BACHELOR OF COMPUTER APPLICATIONS

SEMESTER 3

DCA2101

COMPUTER ORIENTED NUMERICAL METHODS

Unit 1

Errors and Computation

Table of Contents

INSPIRED BY LIFE

1. INTRODUCTION

Approximations and errors are an integral part of real life. They are everywhere and unavoidable. This is more so in the life of a computational scientist. We cannot use numerical methods and ignore the existence of errors. Errors come in a variety of forms and sizes; some are avoidable, some are not. For example, data conversion and roundoff errors cannot be avoided, but a human error can be eliminated. Some certain errors cannot be eliminated completely; we must at least know the bounds of these errors to make use of our final solution. Numerically, computed solutions are subject to certain errors. It may be fruitful to identify the error sources and then all these errors contribute to the total error in the final result. In this unit we discuss the various forms of approximations and errors, their sources, how they propagate during the numerical process, and how they affect the result as well as the solution process. Numerical computing is an approach for solving complex mathematical problems using only simple basic arithmetic operations. (The basic arithmetic operations performed by the computer are addition, subtraction, multiplication and division).

1.1 OBJECTIVES:

At the end of this unit the student should be able to:

- *Find the importance of numerical methods in real life problems*
- *Know the significance of errors in numerical calculation*
- *Identify the types of errors in numerical computations*

2. MATHEMATICAL PRELIMINARIES

In this section we state without proof, certain mathematical results which would be useful in the sequel.

2.1 Theorem:

If $f(x)$ is continuous in $a \le x \le b$ and if $f(a)$ and $f(b)$ are of opposite signs, then $f(c) = 0$ for at least one number c such that $a < c < b$.

2.2 Rolle's Theorem:

If $f(x)$ is (i) continuous in $[a, b]$. (ii) Differentiable in (a, b) and (iii) $f(a) = f(b)$ then $c \in (a, b)$ such that $f'(c)= 0$.

2.3 Generalized Rolle's Theorem:

Let $f(x)$be a function which is continuous on [a, b] n times differentiable on (a, b). If $f(x)$ vanishes at the $(n+1)$ distinct points $x_0<x_1< \ldots<x_n$ in (a, b), then there exists a number c in (a, b) such that $f^n(c) = 0$.

2.4 Example:

Verify Rolle's theorem for the function $f(x) = |x|$ *in* $(-1, 1)$.

Solution: Here $f(x) = -x$ for $-1 < x < 0$.

$= 0$ *for* $x = 0$

$= x$ *for* $0 < x < 1$

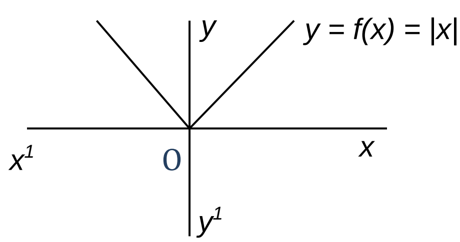
$f(-1) = 1 = f(1)$

Hence $f(-1) = f(1)$

$f^1(x) = -1$ *for* $-1 \le x \le 0$

$f^1(x) = 1$ *for* $0 \le x \le 1$

Therefore $f^1(x)$ does not exist at $x = 0$ and hence $f(x)$ is not differentiable in $(-1, 1)$ Rolle's theorem is not applicable to the function $f(x) = |x|$ *in* $(-1, 1)$



2.5 Intermediate Value Theorem:

Let $f(x)$ be continuous in $[a, b]$ and k be any number between $f(a)$ and $f(b)$. Then there exists a number c in (a, b) such that $f(c) = k$.

Example: $f(x) = x^2 + x - 1$

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

$$f(1) = 1 + 1 - 1 = 1 > 0$$

Here $f(x) = x^2 + x - 1$ is a continuous function and $f(0)$ and $f(1)$ are of different signs, therefore at least one real root lies between 0 and 1.

2.6 Lagrange's Mean value Theorem:

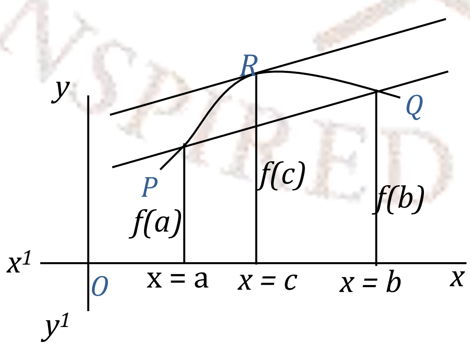
If $f(x)$ is (i) continuous in $[a, b]$ and (ii) differentiable in (a, b) then there exists at least one value 'c' in (a, b) such that $f^1(c) = \frac{f(b)-f(a)}{b-a}$.

2.7 Geometrical Interpretation of Lagrange's Mean Value theorem:

P and Q are two points on the continuous curve $y = f(x)$ corresponding to $x = a$ and $x = b$ respectively. Therefore $P[a, f(a)]$, $Q[b, f(b)]$ are two points on the curve.

Slope of the line joining the points P and Q is $\frac{f(b)-f(a)}{b-a}$, R is a point on the curve between P and Q corresponding to $x = c$, so that $f^1(c)$ is the slope of the tangent line at $R[c, f(c)]$ *where*

$f(c) = \frac{f(b)-f(a)}{b-a}$ means that the tangent at R is parallel to the chord PQ.



Hence this theorem tells that there is at least one point R on the curve PQ where the tangent to the curve is parallel to the chord PQ.

2.8 Taylor's Series for a function of one variable:

If *f(x)* is continuous and possesses continuous derivatives of order n in an interval that includes *x = a*, then in that interval

$$f(x) = f(a) + (x-a) f^{1}(a) + \frac{(x-a)^2}{2!} f''(a) + \ldots + \frac{(x-a)^{n-1}}{(n-1)!} f^{(n-1)}(a) + R_n(x)$$

where $R_n(x)$, the remainder term can be expressed as

$$R_n(x) = \frac{(x-a)^n}{n!} f^{n}(\xi), \quad a < \xi < x.$$

2.11 Maclaurin's Expansion:

Taylor's series at the origin i.e., at a=0

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

3. SIGNIFICANT DIGITS OR SIGNIFICANT FIGURES

There are two kinds of numbers, exact and approximate numbers. The numbers like 1, 2, 3, …, $\frac{1}{2}$ (= 0.5), $\frac{3}{2}$ (= 1.5), … are treated as exact numbers. But there are numbers $\frac{2}{7}$ (= 0.285714….), π (= 3.14159….), $\sqrt{2}$ (= 1.4142….), e (= 2.71828….) which cannot be expressed by a finite number of digits. These may be approximated by numbers 0.2857, 3.1416, 1.4142, 2.7183 respectively by omitting some digits, then these numbers are called approximate numbers. Thus numbers represent the given numbers to a certain degree of accuracy are called approximate numbers. For example, the approximate value of π is 3.1416 or if we desire a better approximation, it is 3.1415926589793, but we cannot write the exact value of π.

3.1 Definition:

The digits that are used to express a number are called significant digits or significant figures. The significant figures of a number are defined as follows:

Rule 1: (Numbers without decimal point): If the number does not have any decimal point, the significant figures of the number are the digits counted from the first non-zero digit on the left to the last non-zero digit on the right. Therefore, the number 12040 has four significant figures.

Rule 2: (Numbers with decimal point): If the number has a decimal point, the significant figures of the number are the digits counted from the first non-zero digit on the left to the last digit on the right side (irrespective of whether it is zero or non-zero). Therefore, the number, *2100.4*, has five significant figures, and the number *0.015*, has two significant figures.

Thus each of the numbers *3.1416*, *0.60125* and *4.0002* contain five significant digits while the numbers *0.00386*, *0.000587* and *0.00205* contain only three significant digits, since zeros only help to fix the position of the decimal point.

3.2 Note:

The following statements describe the notion of significant digits,

1. All non-zero digits are significant

2. All zeros occurring between non-zero digits are significant digits.
3. Trailing zeros following a decimal point are significant. For example ***3.500, 65.00*** and *0.**3210*** have four significant digits each.
4. Zeros between the decimal point and preceding a non-zero digit are not significant.

   The following numbers have four significant digits,

   *0.000**1234**, 0.00**2001**, 0.0**1321***

5. When the decimal point is not written, trailing zeros are not considered to be significant. For example *4500* may be written as $45\times10^2$and contains only two significant digits but *4500.0* contains four significant digits.

Integer numbers with trailing zeros may be written in scientific notation to specify the significant digits. The concept of *accuracy* and *precision* are closely related to significant digits. They are related as follows.

3.3 Example:

i) ***7.560*** has four significant digits

***25**000* has two significant digits

2.00004 has six significant digits

*0.0**4500*** has four significant digits

*0.0**201*** has three significant digits

0.00001 has one significant digit

100.00001 has eight significant digits

ii) *Accuracy* refers to the number of significant digits in a value. *For* example, the number *57.396* is accurate to five significant digits.

Self-Assessment Questions – 1

1. How many significant digits has the following?

   a) 7.56×10^2
   b) 7.560×10^3
   c) 7.5600×10^4.

4. ERRORS

4.1 Inherent errors

It is that quantity of error which present in the statement of the problem itself, before finding its solution. It arises due to the simplified assumptions made in the mathematical modeling of a problem. It can also arise when the data is obtained from certain physical measurements of the parameters of the problem.

These are two components, namely, data errors and conversion errors.

1. **Data error** (also known as *empirical error)* arises when data for a problem are obtained by some experimental means and are, therefore, of limited accuracy and precision. This may be due to some limitations in instrumentation and reading, and therefore may be unavoidable. A physical measurement, such as a distance, a voltage, or a time period cannot be exact. To remember that there is use in performing arithmetic operations to, say, four decimals when the original data themselves are only correct to two decimal places.

   For instance, the scale reading in a weighing machine may be accurate to only one decimal place.
2. **Conversion errors** (also known as *representation errors)* arise due to the limitations of the computer to store the data exactly. We know that the floating point representation retains only a specified number of digits. The digits that are not retained constitute the round off error.

4.2 Numerical errors

Numerical errors are introduced during the process of implementation of a numerical method. They come in two forms, *round off errors* and *truncation errors.* The total numerical error is the summation of these two errors. The total error can be reduced by devising suitable techniques for implementing the solution. We shall see in this section the magnitude of these errors.

4.3 Round off Errors

Every computer has a finite word length and therefore it is possible to store only a fixed number of digits of a given input number. Since computers store information in binary form, storing an exact decimal number in its binary form into the computer memory gives an error. This error is computer dependent. Also, at the end of computation of a particular problem, the final results in the computer, which is obviously in binary form, should be converted into decimal form – a form understandable to the user – before their print out. Therefore, an additional error is committed at this stage too. This error is called local round-off error.

It is clear that $(0.7625)_{10} = (0.11000011\ 0011)_2$. If a particular computer system has a word length of 12 bits only, then the decimal number *0.7625* is stored in the computer memory in binary form as *0.110000110011*. However, it is equivalent to *0.76245*. Thus, in storing the number *0.7625*, we have committed an error equal to *0.00005*, which is the round-off error; inherent with the computer system considered.

4.4 Definition:

We define the error as

Error = True value – Computed value

4.5 Numbers rounded-off to n significant digits

To round-off a number to n significant digits, discard all digits to the right of the n^{th} digits and if this discarded number is

i) less than half a unit in the n^{th} place, leave the n^{th} digit unchanged.
ii) greater than half the n^{th} place, increase the n^{th} digit by unity.
iii) exactly half a unit in the n^{th} place, increase the n^{th} digit by unity if it is odd, otherwise leave it unchanged.

The number thus rounded-off is said to be correct to *n* significant digits.

4.6 Example:

The following numbers rounded-off to four significant digits:

7.8926 to *7.893*

128.614 to *128.6*

3.14159 to *3.142*

0.859321 to *0.8593*

8476.7 to *8477*

In any Numerical computation, we come across the following types of errors.

4.7 Truncation errors

Truncation errors arise from using an approximation in place of an exact mathematical procedure. Typically, it is the error resulting from the truncation of the numerical process. We often use some finite number of terms to estimate the sum of an infinite series. For example,

$S = \sum_{i=0}^{\infty} a_i x^i$ is replaced by the finite sum $\sum_{i=0}^{n} a_i x^i$

Consider the following infinite series:

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

When we calculate the sine of an angle using this series, we cannot use all the terms in the series for computation. We usually terminate the process after a certain term is calculated. The terms "truncated" introduce an error which is called truncation error.

Many of the iterative procedures used in numerical computing are infinite and, therefore, a knowledge of this error is important. Truncation error can be reduced by using a better numerical model which usually increases the number of arithmetic operations.

In numerical integration, the truncation error can be reduced by increasing the number of points at which the function is integrated. But care should be exercised to see that the roundoff error which is bound to increase due to increase in arithmetic operations does not off-set the reduction in truncation error.

Example 4.8:

Find the truncation error in the result of the following function for $x = \frac{1}{5}$ when we use

i) first three terms
ii) first four terms
iii) first five terms

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!}.$$

Solution:

i) The truncation error when first three terms are added

$$= e^x - (1 + x + \frac{x^2}{2!}) = + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!}$$

$$= + \frac{(0.2)^3}{3!} + \frac{(0.2)^4}{4!} + \frac{(0.2)^5}{5!} + \frac{(0.2)^6}{6!}$$

$$= 0.1402755 \times 10^{-2}.$$

ii) Truncation error when first four terms are added.

Truncation error $= e^x - (1 + x + \frac{x^2}{2!} - \frac{x^3}{3!}) = + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!}$

$$= 0.694222 \times 10^{-4}.$$

iii) The truncation error when first five terms are added = 0.275555×10^{-5}.

5. ABSOLUTE, RELATIVE AND PERCENTAGE ERRORS

5.1 Definition:

Absolute error is the numerical difference between its true value of a quantity and its approximate value. If X is the true quantity and X_a is its approximate value then the absolute error E_a is given by

$E_a = \left|\text{True value} - \text{Approximate value}\right| = \left|X - X_a\right|$.

Let ΔX be a number such that $\left|X - X_a\right| \leq \Delta X$, then ΔX is an upper limit on the magnitude of absolute error and is said to measure *absolute accuracy*.

5.2 Definition:

The relative error is the absolute error divided by the true value of the quantity and this is denoted by E_r,

Relative error $E_r = \dfrac{\text{Absolute error}}{\text{True value}} = \dfrac{E_a}{X}$.

Similarly, the quantity $\dfrac{\Delta X}{\left|X\right|} \simeq \dfrac{\Delta X}{\left|X_a\right|}$ measures the *relative accuracy.*

5.3 Definition:

The percentage error E_p is given by $E_p = \dfrac{E_a}{X} \times 100 = E_r \times 100$.

Observations:

1. The relative and percentage errors are independent of the units used while absolute error is expressed in terms of these units.
2. If the number X is rounded to N decimal places, then $\Delta X = \frac{1}{2} \times 10^{-N}$, where ΔX is the absolute accuracy.

5.4 Example:

If the number $X = 0.51$ and is correct to two decimal places, then

$\Delta X = \frac{1}{2} \times 10^{-2} = 0.005$

and the relative accuracy is $\frac{\Delta X}{X} = \frac{0.005}{0.51} \simeq 0.98$

5.5 Example:

If $\frac{2}{3}$ is approximated by *0.667*, find the absolute and relative errors.

Solution:

Absolute error $E_a = \left|\text{True value} - \textit{Approximate}\ \text{value}\right|$

$$= \left|\frac{2}{3} - 0.667\right| = \left|\frac{2-2.01}{3}\right| = \frac{1}{3}\times10^{-3}$$

Relative error $E_r = \frac{\text{Absolute error}}{\text{True value}}$

$$= \frac{(\frac{1}{3}\times10^{-3})}{2/3} = \frac{1}{2}\times10^{-3}$$

6. SUMMARY

We summarize the concept of significant digits and its relation to accuracy. Various types of errors, viz. inherent errors, round-off errors, truncation errors, absolute errors, relative errors and percentage errors are also discussed.

7. TERMINAL QUESTIONS

1. Round-off the following numbers to 4 significant digits. *38.46325, 0.700292, 0.0022431, 19.235101, 2.36425*.
2. Find the percentage error if 625.483 is approximated to 3 significant digits.
3. The number *2.45789* is rounded-off to five significant figures. Find the relative and percentage error.
4. Find the sum of 0.123×10^3 and 0.456×10^2 and write the result in three significant digits.
5. Calculate the value of $\sqrt{102} - \sqrt{101}$. Correct to four significant digits.

8. ANSWERS

Self Assessment Questions

1. (i) Three significant digits.
   (ii) Four significant digits.
   (iii) Five significant digits.

Terminal Questions

1. Refer Section 1.3.3.
2. (Ans. *0.0772*)
3. (Ans. 4.069×10^{-6}, 4.069×10^{-4})
4. (Ans. 0.169×10^3)
5. (Ans. *0.04963*)