1 Errors in Numerical Computing

1.1 Introduction

In practical applications, an engineer would finally obtain results in a numerical form. The aim of numerical analysis is to provide efficient methods for obtaining numerical answers to such problems.

1.1.1 Approximate Value

There are certain numbers whose exact value cannot be written. For the famous number π, we can only write value of π to certain degree of accuracy. For example, π is 3.1416 or 3.14159265. This values of π are called the approximate values of π. The exact value of π, we cannot write. Another such number is the *Euler's number 'e'*. One scenario is using the value of π in calculating the area of circle? *Can you come up with another scenario where approximate value of a number is used instead of its exact value?*

1.1.2 Significant Digits

(Significant Figures) The digits that are significant (important) in a number expressed as digits, are called significant digits.

Rules to identify significant figures in a number:

1. All non-zero digits are significant. The number 21.11 has four significant digits.

2. Zeros between two significant non-zero digits are significant. The number 20001 has five significant digits.

3. Zeros to the left of the first non-zero digit (leading zeros) are not significant. The number 0.0085 has two significant digits.

4. Zeros to the right of the last non-zero digit (trailing zeros) in a number with the decimal point are significant. The number 320. has three significant digits, and 320.00 has five significant digits.

5. When the decimal point is not written, the trailing zeros are not significant. The number 4500 may be written as 45×10^2 has two significant digits. However, the number 4500.0 has five significant digits.

6. Integers with trailing zeros may be written in scientific notation to specify the significant digits.

7.56×10^4	has 3 significant digits
7.560×10^4	has 4 significant digits
7.5600×10^4	has 5 significant digits

The concept of accuracy and precision are closely related to significant digits.

1. **Accuracy**

   This refers to the number of significant digits in a value. For example, the number 57.396 is accurate to five significant digits.

2. **Precision**

   This refers to the number of decimal positions, i.e. the order of magnitude of the last digit in a value. For example, the number 57.396 has a precision of 0.001 or 10^{-3}.

1.1.3 Exercise

1. Which of the following numbers has the greatest precision?

   a). 4.3201 b). 4.32 c). 4.320106

2. What is the accuracy of the following numbers?

   a). 95.763 b). 0.008472 c). 0.0456000 d). 36 e). 3600 f). 3600.00

1.2 Approximations

In numerical computation, we come across numbers which have large number of digits and it will be necessary to bring them to the required number of significant figures. For instance, the finite precision of computer storage, (using fixed number of bits), does not allow us to store the infinite digits of certain fractions like 1/3. So, the representation of 1/3 in the computers is going to an approximation.

1.2.1 Rounding off

Rounding off is the method of approximation used for the numbers. There are two types of rounding off.

Chopping

A number is written up to its certain digits and remaining digits are simply discarded. For example the number 1/3 is chopped to 0.3333.

Symmetric Rounding

A number is adjusted to the nearest representable value. For example 2.678 is rounded to 2.68, and 2.674 rounded to 2.67. Rules for rounding off.

1. To round-off a number to n significant digits , discard all digits to the right of the nth digit, and if the first digit of this discared number is,
2. less than half a unit in the nth place, leave the nth digit unaltered;
3. greater than half a unit in the nth place, increase the nth digit by unity;
4. exactly half a unit in the nth place, increase the nth digit by unity if it is odd; otherwise, leave it unchanged.

1.2.2 Truncation

While rounding off is the approximation in a number, truncation is the approximation in a mathematical procedure. Especially an infinite mathematical procedure or a complicated mathematical procedure is approximate by a finite mathematical procedure. The function $sinx$ is representation in a computer by its *series*; $sinx = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + ...$ But due to finite precision of computer, we represent sin function by only the finite terms of its series. Another scenario of truncation error is representing a continuous function in a computer. Digital systems like computer cannot represent a continuous phenomenon like a continuous function, because digital system is a discrete system. While approximating a function with a step-size h, the truncation error depends upon the step size. Forward difference approximation of a derivative of function has truncation error of order, $O(h)$.

1.3 Error

Both the rounding off and truncation causes error as they reduce a given number to an approximate value. Using approximate instead of exact value of number gives rise to the error.

1.3.1 Types of Error

Absolute Error

The absolute error E_A is the difference between the exact-value X and the approximate-value X_1 of a number.

$$E_A = X - X_1 = \delta X \tag{1.1}$$

Relative Error

The relative error E_R is the ratio of the absolute error to the exact-value.

$$E_R = \frac{E_A}{X} = \frac{X - X_1}{X} \tag{1.2}$$

Percentage Error

The percentage error E_P is

$$E_P = 100(E_R) = \frac{X - X_1}{X} \times 100 \tag{1.3}$$

The number 2.146879 is rounded to three significant digits. Find its errors.

1.3.2 Absolute errors of Sum, Product and Product

Suppose a_1 and a_2 are approximate values of two numbers, with their absolute errors E_A^1 and E_A^2 respectively.

Sum

If E_A is the absolute error of $a_1 + a_2$ then,

$$E_A = (a_1 + E_A^1) + (a_2 + E_A^2) - (a_1 + a_2) = E_A^1 + E_A^2 \tag{1.4}$$

Product

If E_A is the absolute error of $a_1 a_2$ then,

$$E_A = (a_1 + E_A^1)(a_2 + E_A^2) - (a_1 a_2) = a_1 E_A^2 + a_2 E_A^1 + E_A^1 E_A^2 \approx a_1 E_A^2 + a_2 E_A^1 \tag{1.5}$$

Quotient

If E_A is the absolute error of a_1/a_2 then,

$$E_A = \frac{a_1 + E_A^1}{a_2 + E_A^2} - \frac{a_1}{a_2} = \frac{a_2 E_A^1 - a_1 E_A^2}{a_2(a_2 + E_A^2)} = \frac{a_2 E_A^1 - a_1 E_A^2}{a_2 a_2 (1 + E_A^2/a_2)} \approx \frac{a_2 E_A^1 - a_1 E_A^2}{(a_2)^2}$$

This implies,

$$E_A = \frac{a_1}{a_2}\left[\frac{E_A^1}{a_1} - \frac{E_A^2}{a_2}\right] \tag{1.6}$$

1.3.3 Operations on numbers of different absolute accuracies

While dealing with several numbers of different number of significant digits, the following procedure may be adopted:

1. Isolate the number with the greatest absolute error,
2. Round-off all other numbers retaining in them one digit more than in the isolated number,
3. Add up, and
4. Round-off the sum by discarding one digit.

1.3.4 Upper limit of Absolute Error

The number ΔX such that

$$|X_1 - X| \leq \Delta X. \tag{1.7}$$

Then ΔX is an upper limit on the magnitude of the absolute error and is said to measure *absolute accuracy.*

Theorem 1. If the number X is rounded to N decimal places, then $\Delta X = \frac{1}{2}(10^{-N})$.

Verify this theorem by taking $1.23x$*; x varies from 1 to 9, to two decimal places.*

1.4 General Error Formula

Let us consider a function u that depends upon the variables: x, y, z, i.e $u = f(x, y, z)$. Let $\Delta x, \Delta y, \Delta z$ be the errors in x, y, and z, respectively. Then the total error in u is Δu, which is approximated by du as follows: We have the total derivative:

$$du = \frac{\partial u}{\partial x}\, dx + \frac{\partial u}{\partial y}\, dy + \frac{\partial u}{\partial z}\, dz \tag{1.8}$$

Approximating dx by Δx, dy by Δy and dz by Δz we get,

$$\Delta u \approx \frac{\partial u}{\partial x}\, \Delta x + \frac{\partial u}{\partial y}\, \Delta y + \frac{\partial u}{\partial z}\, \Delta z \tag{1.9}$$

The relation 1.9 gives the absolute general error formula. Then the relative error formula is given by $E_R = \frac{\Delta u}{u}$.

1.4.1 Exercise

1. Round off the following numbers to two decimal places.
   48.21461, 2.3742, 52.275

2. Round off the following numbers to four significant figures:
   38.46235, 0.70029, 0.0022218, 19.235101, 2.36425

3. Two numbers 3.1425 and 34.5851 are rounded to 2 decimal places. Find the error in their sum, product and quotient.

4. Find the absolute error in the sum of the numbers 105.6, 27.28, 5.63, 0.1467, 0.000523, 208.5, 0.0235, 0.432, 0.0467, where each number is correct to the digits given.

5. Find the absolute error in the product uv, where $u = 4.536$ and $v = 1.32$, the numbers being correct up to the digits given. Find the relative error of the quotient u/v as well.

6. Find the percentage error in $z = (1/8)\, xy^3$ when $x = 3014 \pm 0.0016$ and $y = 4.5 \pm 0.05$.

7. Prove that in a product of three nonzero numbers, the relative error does not exceed the sum of the relative errors of the numbers.