MODULE 4

ERRORS

Module Content

- Introduction
- Objectives
- Analytic versus numerical methods
- Absolute versus relative errors
- Sources and types of computational errors
- Measures of reducing or eliminating computational errors
- Summary

4.1 Introduction

This module aims at providing the learner with answers to the following important questions; questions raised by many students taking a numerical methods course for the first time.

- What is a numerical solution and how does such a solution differ from an exact (true) or analytical solution?
- Why should one learn numerical methods? Are numerical methods needed?
- What are errors in the context of mathematics? What are the main sources of computational errors? How can one eliminate errors or reduce their effect on numerical solutions?

We begin by explaining the difference between an analytic and a numerical solution. To answer the questions on errors we define the concept of an error and list a number of sources of errors. In the sequel, we also suggest for each type and source of error, practical steps to be taken to reduce the error and hence its impact on numerical solution.

4.2 Objectives

At the end of this module the learner is expected to be able to

- **Define** the concept of a mathematical error

- **Distinguish** between absolute and relative errors
- **Name** the different types of errors
- **List** the chief sources of errors and explain giving examples practical steps that can taken to eliminate or reduce their cumulative effect on the numerical solution
- **Explain and measure** the difference between the size (accuracy) and the seriousness (precision) of an error.

Key Words
Error in an approximation: The difference between the exact value and the approximate value of a quantity.

Absolute error: The error without consideration of the sign (positive or negative).

Relative error: The ratio between the absolute error and the exact quantity.

Initial, truncation and rounding errors: Different types of errors caused by different sources of error.

4.3 Analytic versus Numerical Methods

What is the difference between an analytical method and a numerical method?

An **analytic method** for solving a given mathematical problem is any method based on rigorous mathematical analysis whose application leads to the **exact** solution of the problem. The solution obtained is also known as **analytic solution.**

In contrast to an analytical method, a **numerical method** for solving a mathematical problem is any method based on rigorous mathematical analysis whose application, in most cases, can only lead to an **approximate (non-exact)** solution. The solution obtained this way is known as a **numerical solution.** In some very rare cases, a numerical method may produce the exact solution.

Example

Find the value of x at which the parabola $y = 21x^2 + 8x - 5$ cuts the positive $x-$axis.

The solution being sought satisfies the equation $21x^2 + 8x - 5 = 0$ subject to the condition $x > 0$.

Analytical Method

Because we have a quadratic equation, the usual quadratic formula can be used

$x_{1,2} = \frac{1}{2a}\left[-b \pm \sqrt{b^2 - 4ac}\right]$; where $a = 21, \quad b = 8, \quad c = -5$.

The formula leads to the two possible answers $x_{1,2} = \frac{-8 \pm 22}{42}$.

The required analytical solution ($x > 0$) is $x_1 = \frac{-8 + 22}{42} = \frac{1}{3}$

Numerical Method

One of the popularly used numerical method is the Newton-Raphson method represented by the formula $x_{n+1} = x_n - \frac{f(x_n)}{f(x_n)}$ which we shall meet with in Module 5.

Here we have the function $f(x) = 21x^2 + 8x - 5$ whose derivative is the function $f'(x) = 42x + 8$.

Since $f(0) = -5$ and $f(1) = 24$ we conclude that the curve representing the function crosses the positive $x-$axis inside the interval $0 < x < 1$. We choose the starting value $x_0 = 0.5$ and perform **only two iterations** to get the following approximations, which are clearly getting closer and closer to the exact solution obtained above using an analytical method.

n	x_n	$f(x_n)$	$f'(x_n)$
0	0.5	4.25	29.0
1	0.353448	0.451019	22.844816
2	0.333705		

4.4 Absolute versus Relative Errors

Definitions

Let X^* be an approximation to an exact (true) quantity X. Then,

The error in X^* is defined by the quantity $X - X^*$

The absolute error in X^* is defined by the quantity $\left|X - X^*\right|$.

The relative error in X^* is given by the quantity $\left|\frac{X - X^*}{X}\right|$, and

The percentage error in X^* is defined by $100\left|\frac{X - X^*}{X}\right|\%$

Since the exact (true) value X is normally not known, one replaces it with the approximate value X^* in the denominator of the expression for the relative and percentage errors.

Precision and Accuracy

Measurements and calculations can be characterized with regard to their accuracy and precision.

Precision refers to how big or how small the absolute error $\left|X - X^*\right|$ is. The absolute error is therefore a measure of the precision of an approximation.

Accuracy refers to how closely the approximation X^* agrees with the true value X. Here, what counts is not only the magnitude of the deviation $X - X^*$ but its size relative to the true value X. Accuracy is therefore measured by the relative error $\left|\frac{X - X^*}{X}\right|$.

4.5 Sources and Types of Computational Errors

Types and Sources of Errors

We now list the sources and types of errors and follow this up with a briefly discussion of methods of eliminating or reducing such errors so that the numerical solution we get is not seriously affected by them to the extent of rendering it meaningless.

(i) Initial errors

Any mathematical problem meriting to be solved numerically involves some initial data. Such data may be in the form of coefficients in a mathematical expression or entries in a matrix. If this initial data is not exact, then the deviations from their respective true values are called **initial errors.** In some problems, uncertainties in the initial data can have devastating effect on the final numerical solution to the problem.

(ii) Truncation error

The term truncation error refers to the error introduced by a method because some infinite process is stopped prematurely (truncated) to a fewer number of terms or iterations..

Such errors are essentially algorithmic errors and one can predict the extent of the error that will occur in the method.

Specifically, the solution obtained using some numerical methods may involve infinite processes. For instance, this is the case with all convergent iteration methods and convergent infinite series. Since such infinite processes cannot be carried out indefinitely, one is forced to stop (truncate) the process and hence accept an approximate solution. The error caused though this unavoidable termination of an infinite process is called a **truncation error.**

(iii) Rounding error

Rounding errors are errors introduced during numerical calculations due to the inability of calculating devices to perform exact arithmetic. For example, if we multiply two numbers, each with six decimal digits, the product will have twelve decimal digits. Unfortunately some calculating devices may not be able to display all twelve decimal digits. In such cases we one is forced to work with fewer digits thereby necessitating dropping

some of the (less significant) digits on the right of the product. The error so introduced is called a **rounding error.**

4.6 Measures of Reducing or Eliminating Computational Errors

In the spirit of ***"prevention is better than cure"*** we shall attempt in this section to give practical suggestions of ways to eliminate or reduce the impact of various types of computational errors that are encountered in resorting to numerical methods.

(i) How to reduce Initial Errors

Initial errors can have a devastating effect on numerical solutions.

We illustrate a typical case involving an example taken from ***Francis Sheid, Numerical Analysis, Shaum Outline Series, 1968 page 342*** involving the solution of the following two simultaneous linear equations.

$$\begin{array}{lcll} x & - & y & = & 1 \\ x & - & 1.00001y & = & 0 \end{array}$$

The true (analytical) solution is $x = 100{,}001$, $y = 100{,}000$. In this example, the set of initial data consists of the elements of the coefficient matrix $A = \begin{bmatrix} 1 & -1 \\ 1 & -1.00001 \end{bmatrix}$ and the right hand side vector $\underline{b} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

However, if the entry -1.00001 in the matrix A is changed to -0.99999 while all other data items remain unchanged, the resulting system of equations

$$\begin{array}{lcll} x & - & y & = & 1 \\ x & - & 0.99999y & = & 0 \end{array}$$

has the drastically changed exact (analytical) solution $x = -99{,}999$, $y = -100{,}000$.

This surprising result demonstrates how a small change in the initial data can cause disproportionately large changes in the solution of some problems.

Thus, the only way to reduce or, if possible, to eliminate initial errors is by **making sure that all data given with or computed for use in solving a problem is as accurate as is humanly possible.**

(ii) How to reduce Truncation Errors

Truncation errors are caused by the unavoidable need to stop a convergent infinite process in efforts to get a solution. The size of the truncation error will therefore depend on the particular infinite process (numerical method) being used an on how far we are prepared to carry on with the infinite process.

The truncation error can be reduced either by

(a) Choosing a numerical method with a small truncation error or by
(b) Carrying out the infinite process sufficiently far.

Example

The continuous function $f(x) = x^2 - 3x + 1$ has a root which lies in the interval $0 < x < 1$ (Why?). Using the quadratic formula, the exact value of the root correct to six decimal places is $\rho = 0.381966$. A number of iterative methods exist for approximating such a root. Here we consider two such methods:

The bisection method

$$x_{n+1} = \frac{x_n + x_{n-1}}{2}, \text{ provided that } f(x_{n-1}) f(x_n) < 0.$$

The Newton-Raphson method

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \text{ provided that } f'(x_n) \neq 0.$$

If one performs only three iterations (truncation after three iterations) with each method using the starting values $x_0 = 0$ and $x_1 = 1$ for the bisection

method and $x_1 = 0$ for the Newton-Raphson method, one gets the following sequence of approximations for each method.

Method	**Initial Values**	x_2	x_3	x_4
Bisection	$x_0 = 0$ $x_1 = 1$	0.500000	0.250000	0.375000
Newton Raphson	$x_1 = 0$	0.333333	0.380952	0.381966

These results demonstrate that in stopping the infinite process (iteration) after the third iteration, the truncation error of the Newton Raphson method is much smaller than that of the bisection method.

(iv) How to reduce Rounding Errors

Before we discuss this important last task in our learning activity we shall first introduce a few terms that will frequently be mentioned and used in the process.

- **What are figures or digits?**

In computational mathematics, the words **"figure"** and **"digit"** are synonyms. They are used interchangeably to mean any one of the ten numerals in the set

{0, 1, 2, 3, 4, 5, 6, 7, 8, 9}.

In the decimal system of real numbers, a number N is a **string or an ordered sequence** of figures or digits A typical example is the number

$$N = 00073920600365.00004507000$$

A number can be viewed as a measure of the size or magnitude of some real or imaginary quantity. The position of each digit in the string of digits has direct bearing on the importance or significance of that digit (figure) in the overall measure of the size or magnitude of the quantity the number represents.

Intuitively we know that the leftmost digit 7 in the number N above is more significant than the rightmost digit 7.

- **Which digits in a number are significant**?

The following rules apply in deciding which digits or figures in a given number are significant.

1. Nonzero integers are always significant figures.
2. Zero digits on the leftmost part of a number are not significant.
3. All zero digits positioned between nonzero digits are significant.
4. Zeros at the rightmost end of a number are counted as significant only if the number contains a decimal point.

- **How many significant figures are in a given number?**

The number of significant figures in a given number is found using the following rules:

Rule 1: The number of significant figures in a purely integer number (with no decimal digits) is obtained by counting, starting with the leftmost nonzero digit and ending with the rightmost nonzero digit.

Example 1.3

The number 541500409 **has 9 significant figures**.
The number 002507030 **has 6 significant figures**

Rule 2: The number of significant figures in a number having a decimal part is obtained by counting all the digits, starting with the leftmost nonzero digit.

Example 1.4

The number 6.00213 **has 6 significant figures.**
The number 6.00213000 **has 9 significant figures**

NOTE: All zero digits at the end of a decimal number are significant.

(iii) How to reduce rounding errors

Armed with the concepts of digits/figures and significant figures in a number we can now comfortably discuss ways of reducing rounding errors.

One obvious method of dealing with the problem of rounding errors is to work with the maximum allowable accuracy on our calculating device at each stage in our calculations.

Example 1.5

Find the sum of the numbers 2.35, 1.48, 4.24 using a calculating device that can only handle numbers with two significant figures.

The exact um is $S = 2.35 + 1.48 + 4.24 = 8.07$

If we neglect the second decimal digit from each term and form their sum we find the approximate sum $S_1 = 2.3 + 1.4 + 4.2 = 7.9$

The absolute error in S_1 is: $|S - S_1| = 0.17$

A better approximation of S under the same limitations is

$$S_2 = 2.4 + 1.5 + 4.2 = 8.1$$

The absolute error in S_2 is: $|S - S_2| = 0.03$

This error is significantly smaller than that in S_1.

The immediate question expected to be raised by the learner is ***"How did one arrive at the two digit terms in the sum*** S_2 ***?"***

The answer to the above question is simple. Each term has been obtained from its corresponding three-digit term by rounding.

What does it mean to round a number?

To round a number to a fixed number of figures or digits simply means leaving out (dropping) all digits on the right hand side of the number beyond a certain position.

If a number is rounded simply by dropping all digits beyond a certain position on the right hand side of the number without making any adjustments to the last retained digit, then one speaks of "**rounding off or chopping the number".**

Example 1.6

The sum S_1 has been calculated using terms obtained from the original numbers by rounding off (chopping) the third decimal digit from each term. The term 2.35 was rounded to 2.3, the term 1.48 was rounded to 1.4 and the term 4.24 was rounded to 4.2. In each case, the last retained digit (the first decimal place) has not been adjusted in the process of rounding.

Note

The sum S_2 has also been obtained through rounding. However, the rounding this time is different. Here, not all the three terms have been rounded off!

The term	2.35	has been rounded to	2.4
The term	1.48	has been rounded to	1.5
The term	4.24	has been rounded to	4.2

We observe that in rounding each of the first two terms 2.35 and 1.48, the digit occupying the second decimal position has been dropped but the digit occupying the first decimal position has been adjusted by increasing it by one (unity). The third digit 4.24 has simply been rounded off.

This practice (or as yet unknown rule for rounding numbers) seems to have some significant advantage over rounding off manifested by the above example in which S_2 is more accurate than S_1.

Rules for rounding numbers

In order to reduce the error in rounding numbers, the rejection of digits beyond some predetermined position (n) is accompanied by making adjustments to the digit retained in position $(n-1)$. The adjustment involved either leaving the digit in position (n) unchanged or increasing it by one (unity). The decision to retain or increase by 1 the digit occupying position $(n-1)$ is governed by the following rules.

(a) If the digit in position $(n+1)$ is **greater than** 5 then the digit in position (n) is **increased by 1**.

(b) If the digit in position $(n+1)$ is 5 and at least one other digit to its right is non zero then the digit in position (n) is increased by 1.

(c) If the digit in position $(n+1)$ is **less than** 5 then the digit in position (n) is **left unchanged**.

(d) If the digit in position **(** $n+1$ **)** is 5 and **all other digits to the right of** 5 **are zero**, then

(i) The digit in position **(** n **)** is **increased by 1** if it is an **odd** number $(1,3,5,7,9)$;

(ii) The digit in position **(** n **)** is **retained unchanged** if it is an **even** number $(0,2,4,6,8)$.

Example1.7

Rounding a given number correct to two significant figures

S/N	Number	Rounded to 2 Significant figures	Rule Used
1	8.361	8.4	(a)
2	8.351	8.4	(b)
3	8.350	8.4	(d) (i)
4	8.450	8.4	(d) (ii)
5	8.050	8.0	(d) (ii)
6	8.349	8.3	(c)
7	2.55	2.6	(d) (i)
8	2.65	2.6	(d) (ii)
9	0.0557	0.056	(a)
10	0.0554	0.055	(b)

4.7 Summary

We have covered a lot of ground in highlighting the possible pitfalls in carrying out numerical calculations. The main reason in doing so was to be aware of the possible sources of errors and how either to avoid them or manage them in such a way that errors do not accumulate and render our final numerical result useless. Having an idea of the errors and taking measures to reduce them enhances our confidence in accepting a numerical solution as a reasonable approximation to the solution we seek to find.

Competence in this regard is measured not only by the knowledge we have presented and hopefully acquired, but in taking concrete measures as described here in eliminating or reducing the possible errors that can encroach into our numerical calculations.