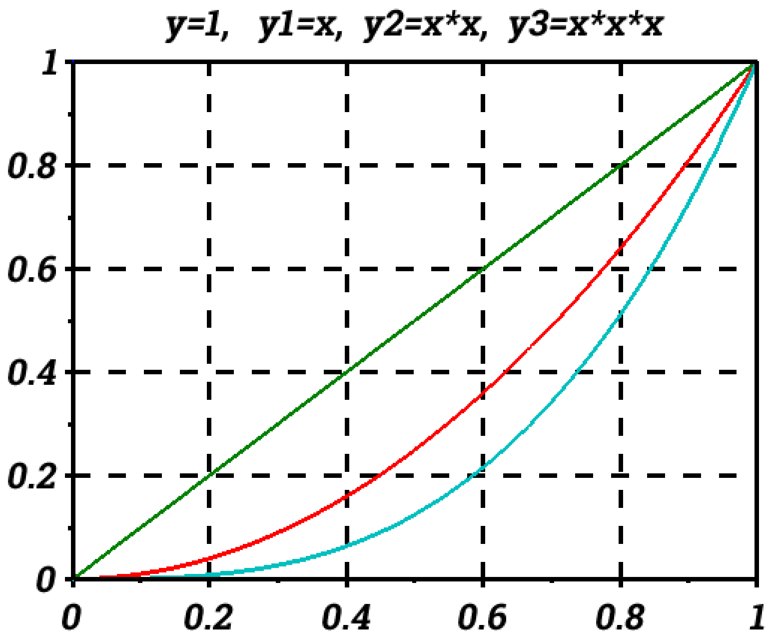
$$\Rightarrow y_{n+1}=x^{n+1}e^x|_0^1-(n+1)\int_0^1 x^n e^x\,dx=e-(n+1)y_n$$

$\Rightarrow y_1=1, \quad y_2=0.7182817, \quad y_{11}=1.422453,$

$y_{15}=39711.43$

in fact $\quad y_1>y_2>y_3>\ldots>0$

The error at nth step is multiplied by (n+1) in computing y_{n+1}.



Conditioning

-A problem is ill conditioned if small changes in the data can produce large changes in the answer.

-For certain types of problems, a condition number can be defined.

-Condition Number for *f(x)*.

-If *x* is perturbed slightly, what is the effect on *f(x)*?

$$\underbrace{f(x+h)-f(x)}_{\text{Error}} = hf'(\xi) \simeq \underbrace{h}_{\text{Change in } x.}\,\underbrace{f'(x)}_{\text{Condition number for this problem}} \qquad \text{(mean value theorem)}$$

Sometimes, the relative error is important.

$$\underbrace{\frac{f(x+h)-f(x)}{f(x)}}_{\text{Relative change in } f(x).} \simeq \frac{hf'(x)}{f(x)} = \overbrace{\left[\frac{xf'(x)}{f(x)}\right]}^{\text{Condition number for this problem}}\underbrace{\left(\frac{h}{x}\right)}_{\text{Relative change in } x.}$$

Example: Let *f(x)=arcsin(x)*. Find a condition number for *f(x).* (Relative change in *f(x)*.)

Solution:

$$\text{Condition Number} = \frac{xf'(x)}{f(x)} = \frac{x}{\sqrt{1-x^2}\arcsin x}$$

For x near 1 $\arcsin x \simeq \frac{\pi}{2}$ $\text{condition number} \to \infty$ Hence, small relative error in *x* may lead to large relative errors in *arcsin(x)* near *x=1*.

Chapter-3: Solution of Nonlinear Equations

-We want to find *x* such that *f(x)=0.*

Example: *f(x)=x-tan(x)=0*

Example: *x-a*sin(x)=b*

-There may be many approximate solutions even yhough the exact solution is unique. (Because of roundoff errors.)

Example: $P_4(x)=x^4-4x^3+6x^2-4x+1=(x-1)^4$

If you use marc-32, you will find many zeros in the interval [0.975, 1.035]

3-1 Bisection (Interval Halving) Method

-If f(x) is a continuous function on the interval [a,b] and if f(a)f(b)<0, then f(x) must have a zero in (a,b).

Bisection Method:

1- Compute c=0.5*(a+b)

2- If f(a)f(c)<0 then f(x) has a zero in [a,c] $\Rightarrow$ $b \leftarrow c$ (assign c to b)

3- Else $\Rightarrow$ $a \leftarrow c$ (assign c to a $f(x)$ has a zero in $[c, b]$)

4- Stop if *f(c)=0*. (*c* is the zero of *f(x)*)

5- Go to step 1.

-It is quite unlikely that f(c) will be exactly 0 in the computer because of roundoff errors.

Stopping Criteria (stop when one of them is satisfied)

1-The maximum number of steps (M)

2- $|f(c)| < \epsilon$ ϵ is a positive number.

3- $|b-c| < \delta$ δ is a positive number.