§ 8.5. Simpson's $\frac{1}{3}$ rd Rule 246
§ 8.6. Weddle's Rule 248
§ 8.7. Gauss Elimination Method 250
§ 8.8. Gauss-Seidel iterative Method 253
§ 8.9. Newton-Raphson method 256
§ 8.10. Regula-Falsi Method 258
§ 8.11. Euler's method 261
§ 8.12. Runga-Kutta Method of order 2. 262
§ 8.13. Runga-Kutta Method of order 4. 264

9. MATLAB
9.1 Introduction 267
9.2. Overview of the MATLAB Environment 267
9.3. The MATLAB System 268
9.4. Use of Matlab, 270
9.5. M-files: Scripts and Functions 275
9.6. Making Plots 276
9.7. Advanced Operations 277
Illustrative Examples 277

10. MATHEMATICA
10.1 Introduction 282
10.2. Use of mathematica 282
10.3. Some Mathematica calculations 284
10.4. Arithmetic Operations 285
10.5. User Defined Functions 285
10.6. Built-In Functions 285
10.7. Mathematical Constants 286
10.8. Incrementing Variables 286
10.9. Conditionals 286
10.10. Loops 287
10.11. Modules 287
10.12. Graphics 288
10.13. Solving equations 290

INDEX 292

QUESTIONS PAPER WBUT EXAMINATION 294

ST. THOMAS' COLLEGE OF ENGG. & TECH. BOOK BANK

1 APPROXIMATION IN NUMERICAL COMPUTATION

1.1 Introduction:

Numerical analysis is the subject of study to find the numerical solutions of mathematical problems by computational methods. In this context we shall consider various numerical methods for computing approximate numerical results of different mathematical problems and analyze the errors in the result due to the errors in the given data or the errors in methods or both. In this chapter, we now proceed to understand different type of errors, their estimate and propagation of errors in calculations of different numerical procedure.

1.2. Approximate numbers and significant figures.

(i) Approximate numbers

The numbers like $5, \frac{1}{2}, \frac{1}{25}, 200$ etc are called *exact numbers* because these numbers can be expressed exactly by a finite number of digits. On the other hand the numbers like $\sqrt{3}, 0, \pi$ etc can not be expressed exactly by a finite number of digits. We can write these numbers to a certain degree of accuracy, as $1 \cdot 7320, 2 \cdot 7183, 3 \cdot 1416$ etc which are only approximations to the true values and are called *approximate numbers*. Thus an approximate number is defind as a number approximated to the exact number and there is a slight difference between the exact and approximate numbers.

(ii) Significant digits

The digits which are used to represent a number are called significant digits (figures). Thus 1, 2, 3, ..., 9 are significant digits and 0 is significant digit except if it is used to fix the decimal place or to discard digits or to fill the unknown places. For example, the number 0.002396 has significant digits 2, 3, 9, 6, the zeros used here are not significant because they only fix the decimal places. But, for the number 0.01205, the significant digits are 1, 2, 0, 5. Here first zero after the decimal point is not significant.

1.3. Rounding off numbers.

There are numbers with large number of digits, viz, $\sqrt{2} = 1.414213\ldots$. For practical computation, it is necessary to cut-off some unwanted digits and retain only the desired such as 1.414 or 1.4142. This process of dropping unnecessary digits is called *rounding off.*

The general rules for rounding off a number to n significant figures are as follows :

Discard all digits to the right of the n^{th} digit and if among these discarded digits the digit in the $(n+1)^{th}$ place is

(i) greater than 5 then the digit in the n^{th} place is increased by 1.

(ii) less than 5 then the digit in the n^{th} place is left unchanged

(iii) exactly 5 then the convention is to leave the nth digit unaltered if it is even and to increase it by 1 if it is odd.

For example, the following numbers are rounded off to four significant figures :

6.02887	becomes	6.029
2.5632	becomes	2.563
79.3998	becomes	79.40
8.42853	becomes	8.428
8.42756	becomes	8.428

1.4. Errors and their computation

Let us start with some simple definitions about error. The difference between the true value of a quantity and the approximate value computed or obtained by measurement is called the *error.* Thus, if x_T and x_A be the true and approximate value of the solution in solving a problem, then the quantity $x_T - x_A$ gives the *error* in x_A. The *absolute error*, E_a invloved in x_A is given by

$$E_a = |x_T - x_A| \quad \ldots \quad (1)$$

The *relative error*, E_r in x_A is defined by

$$E_r = \frac{|x_T - x_A|}{x_T}, \text{ provided } x_T \neq 0 \quad \ldots \quad (2)$$

The percentage error, E_p is 100 times the relative error,

$$\text{i.e., } E_p = E_r \times 100 = \frac{|x_T - x_A|}{x_T} \times 100, \text{ provided } x_T \neq 0 \quad \ldots \quad (3)$$

As an illustration, suppose that the number 4.6285 be rounded off to 4.628 correct to four significant figures. Then we have

$$x_T = 4.6285, x_A = 4.628$$

So the absolute error is

$$E_a = |4.6285 - 4.628| = 0.0005$$

The relative error is given by

$$E_r = \frac{0.0005}{4.6285} = 1.0804 \times 10^{-4}$$

So the percentage error is

$$E_p = 1.0804 \times 10^{-4} \times 100$$
$$= 1.0804 \times 10^{-2}$$

Often we come across with two other type of errors in numerical computation neglecting a gross mistake. The error which is inherent in a numerical method itself or in the statement of a given problem is called *truncation error.* Another type of error is the *computational error* which arises during arithmetic computation due to the finite representation of numbers.

Truncation errors

The error which arise due to approximation of the result or due to the replacement of an infinite process by a finite one are called *truncation errors.* For exmaple, if a function $f(x)$ be expressed in the form of an infinite series, then for computation, we have to truncate the series at a certain stage causing an errorr, called truncation error.

Rounding errors

This is a one type of computation error which arise due to the process of rounding off the number during the computation.

Such errors are unavoidable in most of the calculation due to the limitations of the computing aids. In desk calculators we can reduce the round off errors by using more significant figures at each step of the computation, at least one more significant data than that of the given data, must be retained and rounding off is to be performed in the last operation.

For example, when a result 2.01536 is rounded off to four decimal places, then

$$x_T = 2.01536$$
$$x_A = 2.0154$$

So, in this case the rounded off error is given by

$$x_T - x_A = 2.01536 - 2.0154$$
$$= -0.00004.$$

1. 5. Fixed and floating point arithmetic

The first step in the computation with digital computers is to convert the decimal numbers to another number system with base b (say, binary number system with base 2) understandable to that particular computer and then to store these converted numbers in computer memory, which is a collection of small cells. Each cell can accommodate a *word*, that consists of the same number of characters, the left most being a sign and the others digits. Negative numbers are stored as absolute values plus a sign. The number of characters (sign plus digits) in a word that can be stored in a cell is called the *word length*. The word length varies from one computer to another. Real numbers can be stored in the computer word in two forms :

(i) Fixed point form

(ii) Floating point form

In fixed point form, a n digit number is assumed to have its decimal point at the left-hand end of the word. So all numbers are assumed to be less than 1 in magnitude. The fixed-point number with base b and n digits word length can be written as

$$\pm\left(a_1 b^{-1} + a_2 b^{-2} + \ldots + a_n b^{-n}\right) \quad \ldots \quad (4)$$

where $0 \le a_i < b, i = 1, 2, \ldots n$

To avoid the difficulty, of keeping every number less than 1 in magnitude during computation, most computer use normalised floating point form for a real number. A normalised floating point form of a real number is

$$\pm \cdot d_1 d_2 \ldots d_n \times b^k \quad \ldots \quad (5)$$

where b is the base of the number system used in the computer, $d_1, d_2, \ldots, d_n$ are all digits $(d_1 \neq 0)$ in the b-base system and the number k, called the exponent or characteristic, is such that $m \le k \le M$. The values of the numbers m and M vary with the computer. The fractional part $\cdot d_1 d_2 \ldots d_n$ is called the *mantissa*. In this case, the number is said to have n *significant digits*.

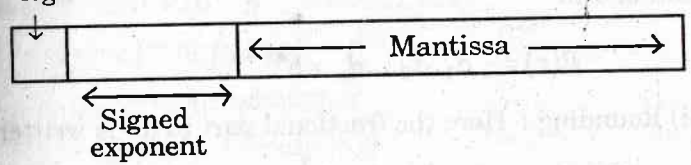

Fig.1

The above figure shows one way that a floating point number could be stored in a word. The first bit is reserved for the sign, the next series of bits for the signed exponent and the last bits for the mantissa.

Note that the mantissa is usually normalized if it has leading zero digits. For example, suppose the quantity $\frac{1}{34} = 0.02941176..$ was stored in a floating point base 10 system that allowed only five decimal places to be stored. Thus, $\frac{1}{34}$ would be stored as

$$0.02941 \times 10^0$$

But, in that case, the number can be normalized to remove the leading zero by multiplying the mantissa by 10 and lowering the exponent by 1 to give

$$0.29411 \times 10^{-1}$$

Thus, we retain an additional significant figure when the number is stored.

Floating point representation allows both fractions and very large numbers to be expressed on the computer. However, it has some disadvantage.

For example, floating point numbers take up more space and take longer to process than integer numbers. More significantly, their use introduces a source of error because the matissa holds only a finite number of significant figures. Thus a round off error is introduced. Thus, if a number x has the representation in the form

$$x = \cdot\ d_1 d_2 \ldots d_n d_{n+1} \ldots \times b^k \qquad \ldots \quad (6)$$

then the floating-point number $fl(x)$ in n-digit mantissa standard form can be obtained in the following two ways :

(i) Chopping (truncation) : Here we neglect $d_{n+1}, d_{n+2}, \ldots$ in (6) and obtain

$$fl(x) = \cdot\ d_1, d_2 \ldots d_n \times b^k \qquad \ldots \quad (7)$$

(ii) Rounding : Here the fractional part in (6) is written as

$$\cdot\ d_1\, d_2 \ldots d_n d_{n+1} + \frac{1}{2} b$$

and the first n digits are taken to write the floating point number.

For example, let us consider the fixed point number $N = 28.516789$ whose floating point form is $\cdot 28516789 \times 10^2$. If the word length of the computer is 5, chopping gives the number $N' = \cdot 28156 \times 10^2$, or $28 \cdot 156$ whereas rounding gives the number $N' = \cdot 28157 \times 10^2$ or 28.157

The errors due to chopping or rounding are called *storage errors* that occur at the input stage. At the intermediate stages of arithmetic operations, varied amount of errors occur depending on the nature of operations performed which are called *computational errors.*

When two floating point number are added or subtracted, the digits in the number with the smaller exponent must be shifted to align the decimal points. This shifting can lose some of the significant digits of one of the values.

For example, let $x = 387.4$, $y = 0.01234$

$\therefore\ x - y = 387.38766$ and $x + y = 387.41234$

If this arithmetic operation is done by a hypothetical computer with word length 4, then

$x' = 0.3874 \times 10^3$

and $y' = 0.0000 \times 10^3 \quad [\because\ y = 0.01234 = 0.00001234 \times 10^3]$

Then $x' \pm y' = 0.3874 \times 10^3$

Thus we see that a smaller value has no effect in addition or subtraction and so the error is serious.

Next, let $x = 3780$, $y = .321$

Then $x' = .378 \times 10^4$, $y' = .0000321 \times 10^4$

$\therefore\ x' - y' = .3779679 \times 10^4$

$= .377 \times 10^4$ (chopping)

$= .378 \times 10^4$ (rounding)

Thus shifting to align the decimal points has completely lost the significant digits of the subtrahend.

1.6. Propagation of Errors

Let us now consider the effect of calculations with numbers which involve errors. We first consider arithmetic operations $+, -, \times$ and $\div$. Let ω denote any one of these operations and ω^* be the computed version of that operation. Then, if x_A and y_A are the approximations in the calculations containing errors corresponding to the ture values x_T and y_T respectively, we can write

$$x_T = x_A + \varepsilon,\ y_T = y_A + \varepsilon'$$

where ε, ε' are the corresponding errors. Thus, if $x_A \omega^* y_A$ is the actually computed number, then for its error, we have

$$x_T \omega y_T - x_A \omega^* y_A = (x_T \omega y_T - x_A \omega y_A) + (x_A \omega y_A - x_A \omega^* y_A) \qquad \ldots \quad (8)$$

The first term on the right hand side of (8) within the bracket is known as the *propagated error* while the second term within the bracket is called the *rounding or chopping error.*

We now discuss some particular cases of propogated error

Case (i) For addition and subtraction, we have

$$(x_T \pm y_T) - (x_A \pm y_A) = (x_T - x_A) \pm (y_T - y_A)$$

$$= \varepsilon \pm \varepsilon^1 \quad \ldots \quad (9)$$

Thus error $(x_A \pm y_A)$ = error $(x_A) \pm$ error (y_A)

Case (ii) For multiplication, we have

$$x_T y_T - x_A y_A = x_T y_T - (x_T - \varepsilon)(y_T - \varepsilon')$$

$$= x_T \varepsilon' + y_T \varepsilon - \varepsilon\varepsilon' \quad \ldots \quad (10)$$

Thus error $(x_A y_A) = x_T$ error $(y_A) + y_T$ error (x_A)

$-$ error (x_A) error (y_A)

So the relative error in $x_A y_A$ is

$$E_r(x_A y_A) = \frac{x_T y_T - x_A y_A}{x_T y_T}$$

$$= \frac{\varepsilon}{x_T} + \frac{\varepsilon'}{y_T} - \frac{\varepsilon}{x_T}\frac{\varepsilon'}{y_T}$$

$$= E_r(x_A) + E_r(y_A) - E_r(x_A)E_r(y_A)$$

If $E_r(x_A); E_r(y_A) << 1$, then

$$\therefore E_r(x_A y_A) \simeq E_r(x_A) + E_r(y_A) \quad \ldots \quad (11)$$

Case (iii) For division, we get by proceeding along the same lines in multiplication,

$$E_r(x_A / y_A) \simeq E_r(x_A) - E_r(y_A), \text{ provided } |E_r(y_A)| << 1 \quad \ldots \quad (12)$$

Notes (i) The relative errors in multiplication or division do not propagate very rapidly

(ii) $E_r(x_A \pm y_A)$ can be quite poor in comparison with $E_r(x_A)$ and $E_r(y_A)$

As an illustration, suppose that $x_T = \pi$, $x_A = 3.1416$ and $y_T = \frac{22}{7}$, $y_A = 3.1429$

Then $x_T - x_A \simeq -7.35 \times 10^{-6}$

$y_T - y_A \simeq -4.29 \times 10^{-5}$

$\therefore \varepsilon_r(x_A) = -2.34 \times 10^{-6}$ and $\varepsilon_r(y_A) \simeq -1.36 \times 10^{-5}$

Also $(x_T + y_T) - (x_A + y_A) \simeq -5.02 \times 10^{-5}$

and

$(x_T - y_T) - (x_A - y_A) \simeq 3.55 \times 10^{-5}$

Hence $E_r(x_A + y_A) \simeq -7.99 \times 10^{-6}$

and

$E_r(x_A - y_A) \simeq -2.8 \times 10^{-2}$.

Thus although the error in $(x_A - y_A)$ is quite small, its relative error is not so and is much larger than $E_r(x_A)$ or $E_r(y_A)$. But the relative error in $(x_A + y_A)$ is quite small.

1.7. General formula for errors

Consider the differentiable function $u = f(u_1, u_2, \ldots, u_n)$ of several independent variables $u_1, u_2, \ldots, u_n$ subject to the errors $\Delta u_1, \Delta u_2, \ldots, \Delta u_n$ respectively. Then the errors in $u_i (i = 1, 2, \ldots, n)$ cause an error Δu in u and is given by

$$u + \Delta u = f(u_1 + \Delta u_1, u_2 + \Delta u_2, \ldots, u_n + \Delta u_n)$$

$$= f(u_1, u_2, \ldots, u_n) + \Delta u_1 \frac{\partial f}{\partial u_1} + \Delta u_2 \frac{\partial f}{\partial u_2} + \ldots + \Delta u_n \frac{\partial f}{\partial u_n}$$

$$+ \frac{1}{2}\left[(\Delta u_1)^2 \frac{\partial^2 f}{\partial u_1^2} + (\Delta u_2)^2 \frac{\partial^2 f}{\partial u_2^2} + \ldots + (\Delta u_n)^2 \frac{\partial^2 f}{\partial u_n^2} + 2\Delta u_1 . \Delta u_2 \frac{\partial^2 f}{\partial u_1 \partial u_2} + \ldots\right] + \ldots$$

Noting that the errors $\Delta u_1, \Delta u_2, \ldots, \Delta u_n$ are relatively small, we neglect their squares, products and higher powers. Thus we get

$$u + \Delta u \simeq u + \frac{\partial f}{\partial u_1}\Delta u_1 + \frac{\partial f}{\partial u_2} . \Delta u_2 + \ldots + \frac{\partial f}{\partial u_n}\Delta u_n$$

i.e., $\Delta u \simeq \frac{\partial f}{\partial u_1}\Delta u_1 + \frac{\partial f}{\partial u_2}\Delta u_2 + \ldots + \frac{\partial f}{\partial u_n}\Delta u_n$ ... (13)

This is the *general formula for computing the error* of a function $u = f(u_1, u_2, \ldots, u_n)$.

The relative error of u is given by

$$E_r = \frac{\Delta u}{u} \simeq \frac{\partial f}{\partial u_1}\cdot\frac{\Delta u_1}{u} + \frac{\partial f}{\partial u_2}\frac{\Delta u_2}{u} + \ldots + \frac{\partial f}{\partial u_n}\cdot\frac{\Delta u_n}{u} \quad \ldots (14)$$

1.7.1. Error in addition of numbers

Let $u = \sum_{i=1}^{n} u_i$, where $u_i (i = 1, 2, \ldots n)$ are n approximate numbers

Then we have

$$\Delta u = \sum_{i=1}^{n}(u_i + \Delta u_i) - \sum_{i=1}^{n} u_i = \sum_{i=1}^{n}\Delta u_i$$

$$\therefore |\Delta u| \le \sum_{i=1}^{n}|\Delta u_i|$$

Thus the absolute error of a sum of approximate numbers is less than on equal to the sum of the absolute errors of these numbers

Also $\left|\frac{\Delta u}{u}\right| \le \left|\frac{\Delta u_1}{u}\right| + \left|\frac{\Delta u_2}{u}\right| + \ldots + \left|\frac{\Delta u_n}{u}\right|$

This gives the maximum relative error of u.

1.7.2. Error in subtraction of numbers

Let $u = u_1 \sim u_2$, where u_1 and u_2 being two approximate numbers.

Then

$$\Delta u = (u_1 + \Delta u_1) \sim (u_2 + \Delta u_2) - (u_1 \sim u_2) = \Delta u_1 \sim \Delta u_2$$

$$\therefore |\Delta u| \le |\Delta u_1| + |\Delta u_2|$$

Thus the absolute error of the difference of two approximate numbers is less than or equal to the sum of the absolute errors of these numbers.

The maximum relative error of u is given by

$$E_r = \left|\frac{\Delta u}{u}\right| \le \left|\frac{\Delta u_1}{u}\right| + \left|\frac{\Delta u_2}{u}\right|$$

1.7.3. Error in multiplication of numbers.

Let $u = u_1 u_2 \ldots u_n$, where $u_i (i = 1, 2, \ldots n)$ being n approximate numbers

Then we have

$$u + \Delta u = (u_1 + \Delta u_1)(u_2 + \Delta u_2)\ldots(u_n + \Delta u_n)$$
$$= u_1 u_2 \ldots u_n + u_2 u_3 \ldots u_n \Delta u_1 + u_1 u_3 u_4 \ldots u_n \Delta u_2 + \ldots + u_1 u_2 \ldots u_{n-1}\Delta u_n + \ldots$$

Since the errors $\Delta u_1, \Delta u_2, \ldots, \Delta u_n$ are relatively small, we neglect their products and, therefore, we have

$$u + \Delta u \simeq u + u_2 u_3 \ldots u_n \Delta u_1 + u_1 u_3 \ldots u_n \Delta u_2 + \ldots + u_1 u_2 \ldots u_{n-1}\Delta u_n$$

so that

$$\frac{\Delta u}{u} \simeq \frac{\Delta u_1}{u_1} + \frac{\Delta u_2}{u_2} + \ldots + \frac{\Delta u_n}{u_n}$$

and hence

$$E_r = \left|\frac{\Delta u}{u}\right| \le \left|\frac{\Delta u_1}{u_1}\right| + \left|\frac{\Delta u_2}{u_2}\right| + \ldots + \left|\frac{\Delta u_n}{u_n}\right|$$

Thus the relative error of the product of n approximate numbers is less than or equal to the sum of the relative errors of these numbers.

1.7.4. Error in division of numbers.

Let $u = \frac{u_1}{u_2}$, u_1 and u_2 being two approximate numbers. Then we get

$$u + \Delta u = \frac{u_1 + \Delta u_1}{u_2 + \Delta u_2} = \frac{u_1 + \Delta u_1}{u_2}\left(1 + \frac{\Delta u_2}{u_2}\right)^{-1}$$
$$= \frac{u_1}{u_2}\left(1 + \frac{\Delta u_1}{u_1}\right)\left(1 - \frac{\Delta u_2}{u_2} + \ldots\right)$$
$$= u\left(1 + \frac{\Delta u_1}{u_1} - \frac{\Delta u_2}{u_2} + \ldots\right)$$

$$\therefore 1+\frac{\Delta u}{u}=1+\frac{\Delta u_1}{u_1}-\frac{\Delta u_2}{u_2}+\ldots$$

Neglecting the products of the relatively small errors Δu_1 and Δu_2, we have

$$\frac{\Delta u}{u}\simeq\frac{\Delta u_1}{u_1}-\frac{\Delta u_2}{u_2}$$

$$\therefore E_r=\left|\frac{\Delta u}{u}\right|\le\left|\frac{\Delta u_1}{u_1}\right|+\left|\frac{\Delta u_2}{u_2}\right|$$

Hence the relative error of the quotient of two approximate numbers is less than or equal to the sum of the relative errors of the number.

Illustrative Examples

Ex. 1. Round off the numbers 0.004935, 870.91, 100.53, –439.85 to three significant digits.

Solution. 0.00494, 871, 100, –440

Ex. 2. Round off

(a) 123456 to 3 significant figures

(b) 31.520457 to 1 significant figures

[W.B.U.T., CS-312, 2004]

Solution. (a) 12.3×10^4 (b) 3×10^2

Ex.3. Find the significant figures for the number

(i) 0.007501 (ii) 0.07510 (iii) 980.37

(iv) 109.00 (v) 10000

Solution. The significant figures are

(i) 7, 5, 0, 1 (ii) 7, 5, 1 (iii) 9, 8, 0, 3, 7

(iv) 1, 0, 9 (v) 1

Ex. 4. Find the number of significant digits of

(i) 0.1204 (ii) 0.002560

(iii) 3100, (iv) –56.0270

Solution. (i) 4 (ii) 3 (iii) 2 (iv) 5

Ex.5. Round off the following number to 4 decimal places :

(i) 3.567019, 9.77385, 36.00895, 0.00126

(ii) –6.00255, 3.08914×10^2, 0.28997, 100.567

Solution. (i) 3.5670 9.7738 36.0090 0.0013

(ii) –6.0026 3.0891×10^2 0.2900 1.0057×10^2

Ex. 6. If $\pi=3.14$ is used in place of 3.14156 find the absolute and relative errors.

Solution. Here, the true value, $x_T=3.14156$ and approximate value $x_A=3.14$

Absolute error, $E_a=|x_T-x_A|$

$$=|3.14156-3.14|$$

$$=0.00156$$

and the relative error, $E_r=\dfrac{|x_T-x_A|}{x_T}$

$$=\frac{0.00156}{3.14156}$$

$$=4.966\times10^{-4}$$

Ex. 7. If the value of $e=2.71828$ is replaced by 2.71937, what is the percentage error ?

Solution. Here $x_T=2.71828, x_A=2.71937$

So the required percentage error is

$$E_p=\frac{|x_T-x_A|}{x_T}\times100$$

$$=\frac{|2.71828-2.71937|}{2.71828}\times100$$

$$=4.01\times10^{-2}.$$

Ex. 8. Write down approximate representation of $\frac{2}{3}$ correct upto four significant digits; find the absolute, relative and percentage error.

Solution. Here $\frac{2}{3} = 0.66666.....$

$\therefore \quad \frac{2}{3} = 0.6667$, correct upto four significant digits

Let $x_T = \frac{2}{3}, x_A = 0.6667$

So the absoluter error is

$$E_a = |x_T - x_A| = \left|\frac{2}{3} - 0.6667\right|$$

$$= \frac{0.0001}{3}$$

$$= 3.3 \times 10^{-5}$$

The relative error is

$$E_r = \frac{E_a}{x_T} = \frac{0.0001}{3} / \frac{2}{3} = 5 \times 10^{-5}$$

$\therefore$ The percentage error is

$$E_p = E_r \times 100$$
$$= 5 \times 10^{-3}$$

Ex. 9. Find the percentage error in approximating $\frac{4}{3}$ to 1.3333

[W.B.U.T., M(CS)-312, 2010]

Solution. Let $x_T = \frac{4}{3}, \quad x_A = 1{\cdot}3333$

So the percentage error is

$$\frac{|x_T - x_A|}{x_T} \times 100$$

$$= \frac{\left|\frac{4}{3} - 1.3333\right|}{\frac{4}{3}} \times 100$$

$$= 0.0025$$

Ex. 10. Find the relative error in computations of $x - y$ for $x = 12.05$ and $y = 8.02$ having absolute errors $\Delta x = 0.005$ and $\Delta y = 0.001$

Solution. Let $u = x - y$

$$\therefore \quad u = 12.05 - 8.02 = 4.03$$

The maximum possible absolute error is

$$|\Delta u| = |\Delta x| + |\Delta y|$$
$$= 0.005 + 0.001$$
$$= 0.006$$

Hence the relative error in computation of $u = x - y$ is

$$\frac{|\Delta u|}{u} = \frac{0.006}{4.03}$$

$\simeq 0.0015$, correct upto four decimal places.

Ex. 11. Find absolute, relative and percentage error in computation of $f(x) = 3\sin x - 2x^2 - 9$ for $x = 0$ when the error in x is 0.003.

Solution. Here $f(x) = 3\sin x - 2x^2 - 9$

$$\therefore \Delta f(x) = \frac{df}{dx} \cdot \Delta x$$

$$= (3\cos x - 4x)\Delta x$$

At $x = 0, \ \Delta x = 0.003$

So the error in computation of $f(x)$ is

$$\Delta f(x) = (3 \times \cos 0 - 4 \times 0)0.003$$

$$= 3 \times 0.003$$

$$= 0.009$$

Thus the absolute error is 0.009

So the relative error is

$$\frac{|\Delta f(x)|}{|f(0)|} = \frac{0.009}{9} = 001$$

Hence the percentage error is

$$0.001\times100 = 0.1$$

Ex. 12. If $u = \frac{x^2y}{2}, \Delta x = 0.01, \Delta y = 0.02$ at $x = 2, y = 1$ compute the maximum absolute and relative errors in evaluating u.

Solution. We have

$$\frac{\partial u}{\partial x} = xy, \frac{\partial u}{\partial y} = \frac{x^2}{2}$$

$$\therefore \Delta u \simeq \frac{\partial u}{\partial x}.\Delta x + \frac{\partial u}{\partial y}.\Delta y$$

$$= xy\Delta x + \frac{x^2}{2}.\Delta y$$

$$\therefore |\Delta u| \le |xy||\Delta x| + \frac{x^2}{2}|\Delta y|$$

$$= 2\times0.01 + \frac{4}{2}\times0.02$$

$$= 0.06$$

So the maximum absolute error in u is 0.06 and hence the maximum relative error of u is given by

$$(E_r)_{m\times n} = \frac{0.06}{u} = \frac{0.06}{\frac{2^2\times1}{2}} = 0.03$$

Ex. 13. Obtain the relative error in $u = x_1^n x_2^n x_3^n$ in terms of the relative errors of x_1, x_2, x_3

Solution. We have

$$u = x_1^n x_2^n x_3^n$$

$$\therefore \log u = n\log x_1 + n\log x_2 + n\log x_3$$

$$\therefore \frac{1}{u}\frac{\partial u}{\partial x_1} = \frac{n}{x_1}, \frac{1}{u}\frac{\partial u}{\partial x_2} = \frac{n}{x_2}, \frac{1}{u}\frac{\partial u}{\partial x_3} = \frac{n}{x_3}$$

So the relative error of u is given by

$$E_r = \frac{\Delta u}{u} \simeq \frac{\partial u}{\partial x_1}\cdot\frac{\Delta x_1}{u} + \frac{\partial u}{\partial x_2}\cdot\frac{\Delta x_2}{u} + \frac{\partial u}{\partial x_3}\cdot\frac{\Delta x_3}{u}$$

$$= n\frac{\Delta x_1}{x_1} + n\frac{\Delta x_2}{x_2} + n\frac{\Delta x_3}{x_3}$$

As the errors $\Delta x_1, \Delta x_2$ and Δx_3 may be positive or negative, so we take the absolute values of the terms on the right side. Thus we get

$$(E_r)_{\max} \le n\left|\frac{\Delta x_1}{x_1}\right| + n\left|\frac{\Delta x_2}{x_2}\right| + n\left|\frac{\Delta x_3}{x_3}\right|$$

Ex. 14. Find an upper limit of the relative error in the measure of $w = \frac{x^\alpha y^\beta}{z^\gamma}$

Solution. We have

$$w = \frac{x^\alpha y^\beta}{z^\gamma}$$

$$\therefore \log w = \alpha\log x + \beta\log y - \gamma\log z$$

$$\therefore \frac{1}{w}\frac{\partial w}{\partial x} = \frac{\alpha}{x}, \quad \text{i.e.,} \quad \frac{\partial w}{\partial x} = \frac{\alpha w}{x}$$

Similarly $\frac{\partial w}{\partial y} = \frac{\beta w}{y}, \frac{\partial w}{\partial z} = -\frac{\gamma w}{z}$

Now, $\Delta w = \frac{\partial w}{\partial x}\cdot\Delta x + \frac{\partial w}{\partial z}\cdot\Delta y + \frac{\partial w}{\partial z}\cdot\Delta z$

gives

$$\frac{\Delta w}{w} = \frac{\alpha}{x}\Delta x + \frac{\beta}{y}\Delta y - \frac{\gamma}{z}\Delta z$$

Thus the relative error of ω is given by

$$E_r = \left|\frac{\Delta w}{w}\right| \le \alpha\left|\frac{\Delta x}{x}\right| + \beta\left|\frac{\Delta y}{y}\right| + \gamma\left|\frac{\Delta z}{z}\right|$$

Hence the upper limit of the relative error is

$$\alpha\left|\frac{\Delta x}{x}\right|+\beta\left|\frac{\Delta y}{y}\right|+\gamma\left|\frac{\Delta z}{z}\right|.$$

Exercise

I. Short Answer Questions

1. Find the number of significant digit

(i) 120.00 (ii) 12340 (iii) 0.2050

(iv) 20000 (v) –12.970 (vi) 89.7010

2. Round off the following numbers to 3 significant figures :

(i) 0.0063, 2.138, –46.285, 77.75

(ii) 2, 12, 1506, 19928

3. Round off the following numbers :

(i) 170.570 upto four significant figures

(ii) 21753 upto three significant figures

(iii) –79.861 upto three significant figures

(iv) 32056 upto two significant figures

(v) 0.2502 upto one significant figures

4. Round off the following number to 3 decimal places

(i) 2.47235, 0.003568, 42.3085, 9.77345

(ii) 1.2755, 6.4452×10^3, 0.999500, 0.24813

(iii) 98.9268, 0.00282, –72.0506, –0.0056

5. If $\frac{5}{3}$ is approximated to 1.6667, find the absolute error. *[W.B.U.T., CS-312, 2007]*

6. Find the percentage error in approximate representation of $\frac{7}{6}$ by 1.16

7. If 0.8333 is taken to be an approximate value of $\frac{5}{6}$, find the percentage error

8. Find the percentage error in approximate representation of $\frac{4}{1}$ by 1.33.

9. Find the absolute and relative error when 5.0214 is round off to 3 significant figures.

10. Round off the number 8.03567 to four significant digits and compute the percentage error.

ST. THOMAS' COLLEGE OF ENGG. & TECH. BOOK BANK

Answers

1. (i) 2, (ii) 4, (iii) 3, (iv) 1, (v) 4, (vi) 5

2. (i) $630\times10^{-5}, 214, -46.3, 77.8$

(ii) $200\times10^{-2}, 120\times10^{-1}, 151\times10^{2}, 1.99\times10^{4}$

3. (i) 170.6 (ii) 21.8×10^3 (iii) –79.9 (iv) 3.2×10^4 (v) 0.3

4. (i) 2.472, 0.004, 42.308 (ii) 1.276, 6.445×10^3, 1.000, 0.248

(iii) 98.927, 0.003, –72.051, –0.006

5. 0.000033 **6.** 0.571 **7.** 0.004 **8.** 0.25

9. 0.0014, 27.88×10^{-5} **10.** 8·036, $4\cdot1\times10^{-7}$

II. Long Answer Questions

1. Find the absolute, relative and percentage error, if $\frac{1}{3}$ is approximated by 0.333.

2. If $\frac{5}{6}$ is represented by the approximate number 0.8333, compute absolute, relative and percentage errors.

3. If 3.45234 be an approximate value of 3.45678, find the absolute, relative and percentage errors.

4. If $\Delta r = \Delta h = 0.1$, find the absolute and relative errors upto three significant figures in $v = \frac{1}{3}\pi r^2 h$ when $r = 2$ and $h = 3$.

5. If $u = \frac{x^2}{y}$ and errors in x, y be 0.01, compute the relativ

maximum error in u when $x = y = 1$.

6. Find the maximum absolute error in computing $u = \frac{x^3y^2}{z}$ when $x = y = z = 0.1$, $\Delta x = \Delta y = \Delta z = 0.002$

7. Calculate the relative error in computation of $x - y$ fo $x = 3.21, y = 2.12$ having absolute errors $\Delta x = 0.003$ an $\Delta y = 0 \cdot 001$

8. Find the error in calculating the area of a circle of redius when error in radius is 0.1.

9. Given $u = x^3y^2$, if x_0, y_0 be the approximate values of x, respectively and $\Delta x_0, \Delta y_0$ are the absolute errors in them determine the relative error in u.

10. Obtain the relative error in $u = x_1^m x_2^m \ldots x_n^m$ in terms of th relative errors of $x_1, x_2, \ldots, x_n$.

Answers

1. $3.333 \times 10^{-4}, 0.001, 0.1$ **2.** $33.33 \times 10^{-6}, 4.0 \times 10^{-5}, 4.0 \times 10^{-3}$

3. 0.00444, 0.128443×10^{-2}, 0.128443 **4.** 0.168, 0.0135

5. 0.03 **6.** 0.012 **7.** $0 \cdot 004$

8. 55 **9.** $3\left|\frac{\Delta x_0}{x_0}\right| + \frac{3}{2}\left|\frac{\Delta y_0}{y_0}\right|$

10. $m\left|\frac{\Delta x_1}{x_1}\right| + m\left|\frac{\Delta x_2}{x_2}\right| + \cdots + m\left|\frac{\Delta x_n}{x_n}\right|$

III. Multiple Choice Questions

1. The number of significant figures in 0.03409 is

(a) five (b) six (c) seven (d) four

[W.B.U.T., CS-312, 2003]

2. Which are the following digits are not significant of the number 0.025

(a) 0 (b) 2 (c) 5 (d)none

3. Which are the following digits are not significant digit of the number 1.307

(a)3 (b) 1 (c) 0 (d) none

4. The number of significant digit in the number 3.0056 is

(a) 3 (b) 4 (c) 5 (d) 2

5. The number of significant figures in 6,00,000 is

(a) 1 (b) 7 (c) 0 (d) 6

[W.B.U.T., M(CS)-312, 2009]

6. After rounding off to three places of decimals the number 57.1092 becomes

(a) 57.109 (b) 57.100 (c) 57.110 (d) 0.109

7. After being rounded off to two places of decimals the number 8.1083 becomes

(a) 8.10 (b) 0.11 (c) 8.11 (d) none

8. After being rounded off to three places of decimal the number 0.199561 becomes

(a) 0.199 (b) 0.190 (c) 0.200 (d) 0.210

9. After being rounded off to one place of decimal the number 35.956 becomes

(a) 36.0 (b) 3.6 (c) 35.9 (d) 35.0

10. The number 3.4506531 when rounded off to 4 places of decimal will give

(a) 3.4506 (b) 3.4507 (c) 3.451 (d) none

[W.B.U.T., CS-312, 2003]

11. The significant digit of 0.0001234 is

(a) 7 (b) 4 (c) 8 (d) 6

[W.B.U.T., M(CS)-312, 2010]

12. After being rounded off to three significant figure the number 125.42 becomes

(a) 125 (b) 126 (c) 125.420 (d)0.102148

13. Rounding off the number 0.0063945 correct upto 4 significant figure is

(a) 0.0064 (b) 0.0063 (c) 0.006395 (d) 0.006394

[W.B.U.T., CS-312, 2008]

14. When 9.8 is the approximate value of 9.79, the percentage error becomes

(a) –0.01 (b) 0.01 (c) 0.00102145 (d) 0.102145

15. When 0.0081 is the approximate value of 0.00809, the error is

(a) 0.001 (b) 0.00001 (c) –0.00001 (d) none

16. When 0.1 approximates the value 0.09, the relative error is

(a) $\frac{1}{9}$ (b) 0.11111 (c) 0.11 (d) none

17. The percentage error in approximating $\frac{4}{3}$ to 1.3333 is

(a) 0.0025% (b) 25% (c) 0.00025% (d) 0.25%

[W.B.U.T., M(CS)-312, 2010]

18. If $\frac{5}{3}$ is approximated to 1.6667, then absolute error is

(a) 0.000033 (b) 0.000043

(c) 0.000045 (d) 0.000051

[W.B.U.T., CS-312, 2007]

19. The kind of error when 3.14 is approximate values of π is

(a) inherent error (b) truncation error

(c) round off error (d) percentage error

20. Round-off error is a form of

(a) truncation error (b) numerical error

(c) inherent error (d) none of these

[W.B.U.T., CS-312, 2006]

21. Relative error is measured by which of the following expressions?

(a) $\text{mod}\left[\frac{\text{absolute error}}{\text{approximate error}}\right]$

(b) $\text{mod}\left[\frac{\text{absolute error}}{\text{exact error}}\right]$

(c) mod (exact value–approximate value)

(d) none of these

[W.B.U.T., CS-312, 2008]

22. The ratio of absolute error and of the true value is called

(a) relative error (b) absolute error

(c) truncation error (d) inherent error

23. If a be the actual value and e be its estimated value, the formula for relative error is

(a) $\frac{a}{c}$ (b) $\frac{|a-e|}{c}$ (c) $\frac{(a-e)}{a}$ (d) $\frac{|a-e|}{a}$

[W.B.U.T., CS-312, 2006]

24. If E_a is the absolute error in a quantity whose true and approximate value are given by x_t and x_a, then the relative error is given by

(a) $\left|\frac{E_a}{x_a}\right|$ (b) $\left|\frac{E_a}{x_t}\right|$ (c) $\left|\frac{E_a}{x_t - x_a}\right|$ (d) $|E_a|$

[W.B.U.T., CS-312, 2007]

25. The error which is inherent in a numerical method itself, is called computational error

(a) True (b) False

[W.B.U.T., CS-312, 2002]

26. The error in a tabulated value does not propagate in successive differences

(a) True (b) False

[*W.B.U.T., CS-312, 2002*]

27. In the problem "Find area of circle having radius 2; given $\pi = 3.14$ ", state the kind of error of the approximation 3.14 for π

(a) truncation error (b) round off error

(c) inherent error (d) relative error

Answers

1.a	**2.**a	**3.**d	**4.**c	**5.**a	**6.**a	**7.**c	**8.**c	**9.** a	**10.**a
11.b	**12.**a	**13.**d	**14.** d	**15.**c	**16.**a	**17.**a	**18.**a	**19.**c	**20.**b
21.b	**22.**a	**23.**d	**24.**b	**25.**b	**26.**b	**27** c			

2 CALCULUS OF FINITE DIFFERENCES

2.1 Introduction.

The calculus of finite difference deals with the changes in the values of the dependent variable with the change of the independent variable. Finite difference method is successfully applied in numerical analysis in interpolation, numerical differentiation and integration etc. In this chapter, we introduce and discuss different types of difference operators, their relations, fundamental theorem of difference calculus and their applications.

2.2. Finite differences.

Let $y = f(x)$ be a real-valued function of x defined in an interval $[a,b]$ and its values are known for $(n+1)$ equally spacing points x_i $(i = 0, 1, 2, \cdots, n)$ such that $x_i = x_0 + ih$ $(i = 0, 1, 2, \cdots, n)$ where $x_0 = a, x_n = b$ and h is the *space length*. Then x_i $(i = 0, 1, 2, \cdots, n)$ are called *nodes* and the corresponding values y_i are termed as *entries*.

We now introduce the concept of various type differences in order to find the values of $f(x)$ or its derivative for some intermediate values of x in $[a,b]$.

2.3. Forward Differences.

The differences $y_1 - y_0, y_2 - y_1, y_3 - y_2, \cdots, y_n - y_{n-1}$ for the entries $y_0, y_1, y_2, \cdots, y_{n-1}, y_n$ are called *first forward differences* and are denoted by $\Delta y_0, \Delta y_1, \Delta y_2, \cdots, \Delta y_{n-1}$ respectively. Thus we have

$$\Delta y_i = y_{i+1} - y_i \quad (i = 0, 1, 2, \cdots, n-1) \qquad \ldots \quad (1)$$

where Δ is called *forward difference operator*. In general forward difference operator is defined by

$$\Delta f(x) = f(x+h) - f(x) \qquad \ldots \quad (2)$$

Similarly, the higher order forward differences are define as

$$\Delta^2 y_i = \Delta y_{i+1} - \Delta y_i$$