Preliminaries

Peng Yu

Tel: 0755 8801 8911
Email: yup6@sustech.edu.cn

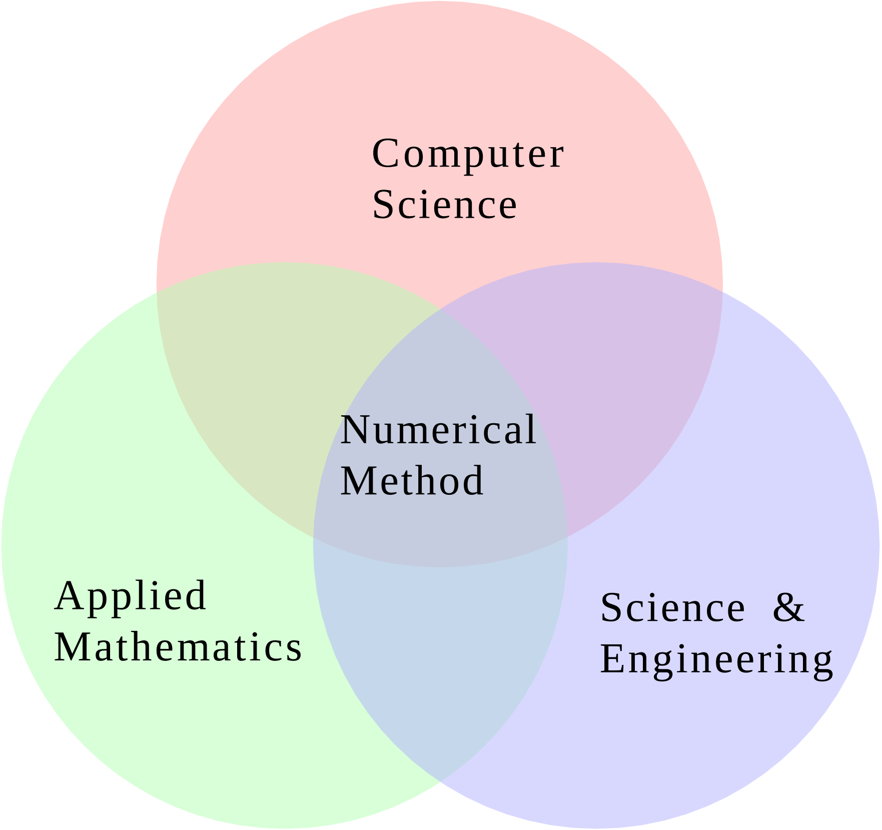
Preliminaries

- Numerical Method
- Review of Calculus
- Binary Numbers
- Error Analysis

Numerical Method

Numerical Method

- Design and analysis of algorithms for solving **mathematical problems** arising in **science and engineering** **numerically**:



- Also called numerical analysis, scientific computing, or computational mathematics

Numerical Method

- Distinguishing features of numerical method
  - ✓Deals with *continuous* quantities (e.g., time, distance, velocity, temperature, density, pressure) typically measured by real numbers
  - ✓Considers effects of *approximations*

- Why numerical method?
  - ✓Predictive simulation of natural phenomena
  - ✓Virtual prototyping of engineering designs
  - ✓Analyzing data

Mathematical Problems

- Given mathematical relationship $y = f(x)$, typical problems include
  - ✓Evaluate a function: compute output y for given input x
  - ✓Solve an equation: find input x that produces given output y
  - ✓Optimize: find x that yields extreme value of y over given domain

- Specific type of problem and best approach to solving it depend on whether variables and function involved are
  - ✓discrete or continuous
  - ✓linear or nonlinear
  - ✓finite or infinite dimensional
  - ✓purely algebraic or involve derivatives or integrals

General Problem-Solving Strategy

- Replace difficult problem by easier one having same or closely related solution
  - ✓infinite dimensional → finite dimensional
  - ✓differential → algebraic
  - ✓nonlinear → linear
  - ✓complicated → simple
- Solution obtained may only **approximate** that of original problem
- Our goal is to estimate accuracy and ensure that it suffices

Review of Calculus

Limits and Continuity

Definition 1.1. Assume that *f(x)* is defined on an open interval containing $x = x_0$, except possibly at $x = x_0$ itself. Then *f* is said to have the ***limit L*** at $x = x_0$, and we write

$$\lim_{x \to x_0} f(x) = L, \tag{1}$$

if given any $\epsilon > 0$ there exists a $\delta > 0$ such that $|f(x) - L| < \epsilon$ whenever $0 < |x - x_0| < \delta$. When the h-increment notation $x = x_0 + h$ is used, equation (1) becomes

$$\lim_{h \to 0} f(x_0 + h) = L. \tag{2}$$

Limits and Continuity

Definition 1.2. Assume that $f(x)$ is defined on an open interval containing $x = x_0$, then f is said to be continuous at $x = x_0$ if

(3) $$\lim_{x \to x_0} f(x) = f(x_0).$$

The function f is said to be continuous on a set S if it is continuous at each point $x \in S$.

The notation $C^n(S)$ stands for the set of all functions f such that f and its first n derivatives are continuous on S.

When S is an interval, say $[a, b]$, then the notation $C^n[a, b]$ is used.

Limits and Continuity

Definition 1.3. Suppose that $\{x_n\}_{n=1}^{\infty}$ is an infinite sequence. Then the sequence is said to have the $limit\ L$, and we write

(4) $$\lim_{n\to\infty} x_n = L,$$

if given any $\epsilon > 0$, there exists a positive integer $N = N(\epsilon)$ such that $n > N$ implies that $|x_n - L| < \epsilon$.

When a sequence has a limit, we say that it is a $\boldsymbol{convergent\ sequence.}$ Another commonly used notation is "$x_n \to L$ as $n \to \infty$.” Equation (4) is equivalent to

(5) $$\lim_{n\to\infty} (x_n - L) = 0.$$

Thus we can view the sequence $\{\epsilon_n\}_{n=1}^{\infty} = \{x_n - L\}_{n=1}^{\infty}$ as an $\boldsymbol{error\ sequence.}$ The following theorem relates the concepts of continuity and convergent sequence.

Limits and Continuity

Theorem 1.1. Assume that $f(x)$ is defined on the set S and $x_0 \in S$. The following statements are equivalent:

(6) (a) The function f is continuous at x_0.
(b) If $\lim_{n\to\infty} x_n = x_0$, then $\lim_{n\to\infty} f(x_n) = f(x_0)$.

Theorem 1.2. (Intermediate Value Theorem). Assume that $f \in C[a, b]$ and L is any number between $f(a)$ and $f(b)$. Then there exists a number c, with $c \in (a, b)$, such that $f(c) = L$.

Theorem 1.3. (Extreme Value Theorem for a Continuous Function). Assume that $f \in C[a, b]$. Then there exists a lower bound M_1, an upper bound M_2, and two numbers $x_1, x_2 \in [a, b]$ such that

(7) $$M_1 = f(x_1) \le f(x) \le f(x_2) = M_2 \text{ whenever } x \in [a, b].$$

We sometimes express this by writing

(8) $$M_1 = f(x_1) = \min_{a \le x \le b}\{f(x)\} \text{ and } M_2 = f(x_2) = \max_{a \le x \le b}\{f(x)\}.$$

Differentiable Functions

Definition 1.4. Assume that $f(x)$ is defined on an open interval containing x_0. Then *f* is said to be *differentiable* at x_0 if

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

exists. When this limit exists, it is denoted by $f'(x_0)$ and is called the ***derivative*** of f at x_0. An equivalent way to express this limit is to use the *h*-increment notation:

(10) $$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} = f'(x_0).$$

A function that has a derivative at each point in a set *S* is said to be ***differentiable*** on *S*.

Theorem 1.4. If $f(x)$ is differentiable at $x = x_0$, then $f(x)$ is continuous at $x = x_0$.

It follows from Theorem 1.3 that if a function *f* is differentiable in a closed interval [a, b], then its extreme values occur at the endpoints of the interval or at the critical points (solution of $f'(x) = 0$) in the open interval (a, b).

Differentiable Functions

Theorem 1.5 (Rolle's Theorem). Assume that $f \in C[a,b]$ and that $f'(x)$ exists for all $x \in (a,b)$. If $f(a) = f(b)$, then there exists a number c, with $c \in (a,b)$, such that $f'(c) = 0$.

Theorem 1.6 (Mean Value Theorem). Assume that $f \in C[a,b]$ and that $f'(x)$ exists for all $x \in (a,b)$. Then there exists a number c, with $c \in (a,b)$, such that

$$f'(c) = \frac{f(b)-f(a)}{b-a}. \tag{11}$$

Geometrically, the mean value theorem says that there is at least one number $c \in (a, b)$ such that the slope of the tangent line to the graph of $y = f(x)$ at the point $(c, f(c))$ equals the slope of the secant line through the points $(a, f(a))$ and $(b, f(b))$.

Theorem 1.7 (Generalized Rolle's Theorem). Assume that $f \in C[a,b]$ and that $f'(x), f''(x), \ldots, f^{(n)}(x)$ exists over (a, b) and $x_0, x_1, \ldots, x_n \in [a,b]$. If $f\left(x_j\right) = 0$ for $j = 0,1,\ldots,n$, then there exists a number c, with $c \in (a,b)$, such that $f^{(n)}(c) = 0$.

Integrals

Theorem 1.8 (First Fundamental Theorem). If f is continuous over $[a, b]$ and F is any antiderivative of f on $[a, b]$, then

(12) $$\int_a^b f(x)\,dx = F(b) - F(a) \text{ where } F'(x) = f(x).$$

Theorem 1.9 (Second Fundamental Theorem). If f is continuous over $[a, b]$ and $x \in (a, b)$, then

(13) $$\frac{d}{dx}\int_a^x f(t)\,dt = f(x).$$

Integrals

Theorem 1.10 (Mean Value Theorem for Integrals). Assume that $f \in C[a,b]$. Then there exists a number c, with $c \in (a,b)$, such that

$$\frac{1}{b-a}\int_a^b f(x)\,dx = f(c).$$

The value $f(c)$ is the average value of f over the interval [a, b].

Theorem 1.11 (Weighted Integral Mean Value Theorem). Assume that $f, g \in C[a,b]$ and $g(x) \geq 0$ for $x \in (a,b)$. Then there exists a number c, with $c \in (a,b)$, such that

$$\int_a^b f(x)g(x)\,dx = f(c)\int_a^b g(x)\,dx. \tag{14}$$

Series

Definition 1.5. Let $\{a_n\}_{n=1}^{\infty}$ be a sequence. Then $\sum_{n=1}^{\infty} a_n$ is an infinite series. The nth partial sum is $S_n = \sum_{k=1}^{n} a_k$. The infinite series ***converges*** if and only if the sequence $\{S_n\}_{n=1}^{\infty}$ converges to a limit S,

$$\lim_{n\to\infty} S_n = \lim_{n\to\infty} \sum_{k=1}^{n} a_k = S.$$

If a series does not converge, we say that it ***diverges***.

Example

Consider the infinite sequence $\{a_n\}_{n=1}^{\infty} = \left\{\frac{1}{n(n+1)}\right\}_{n=1}^{\infty}$. Then the nth partial sum is

$$S_n = \sum_{k=1}^{n} \frac{1}{k(k+1)} = \sum_{k=1}^{n} \left(\frac{1}{k} - \frac{1}{k+1}\right) = 1 - \frac{1}{n+1}$$

Therefore, the *sum* of the infinite series is

$$S = \lim_{n\to\infty} S_n = \lim_{n\to\infty} \left(1 - \frac{1}{n+1}\right) = 1$$

Series

Theorem 1.12 (Taylor's Theorem). Assume that $f \in C^{n+1}[a,b]$ and let $x_0 \in [a,b]$. Then, for every $x \in (a,b)$, there exists a number $c = c(x)$ (the value of c depends on the value of x) that lies between x_0 and x such that

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

where

(17) $$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

and

(18) $$R_n(x) = \frac{f^{n+1}(c)}{(n+1)!}(x - x_0)^{n+1}$$

Corollary 1.1. If $P^n(x)$ is the Taylor polynomial of degree n given in Theorem 1.12, then

(19) $$P_n^{(k)}(x_0) = f^{(k)}(x_0) \quad \text{for } k = 0, 1, \ldots, n.$$

Binary Numbers

- Decimal number system (base 10)
- Binary number system (base 2)
- Computer converts inputs to base 2 (or perhaps base 16), then performs base 2 arithmetic, and finally, translates the answer into base 10 before it displays a result.

$$\sum_{k=1}^{100,000} 0.1 = 9999.99447\,.$$

Base 2 Numbers

$$\begin{aligned}1563 = &(1 \times 2^{10}) + (1 \times 2^{9}) + (0 \times 2^{8}) + (0 \times 2^{7}) + (0 \times 2^{6}) \\ &+ (0 \times 2^{5}) + (1 \times 2^{4}) + (1 \times 2^{3}) + (0 \times 2^{2}) + (1 \times 2^{1}) \\ &+ (1 \times 2^{0})\end{aligned}$$

$$1563 = 11000011011_{two}$$

$$\begin{aligned}
1563 &= 2 \times 781 + 1, & b_0 &= 1 \\
781 &= 2 \times 390 + 1, & b_1 &= 1 \\
390 &= 2 \times 195 + 0, & b_2 &= 0 \\
195 &= 2 \times 97 + 1, & b_3 &= 1 \\
97 &= 2 \times 48 + 1, & b_4 &= 1 \\
48 &= 2 \times 24 + 0, & b_5 &= 0 \\
24 &= 2 \times 12 + 0, & b_6 &= 0 \\
12 &= 2 \times 6 + 0, & b_7 &= 0 \\
6 &= 2 \times 3 + 0, & b_8 &= 0 \\
3 &= 2 \times 1 + 1, & b_9 &= 1 \\
1 &= 2 \times 0 + 1, & b_{10} &= 1
\end{aligned}$$

Binary Fractions

$$R = (d_1 \times 2^{-1}) + (d_2 \times 2^{-2}) + \cdots + (d_n \times 2^{-n}) + \cdots,$$

$$\Downarrow$$

$$R = 0.d_1 d_2 \cdots d_n \cdots_{two}.$$

Example:

$$\frac{7}{10} = 0.1\overline{0110}_{two}.$$

$$\begin{array}{lll}
2R = 1.4 & d_1 = \mathrm{int}(1.4) = 1 & F_1 = \mathrm{frac}(1.4) = 0.4 \\
2F_1 = 0.8 & d_2 = \mathrm{int}(0.8) = 0 & F_2 = \mathrm{frac}(0.8) = 0.8 \\
2F_2 = 1.6 & d_3 = \mathrm{int}(1.6) = 1 & F_3 = \mathrm{frac}(1.6) = 0.6 \\
2F_3 = 1.2 & d_4 = \mathrm{int}(1.2) = 1 & F_4 = \mathrm{frac}(1.2) = 0.2 \\
2F_4 = 0.4 & d_5 = \mathrm{int}(0.4) = 0 & F_5 = \mathrm{frac}(0.4) = 0.4 \\
2F_5 = 0.8 & d_6 = \mathrm{int}(0.8) = 0 & F_6 = \mathrm{frac}(0.8) = 0.8 \\
2F_6 = 1.6 & d_7 = \mathrm{int}(1.6) = 1 & F_7 = \mathrm{frac}(1.6) = 0.6
\end{array}$$

Binary Shifting

If a rational number that is equivalent to an infinite repeating binary expansion is to be found, then a shift in the digits can be helpful.

Example

(23) $$S = 0.00000\overline{11000}_{two}.$$

Multiplying both sides of (23) by 2^5 will shift the binary point five places to the right, and $32S$ has the form

(24) $$32S = 0.\overline{11000}_{two}$$

Similarly, multiplying both sides of (23) by 2^{10} will shift the binary point 10 places to the right, and $1024S$ has the form

(25) $$1024S = 11000.\overline{11000}_{two}.$$

The result of taking the difference between the left-and right-hand sides of (24) and (25) is $992S = 11000_{two}$ or $992S = 24$,. Therefore,

$$\boldsymbol{S = 3/124}$$

Scientific Notation

A standard way to present a real number, called scientific notation, is obtained by shifting the decimal point and supplying an appropriate power of 10. For example

$$0.0000747 = 7.47 \times 10^{-5},$$

$$31.4159265 = 3.14159265 \times 10,$$

$$9{,}700{,}000{,}000 = 9.7 \times 10^{9}.$$

In computer science, $1\text{K} = 1.024 \times 10^{3}$

Machine Numbers

- Computers use a normalized floating-point binary representation for real numbers.
- This means that the mathematical quantity x is not actually stored in the computer.
- Computer stores a binary approximation to x

$$x \approx \pm q \times 2^{n}.$$

where q is the ***mantissa*** and it is a finite binary expression satisfying the inequality $1/2 \leq q < 1$, n is the ***exponent.***

- In a computer, only a small subset of the real number system is used.

Machine Numbers

- The number of binary digits is restricted in both the numbers q and n.
- An example: $0.d_1d_2d_3d_{4_{two}} \times 2^n$, where $d_1 = 1$ and d_2, d_3, and d_4 are either 0 or 1, and $n \in \{-3, -2, -1, 0, 1, 2, 3, 4\}$.There are eight choices for the mantissa and eight choices for the exponent, and this produces a set of 64 numbers:

Table 1.3 Decimal Equivalents for a Set of Binary Numbers with 4-Bit Mantissa and Exponent of $n = -3, -2, \ldots, 3, 4$

	Exponent							
Mantissa	$n = -3$	$n = -2$	$n = -1$	$n = 0$	$n = 1$	$n = 2$	$n = 3$	$n = 4$
0.1000_{two}	0.0625	0.125	0.25	0.5	1	2	4	8
0.1001_{two}	0.0703125	0.140625	0.28125	0.5625	1.125	2.25	4.5	9
0.1010_{two}	0.078125	0.15625	0.3125	0.625	1.25	2.5	5	10
0.1011_{two}	0.0859375	0.171875	0.34375	0.6875	1.375	2.75	5.5	11
0.1100_{two}	0.09375	0.1875	0.375	0.75	1.5	3	6	12
0.1101_{two}	0.1015625	0.203125	0.40625	0.8125	1.625	3.25	6.5	13
0.1110_{two}	0.109375	0.21875	0.4375	0.875	1.75	3.5	7	14
0.1111_{two}	0.1171875	0.234375	0.46875	0.9375	1.875	3.75	7.5	15

Machine Numbers

- What would happen if a computer had only a 4-bit mantissa and was restricted to perform the computation $\left(\frac{1}{10}+\frac{1}{5}\right)+\frac{1}{6}$?

$$
\begin{array}{rcccl}
 & \frac{1}{10} & \approx 0.1101_{\text{two}} \times 2^{-3} & = & 0.01101_{\text{two}} \times 2^{-2} \\
+ & \frac{1}{5} & \approx 0.1101_{\text{two}} \times 2^{-2} & = & 0.1101_{\text{two}} \times 2^{-2} \\
\hline
 & \frac{3}{10} & & & 1.00111_{\text{two}} \times 2^{-2}.
\end{array}
$$

- The computer must decide how to store the number $1.00111_{\text{two}} \times 2^{-2}$. Assume that it is rounded to $0.1010_{\text{two}} \times 2^{-1}$.

$$
\begin{array}{rcccl}
 & \frac{3}{10} & \approx 0.1010_{\text{two}} \times 2^{-1} & = & 0.1010_{\text{two}} \times 2^{-1} \\
+ & \frac{1}{6} & \approx 0.1011_{\text{two}} \times 2^{-2} & = & 0.01011_{\text{two}} \times 2^{-1} \\
\hline
 & \frac{7}{15} & & & 0.11111_{\text{two}} \times 2^{-1}.
\end{array}
$$

$$\mathbf{\frac{7}{15} \approx 0.1000_{\boldsymbol{two}} \times 2^{0}}$$

- Error

$$\frac{7}{15} - 0.1000_{two} \approx 0.466667 - 0.500000 \approx 0.033333$$

Computer Accuracy

- To store numbers accurately, computers must have floating-point binary numbers with at least 24 binary bits used for the mantissa (seven decimal places);

- A 32-bit mantissa can result in numbers with nine decimal places.

- Suppose that the mantissa q contains 32 binary bits. The condition $1/2 \leq q < 1$ implies that the first digit is $d_1 = 1$. Hence q has the form

$$q = 0.1d_2d_3 \cdot \cdot \cdot d_{31}d_{32\text{two}}$$

- An example: the mantissa contains 32 binary bits,

$$\frac{1}{10} \approx 0.11001100110011001100110011001100_{two} \times 2^{-3}.$$

- Compared with 1/10, the error is

$$0.\overline{1100}_{\text{two}} \times 2^{-35} \approx 2.328306437 \times 10^{-11}.$$

Computer Floating-Point Numbers

- Computers have both an ***integer mode*** and a ***floating-point mode*** for representing numbers.

- Computers that use 32 bits to represent single-precision real numbers use 8 bits for the exponent and 24 bits for the mantissa. Represent real numbers with magnitudes in the range 2.938736E-39 to 1.701412E+38, with six decimal digits of numerical precision.

- Computers that use 48 bits to represent single-precision real numbers might use 8 bits for the exponent and 40 bits for the mantissa. Represent real numbers from 2.9387358771E-39 to 1.7014118346E+38, with 11 decimal digits of precision.

- For 64-bit double-precision real numbers, it might use 11 bits for the exponent and 53 bits for the mantissa, represents number from 5.562684646268003E-309 to 8.988465674311580E+307, with 16 decimal digits of precision.

Error Analysis

Source of Errors

- Before computation
  - ✓modeling
  - ✓empirical measurements
  - ✓previous computations
- During computation
  - ✓truncation or discretization (mathematical approximations)
  - ✓rounding (arithmetic approximations)
- Accuracy of final result reflects all of these
- Uncertainty in input may be amplified by problem
- Perturbations during computation may be amplified by algorithm

Source of Errors

- Example

Computing surface area of Earth using formula $A = 4\pi r^2$ involves several errors

1. Earth is modeled as a sphere, idealizing its true shape
2. Value for radius is based on empirical measurements and previous computations
3. Value for π requires truncating infinite process
4. Values for input data and results of arithmetic operations are rounded by calculator or computer

Absolute Error and Relative Error

Definition 1.7. Suppose that $\hat{p}$ is an approximation to p. The ***absolute error*** is $E_p = |p - \hat{p}|$, and the ***relative error*** is $R_p = |p - \hat{p}|/|p|$, provided that $p \neq 0$.

The absolute error is simply the difference between the true value and the approximate value, whereas the relative error expresses the error as a percentage of the true value.

Relative error is preferred for floating-point representations since it deals directly with the mantissa.

Let $x = 3.141592$ and $x = 3.14$; then the errors are

$$E_x = |x - \widehat{x}| = |3.141592 - 3.14| = 0.001592,$$

$$R_x = \frac{|x - \widehat{x}|}{|x|} = \frac{0.001592}{3.141592} = 0.00507.$$

Let $y = 1{,}000{,}000$ and $y = 999{,}996$; then the errors are

$$E_y = |y - \widehat{y}| = |1{,}000{,}000 - 999, 996| = 4,$$

$$R_y = \frac{|y - \widehat{y}|}{|y|} = \frac{4}{1{,}000{,}000} = 0.000004.$$

Let $z = 0.000012$ and $z = 0.000009$; then the error is

$$E_z = |z - \widehat{z}| = |0.000012 - 0.000009| = 0.000003,$$

$$R_z = \frac{|z - \widehat{z}|}{|z|} = \frac{0.000003}{0.000012} = 0.25.$$

Absolute Error and Relative Error

Definition 1.8. The number $\hat{p}$ is said to ***approximate*** p to d significant digits if d is the largest nonnegative integer for which

$$\frac{|p - \hat{p}|}{|p|} < \frac{10^{1-d}}{2}. \tag{2}$$

Example

If $x = 3.141592$ and $\hat{x} = 3.14$, then $|x - \hat{x}|/|x| = 0.000507 < 10^{-2}/2$. Therefore, $\hat{x}$ approximates x to three significant digits.

If $y = 1{,}000{,}000$ and $\hat{y} = 999{,}996$, then $|y - \hat{y}|/|y| = 0.000004 < 10^{-5}/2$. Therefore, $\hat{y}$ approximates y to six significant digits.

If $z = 0.000012$ and $\hat{z} = 0.000009$, then $|z - \hat{z}|/|z| = 0.25 < 10^{-0}/2$. Therefore, $\hat{z}$ approximates z to one significant digit.

Truncation Error and Round-off Error

Truncation error:

Errors introduced when a more complicated mathematical expression is "replaced" with a more elementary formula. This terminology originates from the technique of replacing a complicated function with a truncated Taylor series.

For example, the infinite Taylor series

$$e^{x^2} = 1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!} + \cdots + \frac{x^{2n}}{n!} + \cdots$$

might be replaced with just the first five terms $1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!}$. This might be done when approximating an integral numerically.

Round-off Error

A computer's representation of real numbers is limited to the fixed precision of the mantissa. True values are sometimes not stored exactly by a computer's representation. This is called $\boldsymbol{round-off\ \ error.}$ In the preceding section the real number $1/10 = 0.0\overline{0011}_{two}$ was truncated when it was stored in a computer may undergo chopping or rounding the last digit.

Loss of Significance

- Consider p = 3.1415926536 and q = 3.1415957341, which are nearly equal and both carry 11 decimal digits of precision. Their difference is formed: $p - q = -0.0000030805$. Since the first six digits of p and q are the same, their difference $p - q$ contains only five decimal digits of precision. This phenomenon is called **loss of significance** or **subtractive cancellation**. This reduction in the precision of the final computed answer can creep in when it is not suspected.

- Example $f(x) = x\left(\sqrt{x+1} - \sqrt{x}\right)$ and $g(x) = \frac{x}{\sqrt{x+1}+\sqrt{x}}$.

$$f(500) = 500\left(\sqrt{501} - \sqrt{500}\right)$$
$$= 500(22.3830 - 22.3607) = 500(0.0223) = 11.1500.$$

$$g(500) = \frac{500}{\sqrt{501}+\sqrt{500}}$$

$$= \frac{500}{22.3830+22.3607} = \frac{500}{44.7437} = 11.1748$$

True: 11.174755300747198..

$O(h^n)$ Order of Approximation

- Sequences $\left\{\frac{1}{n^2}\right\}_{n=1}^{\infty}$ and $\left\{\frac{1}{n}\right\}_{n=1}^{\infty}$ are both converging to zero; which sequence is converging to zero more rapidly?

Definition 1.9. The function $f(h)$ is said to be $\boldsymbol{big\ Oh}$ of $g(h)$, denoted $f(h) = \boldsymbol{O}(g(h))$, if there exist constants C and c such that

$$|f(h)| \leq C|g(h)| \qquad \text{whenever } c \leq h.$$

Example Consider the functions $f(x) = x^2 + 1$ and $g(x) = x^3$. Since $x^2 \leq x^3$ and $1 \leq x^3$ for $x \geq 1$, it follows that $x^2 + 1 \leq 2x^3$ for $x \geq 1$. Therefore, $f(x) = \boldsymbol{O}(g(x))$.

Definition 1.10. Let $\{x_n\}_{n=1}^{\infty}$ and $\{y_n\}_{n=1}^{\infty}$ be two sequences. The sequence $\{x_n\}$ is said to be of order big Oh of $\{y_n\}$, denoted $x_n = \boldsymbol{O}(y_n)$, if there exists constants C and N such that

$$|x_n| \leq C|y_n| \qquad \text{whenever } n \geq N.$$

Example $\frac{n^2-1}{n^3} = \boldsymbol{O}(\frac{1}{n})$, since $\frac{n^2-1}{n^3} \leq \frac{n^2}{n^3} = \frac{1}{n}$ whenever $n \geq 1$.

$O(h^n)$ Order of Approximation

Definition 1.11. Assume that $f(h)$ is approximated by the function $p(h)$ and that there exists a real constant $M > 0$ and a positive integer n so that

$$\frac{|f(h)-p(h)|}{|h^n|} \leq M \qquad \text{for sufficiently small } h.$$

We say that $p(h)$ ***approximates*** $f(h)$ with order of approximation $O(h^n)$ and write

$$f(h) = p(h) + O(h^n).$$

Theorem 1.15. Assume that $f(h) = p(h) + O(h^n)$, $g(h) = q(h) + O(h^m)$, and $r = \min\{m, n\}$. Then

$$f(h) + g(h) = p(h) + q(h) + O(h^r),$$
$$f(h)g(h) = p(h)q(h) + O(h^r),$$

and

$$\frac{f(h)}{g(h)} = \frac{p(h)}{q(h)} + O(h^r) \qquad \text{provided that } g(h) \neq 0 \text{ and } q(h) \neq 0.$$

$O(h^n)$ Order of Approximation

Theorem 1.16 (Taylor' s Theorem). Assume that $f \in C^{n+1}[a, b]$. If both x_0 and $x = x_0 + h$ lie in $[a, b]$, then

$$f(x_0 + h) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} h^k + O(h^{n+1}).$$

Example

$$e^h = 1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + O(h^4) \text{ and } \cos(h) = 1 - \frac{h^2}{2!} + \frac{h^4}{4!} + O(h^6).$$

$$e^h + \cos(h) = 1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + O(h^4) + 1 - \frac{h^2}{2!} + \frac{h^4}{4!} + O(h^6)$$

$$= 2 + h + \frac{h^3}{3!} + \frac{h^4}{4!} + O(h^4) + O(h^6).$$

Since $\frac{h^4}{4!} + O(h^4) + O(h^6) = O(h^4)$

$$e^h + \cos(h) = 2 + h + \frac{h^3}{3!} + O(h^4)$$

$O(h^n)$ Order of Approximation

$$
\begin{aligned}
e^h\cos(h) &= \left(1+h+\frac{h^2}{2!}+\frac{h^3}{3!}+O(h^4)\right)\left(1-\frac{h^2}{2!}+\frac{h^4}{4!}+O(h^6)\right)\\
&= \left(1+h+\frac{h^2}{2!}+\frac{h^3}{3!}\right)\left(1-\frac{h^2}{2!}+\frac{h^4}{4!}\right)\\
&\quad+\left(1+h+\frac{h^2}{2!}+\frac{h^3}{3!}\right)O(h^6)+\left(1-\frac{h^2}{2!}+\frac{h^4}{4!}\right)O(h^4)\\
&\quad+O(h^4)O(h^6)\\
&= 1+h-\frac{h^3}{3}-\frac{5h^4}{24}-\frac{h^5}{24}+\frac{h^6}{48}+\frac{h^7}{144}\\
&\quad+O(h^6)+O(h^4)+O(h^4)O(h^6).
\end{aligned}
$$

Since $O(h^4)O(h^6)=O(h^{10})$ and

$$-\frac{5h^4}{24}-\frac{h^5}{24}+\frac{h^6}{48}+\frac{h^7}{144}+O(h^6)+O(h^4)+O(h^{10})=O(h^4),$$

We get

$$e^h\cos(h)=1+h-\frac{h^3}{3!}+O(h^4),$$

Order of Convergence of a Sequence

Definition 1.12. Suppose that $\lim_{n\to\infty} x_n = x$ and $\{r_n\}_{n=1}^{\infty}$ is a sequence with $\lim_{n\to\infty} r_n = 0$. We say that $\{x_n\}_{n=1}^{\infty}$ ***converges*** to x with the order of convergence $\boldsymbol{O}(r_n)$, if there exists a constant $\boldsymbol{K} > 0$ such that

$$\frac{|x_n - x|}{|r_n|} \leq \boldsymbol{K} \quad \text{for } n \text{ sufficiently large.}$$

This is indicated by writing $x_n = x + \boldsymbol{O}(r_n)$, or $x_n \to x$ with order of convergence $\boldsymbol{O}(r_n)$.

Example Let $x_n = \cos(n)/n^2$ and $r_n = 1/n^2$; then $\lim_{n\to\infty} x_n = 0$ with a rate of convergence $\boldsymbol{O}(1/n^2)$. This follows immediately from the relation

$$\frac{|\cos(n)/n^2|}{|1/n^2|} = |\cos(n)| \leq 1 \quad \text{for all } n.$$

Propagation of Error

- Consider $p = \hat{p} + \epsilon_p$ and q $= \hat{q} + \epsilon_q$

- The sum

$$p + q = \left(\hat{p} + \epsilon_p\right) + \left(\hat{q} + \epsilon_q\right) = (\hat{p} + \hat{q}) + \left(\epsilon_p + \epsilon_q\right).$$

- The multiplication

$$pq = \left(\hat{p} + \epsilon_p\right)\left(\hat{q} + \epsilon_q\right) = \hat{p}\hat{q} + \hat{p}\epsilon_q + \hat{q}\epsilon_p + \epsilon_p\epsilon_q.$$

$$R_{pq} = \frac{pq - \hat{p}\hat{q}}{pq} = \frac{\hat{p}\epsilon_q + \hat{q}\epsilon_p + \epsilon_p\epsilon_q}{pq} = \frac{\hat{p}\epsilon_q}{pq} + \frac{\hat{q}\epsilon_p}{pq} + \frac{\epsilon_p\epsilon_q}{pq}$$

$$\Downarrow$$

$$R_{pq} = \frac{pq - \hat{p}\hat{q}}{pq} \approx \frac{\epsilon_q}{q} + \frac{\epsilon_p}{p} + 0 = R_q + R_p$$

Propagation of Error

- Stable and unstable
- An example:

$$\{x_n\} = \{1/3^n\}$$

$$r_0 = 1 \quad \text{and} \quad r_n = \frac{1}{3} r_{n-1} \qquad \text{for } n = 1, 2, \ldots,$$

$$p_0 = 1, p_1 = \frac{1}{3}, \qquad \text{and} \quad p_n = \frac{4}{3} p_{n-1} - \frac{1}{3} p_{n-2} \qquad \text{for } n = 2, 3, \ldots,$$

$$q_0 = 1, q_1 = \frac{1}{3}, \qquad \text{and} \quad q_n = \frac{10}{3} q_{n-1} - q_{n-2} \qquad \text{for } n = 2, 3, \ldots,$$

- Consider $r_0 = 0.99996$, $p_1 = 0.33332$, $q_1 = 0.33332$

Table 1.5 Error Sequences $\{x_n - r_n\}$, $\{x_n - p_n\}$, and $\{x_n - q_n\}$

n	$x_n - r_n$	$x_n - p_n$	$x_n - q_n$
0	0.0000400000	0.0000000000	0.0000000000
1	0.0000133333	0.0000133333	0.0000013333
2	0.0000044444	0.0000177778	0.0000444444
3	0.0000014815	0.0000192593	0.0001348148
4	0.0000004938	0.0000197531	0.0004049383
5	0.0000001646	0.0000199177	0.0012149794
6	0.0000000549	0.0000199726	0.0036449931
7	0.0000000183	0.0000199909	0.0109349977
8	0.0000000061	0.0000199970	0.0328049992
9	0.0000000020	0.0000199990	0.0984149998
10	0.0000000007	0.0000199997	0.2952449999

Uncertainty in Data

- Data from real-world problems contain uncertainty or error. This type of error is referred to as ***noise***

- An improvement of precision is not accomplished by performing successive computations using noisy data.

- If start with data with d significant digits of accuracy, then the result of a computation should be reported in d significant digits of accuracy.

- Example: for data $p_1 = 4.152$ and $p_2 = 0.07931$, then $p_1 + p_2 = 4.231$, instead of $p_1 + p_2 = 4.23131$.