Wolaita Sodo University
College of Natural & Computational Science
Department of Mathematics

Numerical Analysis I

Prepared By
Biruk Endeshaw Mekonnen

March 8, 2021

Chapter 1

Basic Concepts in Error Estimation

1.1 INTRODUCTION

Numerical technique is widely used by scientists and engineers to solve their problems. A major advantage for numerical technique is that a numerical answer can be obtained even when a problem has no analytical solution. However, result from numerical analysis is an approximation, in general, which can be made as accurate as desired. The reliability of the numerical result will depend on an error estimate or bound, therefore the analysis of error and the sources of error in numerical methods is also a critically important part of the study of numerical technique.

1.2 ACCURACY OF NUMBERS

(i) Exact Number: Number with which no uncertainly is associated to no approximation is taken, are known as exact numbers.
e.g., $5, 21/6, 12/3, \ldots$ etc. are exact numbers.

(ii) Approximate Number: There are numbers, which are not exact.
e.g., $\sqrt{2} = 1.41421...,e = 2.7183....,$ etc. are not exact numbers since they contain infinitely many non-recurring digits. Therefore the numbers obtained by retaining a few digits, are called approximates numbers, e.g., $3.142, 2.718$ are the approximate values of π and e.

(iii) Significant digits: The significant digits are the number of digits used to express a number. The digits $1, 2, 3, 4, 5, 6, 7, 8, 9$ are significant digits. $'0'$ is also a significant figure except when it is used to fix the decimal point or to fill the places of unknown or discarded digits.

For example, each number $5879, 3.487, 0.4762$ contains four significant digits while the numbers $0.00486, 0.000382, 0.0000376$ contains only three significant digits since zeros only help to fix the position of the decimal point.

Similarly, in the number 0.0002070, the first four $'0'$s are not significant digits since they serve only to fix the position of decimal point and indicate the place values of the other digits. The other two $'0'$s are significant.
To be more clear, the number 2.0683 contains five significant digits.

A. The significant figure in a number **in positional notation** consists of

(i) All non-zero digits

(ii) Zero digits which

(a) lie between significant digits;

(b) lie to the right of decimal point and at the same time to the right of a non-zero digit;

(c) are specifically indicated to be significant.

B. The significant figure in a number written **in scientific notation** $(e.g., M \times 10^k)$ consists of all the digits explicitly in M.

3900	two
39.0	two
3.9×10^6	two

Note 1.1. *Significant digits are counted from left to right starting with the non-zero digit on the left.*

(iv) Round off Numbers: There are numbers with many digits, **e.g.,** $\frac{22}{7} = 3.142857143$. In practical, it is desirable to limit such numbers to a manageable number of digits, such as 3.14 or 3.143. This process of dropping unwanted digits is called **rounding-off**. Number are rounded-off according to the following rules:

To round-off a number to n significant digits, discard all digits to the right of nth digit and if this discarded number is

1. Less than 5 in $(n + 1)$th place, leave the nth digit unaltered
   e.g., 8.893 to 8.89
2. Greater than 5 in $(n + 1)$th place, increase the nth digit by unity
   e.g., 5.3456 to 5.346
3. Exactly 5 in $(n+1)$th place, increase the nth digit by unity if it is odd otherwise leave it unchanged.
   e.g., 11.675 to 11.68, 11.685 to 11.68.

Example 1.1. *Round-off the following numbers correct to four significant figures: 58.3643, 979.267, 7.7265, 56.395, 0.065738 and 7326853000.*

Solution: After retaining first four significant figures we have:

a) 58.3643 becomes 58.36

b) 979.267 becomes 979.3

c) 7.7265 becomes 7.726 (digit in the fourth place is even)

d) 56.395 becomes 56.40 (digit in the fourth place is odd)

e) 0.065738 becomes 0.06574 (because zero in the left is not significant)

f) 7326853000 becomes 7327×10^6.

1.3 ERRORS

A computer has a finite word length and so only a fixed number of digits are stored and used during computation. This would mean that even in storing an exact decimal number in its converted form in the computer memory, an error is introduced. This error is machine dependent and is called machine epsilon. After the computation is over, the result in the machine form (with base b) is again converted to decimal form understandable to the users and some more error may be introduced at this stage. In general, we can say that

$$\boxed{\textbf{Error} = \textbf{Exact value - Approximate value}}$$

The errors may be divided into the following different types:

(1) Inherent Error: Errors which are already present in the statement of a problem before its solution are called inherent errors. Such errors arises either due to the fact that the given data is approximate or due to limitations of mathematical tables, calculators, or the digital computer.

Inherent error can be minimized by taking better data, by using high precision computing aids and by correcting obvious errors in the data. Accuracy refers to the number of significant digits in a value, for example, 53.965 is accurate to 5 significant digits.

Precision refers to the number of decimal positions or order of magnitude of the last digit in the value. For example, in 53.965, precision is 10^{-3}.

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

4.3201, 4.32, 4.320106.

Sol. In 4.3201, precision is 10^{-4}
In 4.32, precision is 10^{-2}
In 4.320106, precision is 10^{-6}.
Hence, the number 4.320106 has the greatest precision.

(2) Round-off Error: Rounding errors arise from the process of rounding-off numbers during the computation. They are also called procedural errors or numerical errors. Such errors are unavoidable in most of the calculations due to limitations of computing aids.

The round-off error can be reduced by carrying the computation to more significant figures at each step of computation. At each step of computations, retain at least one more significant figure than that given in the data, perform the last operation, and then round off.

(3) Truncation Error: Truncation errors are caused by using approximate results or by replacing an infinite process with a finite one.

If we are using a decimal computer having a fixed word length of 4 digits, rounding-off of 13.658 gives 13.66, whereas truncation gives 13.65.

e.g., If $S = \sum_{i=1}^{\infty} a_i x_i$ is replaced by or truncated to $S = \sum_{i=1}^{n} a_i x_i$, then the error developed is a truncation error.

A truncation error is a type of algorithm error. Also,

if $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots \infty = X$ (say) is truncated to

$1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} = X'$ (say), then truncation error $= X - X'$

(4) Absolute Error: Absolute error is the numerical difference between the exact or true value of a quantity and its approximate value.

Thus, if X is the true value of a quantity and X' is its approximate value, then $|X - X'|$ is called the absolute error and denoted by E_a. Therefore

$$E_a = |X - X'|$$

(5) Relative Error: The relative error E_r is defined by

$$E_r = \left|\frac{X - X'}{X}\right| = \frac{E_a}{\text{True value}}$$

(6) Percentage error: The percentage error E_r is defined as

$$E_p = 100 \times E_r = 100 \times \left|\frac{X - X'}{X}\right|$$

Note 1.2. 1. *The relative and percentage errors are independent of units used while absolute error is expressed in terms of these units.*

2. *If a number is correct to n decimal places, then the error*

$= \frac{1}{2}(10^{-n})$.

e.g., if the number 3.1416 is correct to 4 decimal places, then the error

$= \frac{1}{2}(10^{-4}) = 0.00005$.

3. *If the first significant digit of a number is k and the number is correct to n significant digits, then the relative error $< \frac{1}{(k \times 10^{n-1})}$.*

Example 1.2. *Suppose 1.414 is used as an approximation to $\sqrt{2}$. Find the absolute and relative errors.*

Solution: True value $\sqrt{2} = 1.41421356$

Approximate value $= 1.414$

$$\begin{aligned}
\text{Error} &= \text{True value} - \text{Approximate value} \\
&= \sqrt{2} - 1.414 = 1.41421356 - 1.414 \\
&= 0.00021356 \\
\text{Absolute error } E_a &= |\text{Error}| \\
&= |0.00021356| = 0.21356 \times 10^{-3}
\end{aligned}$$

$$\text{Relative error } E_r = \frac{E_a}{\text{True value}} = \frac{0.21356 \times 10^{-3}}{\sqrt{2}}$$
$$= 0.151 \times 10^{-3}.$$

Example 1.3. *If* 0.333 *is the approximate value of* $\frac{1}{3}$, *find the absolute, relative, and percentage errors.*

Solution: Given that True value $(X) = \frac{1}{3}$, and its Approximate value $(X') = 0.333$
Therefore, Absolute Error, $E_a = |X - X'|$
$= |\frac{1}{3} - 0.333| = |0.333333 - 0.333| = 0.000333$
Relative Error, $E_r = \frac{0.000333}{0.333333} = 0.000999$ and
Percentage Error, $E_p = E_r \times 100 = 0.000999 \times 100 = 0.099\%$.

Example 1.4. *Calculate the sum of* $\sqrt{3}, \sqrt{5}$ *and* $\sqrt{7}$ *to four significant digits and find its absolute and relative errors.*

Solution: Here $\sqrt{3} = 1.732$, $\sqrt{5} = 2.236$, $\sqrt{7} = 2.646$
Hence Sum $= 6.614$ and
Absolute error $= E_a = 0.0005 + 0.0005 + 0.0005 = 0.0015$
(Because $\frac{1}{2} \times 10^{-3} = 0.0005$). Also the total absolute error shows that the sum is correct up to 3 significant figures. Therefore $S = 6.61$ and
Relative Error, $E_r = \frac{0.0015}{6.61} = 0.0002$.

Exercise 1.1. *An approximate value of* π *is given by* 3.1428571 *and its true value is* 3.1415926. *Find the absolute and relative errors.*

Exercise 1.2. *Calculate the value of* $\sqrt{102} - \sqrt{101}$ *correct to four significant digits.*

1.4 A GENERAL FORMULA FOR ERROR

Let $y = f(x_1, x_2)$ be a function of two variables x_1 and x_2. To determined the error Δy in y due to the errors $\Delta x_1, \Delta x_2$ in x_1, x_2 respectively.
$$y + \Delta y = f(x_1 + \Delta x_1,\ x_2 + \Delta x_2)$$
Using Taylor's series for two variables, to expand the R.H.S of above, we get

$$y + \Delta y = f(x_1, x_2) + \left(\frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2\right) + \frac{1}{2}\left[\frac{\partial^2 f}{\partial x_1^2}(\Delta x_1)^2 + 2\frac{\partial^2 f}{\partial x_1 x_2}\Delta x_1 \Delta x_2 + \frac{\partial^2 f}{\partial x_2^2}(\Delta x_2)^2\right] + \cdots$$

Errors $\Delta x_1, \Delta x_2$ are small so that the terms containing $(\Delta x_1)^2, (\Delta x_2)^2$ and higher powers of $\Delta x_1, \Delta x_2$ are being neglected. Therefore

$$y + \Delta y = f(x_1, x_2) + \frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2$$
$$\Delta y = \frac{\partial y}{\partial x_1}\Delta x_1 + \frac{\partial y}{\partial x_2}\Delta x_2$$

Because $y = f(x_1, x_2)$.
In general, the error Δy in the function $y = f(x_1, x_2, ..., x_n)$ corresponding to the errors Δx_i in

$x_i (i = 1, 2, \ldots\ldots, n)$ is given by

$$\Delta y = \frac{\partial y}{\partial x_1}\Delta x_1 + \frac{\partial y}{\partial x_2}\Delta x_2 + \ldots + \frac{\partial y}{\partial x_n}\Delta x_n \tag{1.1}$$

Equation (1.1) represents the general formula for Errors. If equation (1.1) divided by y we get relative error

$$E_r = \frac{\Delta y}{y} = \frac{\partial y}{\partial x_1}\frac{\Delta x_1}{y} + \frac{\partial y}{\partial x_2}\frac{\Delta x_2}{y} + \cdots + \frac{\partial y}{\partial x_n}\frac{\Delta x_n}{y}$$

On taking modulus both of the sides, we get maximum relative error.

$$\left|\frac{\Delta y}{y}\right| \leq \left|\frac{\partial y}{\partial x_1}\frac{\Delta x_1}{y}\right| + \left|\frac{\partial y}{\partial x_2}\frac{\Delta x_2}{y}\right| + \ldots + \left|\frac{\partial y}{\partial x_n}\frac{\Delta x_n}{y}\right|$$

Also from equation (1.1), by taking modulus we get maximum absolute error.

$$|\Delta y| \leq \left|\frac{\partial y}{\partial x_1}\Delta x_1\right| + \left|\frac{\partial y}{\partial x_2}\Delta x_2\right| + \ldots + \left|\frac{\partial y}{\partial x_n}\Delta x_n\right|$$

1.5 ERRORS IN NUMERICAL COMPUTATIONS

1.5.1 Error in addition of numbers

Let $X = x_1 + x_2 + \ldots\ldots + x_n$

$$\begin{aligned} X + \Delta X &= (x_1 + \Delta x_1) + (x_2 + \Delta x_2) + \ldots\ldots + (x_n + \Delta x_n) \\ &= (x_1 + x_2 + \ldots\ldots + x_n) + (\Delta x_1 + \Delta x_2 + \ldots\ldots + \Delta x_n) \end{aligned}$$

Therefore, $\Delta X = \Delta x_1 + \Delta x_2 + \ldots\ldots + \Delta x_n$; this is an absolute error.

Dividing by X we get, $\frac{\Delta X}{X} = \frac{\Delta x_1}{X} + \frac{\Delta x_2}{X} + \ldots\ldots + \frac{\Delta x_n}{X}$; which is a relative error.

The maximum relative error is

$$\left|\frac{\Delta X}{X}\right| \leq \left|\frac{\Delta x_1}{X}\right| + \left|\frac{\Delta x_2}{X}\right| + \ldots\ldots + \left|\frac{\Delta x_n}{X}\right|.$$

Therefore it shows that when the given numbers are added then the magnitude of absolute error in the result is the sum of the magnitudes of the absolute errors in that numbers.

1.5.2 Error in subtraction of numbers

Let $X = x_1 - x_2$

$$\begin{aligned} X + \Delta X &= (x_1 + \Delta x_1) - (x_2 + \Delta x_2) \\ &= (x_1 - x_2) + (\Delta x_1 - \Delta x_2) \end{aligned}$$

Therefore $\Delta X = \Delta x_1 - \Delta x_2$ is the Absolute error

and $\frac{\Delta X}{X} = \frac{\Delta x_1}{X} - \frac{\Delta x_2}{X}$ is the Relative error.

But we know that $|\Delta X| \leq |\Delta x_1| + |\Delta x_2|$ and $\left|\frac{\Delta X}{X}\right| \leq \left|\frac{\Delta x_1}{X}\right| + \left|\frac{\Delta x_2}{X}\right|$ therefore on taking modulus of relative errors and absolute errors to get its maximum value, we have $|\Delta X| \leq |\Delta x_1| + |\Delta x_2|$ which is the maximum absolute error and $\left|\frac{\Delta X}{X}\right| \leq \left|\frac{\Delta x_1}{X}\right| + \left|\frac{\Delta x_2}{X}\right|$ which gives the maximum relative error in subtraction of numbers.

1.5.3 Error in product of numbers

Let $X = x_1 x_2 \cdots x_n$ then using general formula for error

$$\Delta X = \frac{\partial X}{\partial x_1}\Delta x_1 + \frac{\partial X}{\partial x_2}\Delta x_2 + \ldots + \frac{\partial X}{\partial x_n}\Delta x_n$$

We have $\frac{\Delta X}{X} = \frac{1}{X}\frac{\partial X}{\partial x_1}\Delta x_1 + \frac{1}{X}\frac{\partial X}{\partial x_2}\Delta x_2 + \ldots + \frac{1}{X}\frac{\partial X}{\partial x_n}\Delta x_n$

Now,

$$\frac{1}{X}\frac{\partial X}{\partial x_1} = \frac{x_2 x_3 \cdots\cdots x_n}{x_1 x_2 \cdots\cdots x_n} = \frac{1}{x_1}$$
$$\frac{1}{X}\frac{\partial X}{\partial x_2} = \frac{x_1 x_3 \cdots\cdots x_n}{x_1 x_2 \cdots\cdots x_n} = \frac{1}{x_2}$$
$$\cdots\cdots$$
$$\frac{1}{X}\frac{\partial X}{\partial x_n} = \frac{x_1 x_2 \cdots\cdots x_{n-1}}{x_1 x_2 \cdots\cdots x_{n-1} x_n} = \frac{1}{x_n}$$

Therefore $\frac{\Delta X}{X} = \frac{\Delta x_1}{x_1} + \frac{\Delta x_2}{x_2} + \ldots\ldots + \frac{\Delta x_n}{x_n}$

Therefore maximum Relative and Absolute errors are given by

Relative Error $= \left|\frac{\Delta X}{X}\right| \le \left|\frac{\Delta x_1}{x_1}\right| + \left|\frac{\Delta x_2}{x_2}\right| + \ldots + \left|\frac{\Delta x_n}{x_n}\right|$

Absolute Error $= \left|\frac{\Delta X}{X}\right| X = \left|\frac{\Delta X}{X}\right| \cdot (x_1 x_2 \cdots x_n)$

1.5.4 Error in division of numbers

Let $X = \frac{x_1}{x_2}$ then again using general formula for error

$$\Delta X = \frac{\partial X}{\partial x_1}\Delta x_1 + \frac{\partial X}{\partial x_2}\Delta x_2 + \ldots + \frac{\partial X}{\partial x_n}\Delta x_n$$

We have

$$\frac{\Delta X}{X} = \frac{1}{X}\cdot\frac{\partial X}{\partial x_1}\Delta x_1 + \frac{1}{X}\cdot\frac{\partial X}{\partial x_2}\cdot\Delta x_2$$
$$= \frac{\Delta x_1}{\left(\frac{x_1}{x_2}\right)}\cdot\frac{1}{x_2} + \frac{\Delta x_2}{\left(\frac{x_1}{x_2}\right)}\left(\frac{-x_1}{x_2^2}\right)$$
$$= \frac{\Delta x_1}{x_1} - \frac{\Delta x_2}{x_2}$$

Therefore $\left|\frac{\Delta X}{X}\right| \le \left|\frac{\Delta x_1}{x_1}\right| + \left|\frac{\Delta x_2}{x_2}\right|$ which is relative error.

Absolute Error $= |\Delta X| \le \left|\frac{\Delta X}{X}\right| X$

1.5.5 Error in evaluating x^k

Let $X = x^k$, where k is an integer or fraction

$$\Delta X = \frac{\partial X}{\partial x}\Delta X = k x^{k-1}\Delta x$$
$$\frac{\Delta X}{X} = k\frac{\Delta x}{x}$$
$$\therefore \quad \left|\frac{\Delta X}{X}\right| \le k\frac{\Delta x}{x}$$

The relative error in evaluating $x^k = k\left|\frac{\Delta x}{x}\right|$

Example 1.5. *If* $u = \frac{4x^2y^3}{z^4}$ *and errors in* x, y, z *be* 0.001, *compute the relative maximum error in* u *when* $x = y = z = 1$.

Solution: We know $\Delta u = \frac{\partial u}{\partial x}\Delta x + \frac{\partial u}{\partial y}\Delta y + \frac{\partial u}{\partial z}\Delta z$

Since $\quad \frac{\partial u}{\partial x} = \frac{8xy^3}{z^4}, \quad \frac{\partial u}{\partial y} = \frac{12x^2y^2}{z^4}, \quad \frac{\partial u}{\partial z} = \frac{-16x^2y^3}{z^5}$

Also the errors $\Delta x, \Delta y, \Delta z$ may be positive or negative, we take the absolute value of terms on R.H.S. is,

$$\begin{aligned}(\Delta u)_{max.} &= \left|\frac{8xy^3}{z^4}\Delta x\right| + \left|\frac{12x^2y^2}{z^4}\Delta y\right| + \left|\frac{-16x^2y^3}{z^5}\Delta z\right| \\ &= 8(0.001) + 12(0.001) + 16(0.001) = 0.036\end{aligned}$$

Also, Max. Relative error $= \frac{\Delta u}{u} = \frac{0.036}{4} = 0.009$

Example 1.6. *If* $u = 2V^6 - 5V$, *find the percentage error in* u *at* $V = 1$, *if error in* V *is* 0.05.

Solution: Given

$$\begin{aligned} u &= 2V^6 - 5V \\ \Delta u &= \frac{\partial u}{\partial V}\Delta V = (12V^5 - 5)\Delta V \\ \frac{\partial u}{u} \times 100 &= \left(\frac{12V^5-5}{6V^6-5V}\right)\Delta V \times 100 \\ &= \left(\frac{12-5}{2-5}\right) \times 0.05 \times 100 = -\frac{7}{3} \times 5 = -11.66\% \end{aligned}$$

Hence maximum percentage error $(E_p)_{max} = 11.667\%$.

1.6 SOME IMPORTANT MATHEMATICAL PRELIMINARIES

There are some important mathematical preliminaries given below which would be useful in numerical computation.

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

(b) Intermediate value Theorem: Let $f(x)$ be continuous in $a \leq x \leq b$ and let any number between $f(a)$ and $f(b)$, then there exists a number d in $a < x < b$ such that $f(d) = l$.

(c) Mean Value Theorem for Derivatives If $f(x)$ is continuous in $[a, b]$ and $f'(x)$ exists in (a, b) then there exists at least one value of x, say d, between a and b such that, $f'(d) = \frac{f(b)-f(a)}{b-a}, a < d < b$

(d) Rolle's Theorem: If $f(x)$ is continuous in $a \leq x \leq b, f'(x)$ exists in $a < x < b$ and $f(a) = f(b) = 0$ then there exists at least one value of x, say d, such that $f'(d) = 0, a < d < b$

(e) Generalized Form of Rolle's Theorem: If $f(x)$ is n times differentiable on $a \leq x \leq b$ and $f(x)$ vanishes at the $(n + 1)$ distinct points $x_0, x_1, x_2, \ldots, x_n$ in (a, b), then there exists a number d in $a < x < b$ such that $f^n(d) = 0$.

(f) 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

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

where $R_n(x)$, the remainder term, can be expressed in the form
$R_n(x) = \frac{(x-a)^n}{n!}f^n(d), a < d < x$.

(g) Maclaurin's Expansion:

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

(h) Taylor's Series for a Function of Two Variables:

$$\begin{aligned} f(x_1 + \Delta x_1, x_2 + \Delta x_2) = f(x_1, x_2) &+ \frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2 \\ &+ \frac{1}{2}\left[\frac{\partial^2 f}{\partial x_1^2}(\Delta x_1)^2 + 2\frac{\partial^2 f}{\partial x_1 \partial x_2}\Delta x_1 \Delta x_2 + \frac{\partial^2 f}{\partial x_2^2}(\Delta x_2)^2\right] + \ldots \end{aligned}$$

PROBLEM SET 1.1

1. Prove that the relative error of a product of three non-zero numbers does not exceed the sum of the relative errors of the given numbers.

2. Find the absolute, relative and percentage errors of the approximations as

$$(a)\ \frac{1}{11} \approx 0.1 \qquad (b)\ \frac{1}{11} \approx 0.00 \qquad (c)\ \frac{5}{9} \approx 0.56 \qquad (d)\ \frac{4}{9} \approx 0.44$$

3. If $S = 4x^2y^3z^{-4}$, find the maximum absolute error and maximum relative errors in S. When errors in $x = 1, y = 2, z = 3$ respectively are equal to $0.001, 0.002, 0.003$. [**Ans.** 0.0035, 0.0089]

4. Find the percentage error if the number 5007932 is approximated to four significant figures. [**Ans.** 0.018%]

5. Compute the relative maximum error in the function $u = 7\frac{xy}{x^2}$, when $x = y = z = 1$ and errors in x, y, z be 0.001. [**Ans.** 0.006]