Approximations and Errors in Computing

INTRODUCTION

Approximations and errors are integral part of human being. They are everywhere and unavoidable. This is more in the life of computational scientist.

While using Numerical methods, it is impossible to ignore the numerical errors. Errors come in a variety of forms and sizes. Some errors are avoidable and some are not. For example, data conversion and round off errors can not be avoided, but human error can be eliminated completely. Although certain errors cannot be eliminated completely, we must at least know the bounds of these errors to make use of our final solutions. It is therefore essential to know how errors arise, how they grow during the numerical process, and how they affect the accuracy of a solution.

By careful analysis and proper design and implementation of algorithm, we can restrict their effect quite significantly.

As mentioned earlier, a number of different types of errors occur during the process of numerical computing. All these errors contribute to the total error in final result. A classification of errors encountered in a numerical process is given in a figure below which shows that every stage of the numerical computing cycle contributes to the total error.

Although perfection is what we Attempt for, it is rarely achieved in practice due to a variety of factors. But that must not deter our attempts to achieve near perfection. In this chapter 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 solution process.
EXACT AND APPROXIMATION NUMBERS

There are two kinds of numbers, exact and approximate numbers. For example numbers like 1, 2, 3, ... 1/2, 3/2, \( \sqrt{2} \), \( \pi \), e etc are exact numbers. Approximate numbers are those that represent the numbers to a certain degree of accuracy. We know that all computers operate with a fixed length of numbers. In particular, we have seen that floating point representation requires the mantissa to be a specified number of digits. Some numbers cannot be represented exactly in a given number of decimal digits. For example the quantity \( \pi \) is equal to

\[
3.1415926535897932384642\ldots
\]

Such numbers can never be represented accurately. We may write as 3.14, 3.14159, or 3.141592653. In all cases we have omitted some digits.
Note that transcendental numbers like $\pi$, $e$ and irrational number like $\sqrt{2}$, $\sqrt{5}$ do not have a terminating representation. Some rational numbers also have a repeating pattern. For instance, the rational number $2/7 = 0.285714285714\ldots$

**SIGNIFICANT DIGITS OR SIGNIFICANT FIGURES**

The concept of significant digits has been introduced primarily to indicate the accuracy of the numerical values. The digits that are used to express a number are called the *significant digits or significant figures*. Thus, the number $3.1416$, $0.36567$ and $4.0345$ contain five significant digits each. The number $0.00345$ has thee significant digits, viz, 3, 4, and 5, since zeros serve only to fix the position of the decimal point. How ever in the number $453,000$, the number of significant digits is uncertain, whereas the numbers $4.53 \times 10^5$, $4.530 \times 10^5$ and $4.5300 \times 10^5$ have three, four and five significant figures respectively.

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.
3. Trailing zeros following a decimal point are significant. For example, $3.50$, $65.0$, and $0.230$ have three significant digits.
4. Zeros between the decimal point and preceding non-zero digits are not significant. For example, the following numbers have only four significant digits.
   - $0.0001234 \ (= 1234 \times 10^{-7})$
   - $0.001234 \ (= 1234 \times 10^{-6})$
   - $0.01234 \ (= 1234 \times 10^{-5})$
5. When the decimal point is not written, the trailing zeros are not considered to be significant.

Integer numbers with trailing zeros may be written in scientific notation to specify the significant digits.

More examples:

1. $96.763$ has five significant digits.
2. $0.008472$ has four significant digits.
3. $0.0456000$ has six significant digits.
4. $36$ has two significant digits.
5. For $3600$, the number of significant digits is uncertain.
6. $3600.00$ has six significant digits. Note that zeros were made significant by writing .00 after $3600$.

**Accuracy and Precision**

The concept of *accuracy* and *precision* are closely related to significant digits. Precision refers to the reproducibility of results and measurements in an experiment; while accuracy refers to how close the value is to the actual or true value. Results could be both precise and accurate, neither precise nor accurate, precise and not accurate, or vice
versa. The validity of the results increases as they are more accurate and precise. They are related as follows:

1. Accuracy refers to the number of significant digits in a value. For example, the number 46.395 is accurate to five significant digits.
2. Precision refers to the number of decimal positions, i.e. the order of magnitude of the last digit in a value. The number 45.679 has precision of 0.001 or $10^{-3}$.

**Example:** Which of the following number has greatest precision?

- (a) 4.3201
- (b) 4.32
- (c) 4.320106

**Answer:**
- (a) 4.3201 has precision of $10^{-4}$.
- (b) 4.32 has precision of $10^{-2}$.
- (c) 4.320106 has precision of $10^{-6}$.

The last number has the greatest precision.

**INHERENT ERRORS**

*Inherent errors* are those that present in the data supplied to the model. Inherent errors contain two components, namely, *data errors* and *conversion errors*.

**Data Errors**

*Data errors* arise when data for a problem are obtained by some experiment 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 distance, a voltage, or a time period cannot be exact.

**Conversion error**

*Conversion errors* (also know as representation errors) arises due to the limitation of the computer to store data exactly. We know that the floating point representation retains only specific number of digits. The digits that are not retained produce the round off error.

As we have seen already many numbers cannot be represented exactly in a given number of decimal digits. In some cases a decimal number cannot be represented exactly in binary form. For example, the decimal number 0.1 has a non-terminating binary form like 0.00011001100110011… … but computer retains only specific number of bits.

**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 the implementing the solution.

**Round-off Error**

*Round-off errors* occur when a fixed number of digits are used to represent exact numbers. Since the numbers are stored at every stage of computation, round-off error is introduced at the end of every arithmetic operation. Consequently, even though an individual round-off error could be very small, the cumulative effect of a series of computation can be very significant. It is usual to round-off numbers according to the following rule:

To round-off a number to \( n \) significant digits, discard all digits to the right of the \( n \)th digit, if the first discarded digit is

1. greater than 5, the last retained significant digit is “rounded up” by 1.
2. less than 5, keep the last retained significant digit unchanged.
3. exactly 5, “rounded up” the last retained digit by 1 if it is odd; otherwise, leave it unchanged.

The number thus rounded-off is said to be correct to \( n \) significant digits.

Examples: Following number are rounded-off to four significant digits:

- 2.64570 to 2.646
- 12.0354 to 12.04
- 0.547326 to 0.5473
- 3.24152 to 3.242

In manual computation, the round-off error can be reduced by carrying out computations to more significant figures at each step of the computation. A usual way to do that is: at each step of the computation, retain at least one more significant digit than the given data, perform the last operation and then round-off.

**Truncation Error**

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

\[ S = \sum_{i=0}^{\infty} a_i x_i \] is replaced by the finite sum \[ S = \sum_{i=0}^{n} a_i x_i . \]

The series has been truncated.
Consider the following infinite series expansion of the \( \sin x \):

\[
\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots
\]

When we calculate the sine of an angle using this series, we cannot use all the terms in the series for the computation. We usually terminate the process after a certain term is calculated. The terms “Truncated” introduce an error which is called truncation error. The truncation error can be reduced by using better numerical method which usually increases the number of arithmetic operations.

**MODELLING ERRORS**

Mathematical models are the basis for numerical solutions. They are formulated to represent physical process using certain parameters involved in the situation. In many situations, it is impractical or impossible to include the entire real problem and, therefore, certain simplifying assumptions are made. For example, while developing a model for calculating the force acting on a falling body, we may not be able to estimate the air resistance coefficient (drag coefficient) properly or determine the direction and magnitude of wind force acting on the body, and so on. To simplify the model, we may assume that the force acting due to air resistance is linearly proportional to the velocity of the falling body or we may assume that there is no wind force acting on the body. All such simplifications certainly result in errors in the output from such models.

Since mathematical model is the basis of the numerical process, no numerical method will provide adequate results if the model is erroneously formulated. The modeling errors can be reduced significantly by refining or enlarging the models by incorporating missing features in the model. But this enhancement might make the model more complex and might be impossible to solve numerically or might take enough time to implement the solution process. It is not always true that an enhanced model will provide a better result. We must note that modeling, data quality and computation go hand in hand. An overly refined model with inaccurate data or an inadequate computer may not be meaningful. On the other hand, an oversimplified model may produce a result that is unacceptable. It is, therefore, necessary to keep a balance between the level of accuracy and the complexity of the model. A model must incorporate those features that are essential to reduce the error to an acceptable level.

**BLUNDERS**

Blunders are errors that are due to human imperfection. As the name indicates, such errors may cause a serious disaster in the result. Since these errors are due to human mistakes, it should be possible to avoid them to large extent by acquiring good knowledge of all the aspects of the problem as well as numerical process.

Human errors can occur at any stage of the numerical processing cycle. Some common types of errors are:

1. lack of understanding of the problem.
2. wrong assumption while formulating a model.
3. errors in deriving the mathematical model that does not describe adequately the physical system under study.
4. selecting a wrong numerical method for solving the mathematical model.
5. selecting a wrong algorithm for implementing the numerical method.
6. making mistake in the computer program.
7. mistakes in data input, such as misprint, giving values column-wise instead row-wise to a matrix, forgetting a negative sign, etc.
8. wrong guessing of initial value.

As mentioned earlier, all these mistakes can be avoided through a reasonable understanding of the problem and the numerical solution method, and use of good programming techniques and tools.

### Absolute, Relative and Percentage Errors

Let us now consider some fundamental definitions of error analysis. Regardless of its source, an error is usually quantified in two different but related ways. One is known as **absolute error** and other is called **relative error**.

Let $X$ denotes the true value of a data item and $X_1$ is its approximated value. Then, these two quantities are related as

\[
\text{True value} = \text{Approximate value} + \text{Error}
\]

\[
i.e \quad X = X_1 + E
\]

\[
or, \quad E = X - X_1
\]

The error may be negative or positive depending on the values of $X$ and $X_1$. In error analysis, what is important is the magnitude of the error but not the sign of the error, and therefore, we normally consider what is known as **absolute error** which is denoted by $E_A$ and given by

\[
E_A = |X - X_1|
\]

In many cases, absolute error may not reflect its influence correctly as it does not take into account the order of the magnitude of the value under study. For example, an error of 1 gram is much more significant in the weight of 10 gram of gold chain than in the weight of a bag of rice. In the view of this, we introduce the concept of **relative error** which is nothing but the “normalized” absolute error. The relative error is denoted by $E_R$ and defined by

\[
E_R = \frac{E_A}{|X|} = \left| \frac{X - X_1}{X} \right| = \left| 1 - \frac{X_1}{X} \right|
\]

And the relative percentage error is given by

\[
E_p = E_R \times 100\%
\]
Limiting Absolute Error

Let \( \Delta X > 0 \) be such a number such that \( |X - X_i| \leq \Delta X \), i.e. \( |E_x| \leq \Delta X \). Then, \( \Delta X \) is an upper limit on the magnitude of the absolute error and is said to measure absolute accuracy.

Similarly, \( \frac{\Delta X}{X} \approx \frac{\Delta Y}{Y} \) measures the relative accuracy.

Remark: If the number \( X \) is rounded to \( N \) decimal places, then the absolute error does not exceed the amount \( \Delta X = \frac{1}{2} (10^{-N}) \).

Example: If the number \( X = 1.325 \) is correct to three decimal places, then limiting absolute error is \( \Delta X = \frac{1}{2} (10^{-3}) = 0.0005 \) and maximum relative percentage error is

\[
\frac{\Delta X}{|X|} \times 100\% = \frac{0.0005}{1.325} \times 100\% = 0.03773585\%.
\]

ERROR PROPAGATION

Numerical computing involves a number of computations consisting of basic arithmetic operations. Therefore, it is not the individual round-off errors that are important but the final error on the result. Our major concern is how an error at one point in the process propagates and how it affects the final error. In this section we will discuss the arithmetic of error propagation and its effect.

Addition and Subtraction

Consider addition of two number \( X = X_1 + E_x \) and \( Y = Y_1 + E_y \), where \( E_x \) and \( E_y \) are the errors in \( X_1 \) and \( Y_1 \) respectively.

Then,

\[
\frac{X + Y}{True} = \frac{X_1 + Y_1 + \left( E_x + E_y \right)}{Approx.}\]

Therefore, total error is

\[
E_{x+y} = E_x + E_y
\]

Similarly, for the subtraction

\[
E_{x-y} = E_x - E_y
\]

Note that the addition \( E_x + E_y \) does not mean that error will increase in all cases. It depends on the sign of individual errors. Similarly, in the case with subtraction.

Generally, we do not know the sign of the errors; we can estimate error bounds, that is
\[
E_{x+y} = \left| E_{x+y} \right| \leq \left| E_x \right| + \left| E_y \right| \quad \text{(Triangle Inequality)}
\]

Therefore, the magnitude of the absolute error of a sum (or difference) is less than or equal to the sum of the magnitude of the errors.

Note that while adding up several numbers of different absolute accuracies, the following procedure may be adopted:
1. Isolate the number with greatest absolute error.
2. Round-off all other number retaining in them one digit more than in the isolated number.
3. Add up, and
4. Round-off the sum by discarding last digit.

**Example:** Find the sum of the following numbers:
1.35265, 2.00468, 1.532, 28.201, 31.00123, where each of which are correct to given digits. Also find total absolute error.

**Solution:**
We have two numbers 1.532 and 28.201 having greatest absolute error of 0.0005.
Round-off all other numbers to four decimal digits. These are
1.3527, 2.0047, 31.0012

The sum of all the numbers is given by
\[
S = 1.3527 + 2.0047 + 31.0012 + 1.532 + 28.201
\]
\[
= 64.0916
\]
\[
= 64.092 \text{ (Rounding-off by discarding last digit)}
\]
To find absolute error:
Two numbers have each an absolute error of 0.0005 and three numbers have each an absolute error of 0.00005.
Therefore, absolute error in sum of all five numbers is
\[
E_A = 2 \times 0.0005 + 3 \times 0.00005
\]
\[
= 0.00115
\]
In addition to above absolute error, we have to take into account the rounding-off error in sum S and which is 0.0004.
Therefore, total absolute error in sum is
\[
E_T = 0.00115 + 0.0004
\]
\[
= 0.00155
\]

Thus, 
\[
S = 64.092 \pm 0.00155
\]

**Multiplication**

Let us consider the multiplication of two numbers
\[
XY = (X_1 + E_x)(Y_1 + E_y)
\]
\[
XY = X_1Y_1 + X_1E_y + Y_1E_x + E_xE_y
\]
Errors are normally small and their product will be much smaller. Therefore, if we neglect the product of the errors, i.e. \( E_x E_y \), we get

\[
\frac{XY}{\text{True}} = \frac{X_1 Y_1}{\text{Approx.}} + X_1 E_x + Y_1 E_y
\]

Then, the total error, \( E_{xy} \), is

\[
E_{xy} = X_1 E_x + Y_1 E_y
\]

Therefore,

\[
E_{xy} = \left| X_1 Y_1 \left( \frac{E_x}{X_1} + \frac{E_y}{Y_1} \right) \right|
\]

**Division**

We have,

\[
\frac{X}{Y} = \frac{X_1 + E_x}{Y_1 + E_y}
\]

Multiplying both numerator and denominator by \( Y_1 - E_y \), we get

\[
\frac{X}{Y} = \frac{X_1 + E_x}{Y_1 + E_y} \times \frac{Y_1 - E_y}{Y_1 - E_y}
\]

Rearranging the terms, we get

\[
\frac{X}{Y} = \frac{X_1 Y_1 + Y_1 E_x - X_1 E_y - E_x E_y}{Y_1^2 - E_y^2}
\]

Dropping all terms that involves only product of errors, we have

\[
\frac{X}{Y} = \frac{X_1 Y_1 + Y_1 E_x - X_1 E_y}{Y_1^2}
\]

Thus,

\[
\frac{X}{Y} = \frac{X_1}{Y_1} \left( 1 + \frac{E_x}{X_1} - \frac{E_y}{Y_1} \right)
\]

Thus,

\[
E_{x/y} = \frac{X_1}{Y_1} \left( \frac{E_x}{X_1} - \frac{E_y}{Y_1} \right)
\]

Applying triangle inequality,
\[
E_{x/y} \leq \left| E_{x/y} \right| = \left| \frac{X_1}{Y_1} \right| \left( \left| \frac{E_x}{X_1} \right| + \left| \frac{E_y}{Y_1} \right| \right)
\]

Note that while multiplying (or dividing) any two numbers of different absolute accuracies, the following procedure may be adopted:

5. Isolate the number with greatest absolute error.
6. Round-off all another number so that it has same absolute error as in the isolated number.
7. Multiply (or divide) the numbers
8. Round-off the result so that it has the same significant digits as in the isolated numbers.

**Example:** Find the product of the numbers 56.54 and 12.4 which are both correct to significant figures given

**Solution:** Here, the number 12.4 has greatest absolute error of 0.05 so we round off the second number to one decimal digits, i.e, 56.5

Then, the product is given by
\[ P = 12.4 \times 56.5 = 700.6 \]

Now, round-off the product 3 significant digits because the isolated number 12.4 has three significant digits, we get
\[ P = 701 \]

Absolute error, \( E_A = 0.05 \times 56.5 + 0.05 \times 12.4 \]
\[ = 3.445 \]

Round-off error = 0.4

Total absolute error, \( E_T = 3.445 + 0.4 = 3.845 \)

**A General Error Formula**

Here we derive a general formula for the error committed in using a certain formula or a functional relation.

Let \( u = f(x_1, x_2, \ldots, x_n) \) be a function of several variables \( x_i, i = 1, 2, \ldots, n \) and let \( \Delta x_i \) be the error in each \( x_i \). Then the error \( \Delta u \) in \( u \) is given by

\[
\Delta u = f(x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots, x_n + \Delta x_n) - f(x_1, x_2, \ldots, x_n)
\]

Expanding the first term in right hand side by Taylor’s series, we obtain

\[
\Delta u = \frac{\partial f}{\partial x_1} \Delta x_1 + \frac{\partial f}{\partial x_2} \Delta x_2 + \ldots + \frac{\partial f}{\partial x_n} \Delta x_n + \text{Terms involving } (\Delta x_i)^2 \text{ and higher powers of } \Delta x_i.
\]

Assuming that the errors involving in \( x_i \) are small enough that the square power and higher powers of \( \Delta x_i \) can be neglected. Then above relation gives
\[
\Delta u = \frac{\partial f}{\partial x_1} \Delta x_1 + \frac{\partial f}{\partial x_2} \Delta x_2 + \cdots + \frac{\partial f}{\partial x_n} \Delta x_n.
\]

The maximum absolute error is
\[
(\Delta u)_{\text{max}} = \left| \frac{\partial f}{\partial x_1} \Delta x_1 \right| + \left| \frac{\partial f}{\partial x_2} \Delta x_2 \right| + \cdots + \left| \frac{\partial f}{\partial x_n} \Delta x_n \right|.
\]

The formula for the relative error follows
\[
E_R = \frac{\partial f}{\partial x_1} \frac{\Delta x_1}{u} + \frac{\partial f}{\partial x_2} \frac{\Delta x_2}{u} + \cdots + \frac{\partial f}{\partial x_n} \frac{\Delta x_n}{u}.
\]

The maximum relative error is given by
\[
(E_R)_{\text{max}} = \left| \frac{\partial f}{\partial x_1} \frac{\Delta x_1}{u} \right| + \left| \frac{\partial f}{\partial x_2} \frac{\Delta x_2}{u} \right| + \cdots + \left| \frac{\partial f}{\partial x_n} \frac{\Delta x_n}{u} \right|.
\]

Example: If \( u = \frac{4x^2y^3}{z^4} \) and errors in \( x, y, z \) be 0.001. Compute the maximum absolute error and relative error in evaluating \( u \) when \( x = y = z = 1 \).

Solution:

The given function is \( u = \frac{4x^2y^3}{z^4} \)

The maximum absolute error is given by
\[
(\Delta u)_{\text{max}} = \left| \frac{\partial u}{\partial x} \Delta x \right| + \left| \frac{\partial u}{\partial y} \Delta y \right| + \left| \frac{\partial u}{\partial z} \Delta z \right|
\]

And, the maximum relative error is
\[
(E_R)_{\text{max}} = \frac{(\Delta u)_{\text{max}}}{u}
\]

Here,
\[
\frac{\partial u}{\partial x} = \frac{8xy^3}{z^4}, \text{ so } \frac{\partial u}{\partial x}_{(1,1,1)} = 8;
\]
\[
\frac{\partial u}{\partial y} = \frac{12x^2y^2}{z^4}, \text{ so } \frac{\partial u}{\partial y}_{(1,1,1)} = 12;
\]
\[
\frac{\partial u}{\partial z} = \frac{-16x^2y^3}{z^5}, \text{ so } \frac{\partial u}{\partial z}_{(1,1,1)} = -16;
\]
The Errors in x, y, z are respectively $\Delta x = \Delta y = \Delta z = 0.0001$.

Therefore, the maximum absolute error is

$$
(\Delta u)_{\max} = |8 \times 0.0001| + |12 \times 0.0001| + |16 \times 0.0001|
$$

\[= 0.0035 \]

The maximum relative error is

$$
(E_r)_{\max} = \frac{0.0036}{4} = 0.0009
$$

The maximum percentage error is

$$
(E_p)_{\max} = (E_r)_{\max} \times 100\% = 0.09\%
$$

**Error in Series Approximation**

Let $f(x)$ be a continuously differentiable function on an interval subset of $\mathbb{R}$. If the value of the function $f$ at an interior point $x_i$ is known, i.e., $f(x_i)$ is known, then the value of the function at next successive point $x_i + h$ is approximated by the Taylor’s series given by

$$
f(x_i + h) = f(x_i) + \frac{h}{1!} f'(x_i) + \frac{h^2}{2!} f''(x_i) + \ldots + \frac{h^n}{n!} f^{(n)}(x_i) + R_{n+1}(h),
$$

where $R_{n+1}(h) = \frac{h^{n+1}}{(n+1)!} f^{(n+1)}(\xi), \quad x_i < \xi < x_i + h$.

The last term $R_{n+1}(h)$ is called the remainder term which for a convergent series, tends to zero as $n \to \infty$. Thus, if $f(x_i + h)$ is approximated by first $n$ terms of the series, the maximum error committed by using this approximation, called $n^{th}$ order approximation, is given by the remainder term $R_{n+1}(h)$.

Conversely, if the accuracy required is specified in advance then it would find the number of terms such that the finite series yields the required accuracy.

The above series can also be written as

$$
f(x_i + h) = f(x_i) + \frac{h}{1!} f'(x_i) + \frac{h^2}{2!} f''(x_i) + \ldots + \frac{h^n}{n!} f^{(n)}(x_i) + O(h^{n+1})
$$

Where, $O(h^{n+1})$ means, the truncation error is of the order of $h^{n+1}$.

For example,

i. $f(x_i + h) = f(x_i) + O(h)$ is zero-order approximation.

ii. $f(x_i + h) = f(x_i) + \frac{h}{1!} f'(x_i) + O(h^2)$ is first-order approximation and so on.